JMIR Medical Informatics

Impact Factor (2022): 3.2
Volume 9 (2021), Issue 6 ISSN 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Viewpoints

Review

Original Papers

Viewpoint

Blockchain Applications in Health Care and Public Health: Increased Transparency

Pedro Elkind Velmovitsky[1], BSc, MSc; Frederico Moreira Bublitz[1,2], BSc, MSc, PhD; Laura Xavier Fadrique[1], MSc, PMP; Plinio Pelegrini Morita[1,3,4,5,6], PEng, MSc, PhD

[1]School of Public Health and Health Systems, University of Waterloo, Waterloo, ON, Canada

[2]Center for Strategic Technologies in Health (NUTES), State University of Paraiba (UEPB), Campina Grande, Brazil

[3]Institute of Health Policy, Management, and Evaluation, University of Toronto, Toronto, ON, Canada

[4]Research Institute for Aging, University of Waterloo, Waterloo, ON, Canada

[5]Department of Systems Design Engineering, University of Waterloo, Waterloo, ON, Canada

[6]eHealth Innovation, Techna Institute, University Health Network, Toronto, ON, Canada

Corresponding Author:
Plinio Pelegrini Morita, PEng, MSc, PhD
School of Public Health and Health Systems
University of Waterloo
200 University Ave W
Waterloo, ON, N2L 3G1
Canada
Phone: 1 15198884567 ext 41372
Email: plinio.morita@uwaterloo.ca

Abstract

Background: Although big data and smart technologies allow for the development of precision medicine and predictive models in health care, there are still several challenges that need to be addressed before the full potential of these data can be realized (eg, data sharing and interoperability issues, lack of massive genomic data sets, data ownership, and security and privacy of health data). Health companies are exploring the use of blockchain, a tamperproof and distributed digital ledger, to address some of these challenges.

Objective: In this viewpoint, we aim to obtain an overview of blockchain solutions that aim to solve challenges in health care from an industry perspective, focusing on solutions developed by health and technology companies.

Methods: We conducted a literature review following the protocol defined by Levac et al to analyze the findings in a systematic manner. In addition to traditional databases such as IEEE and PubMed, we included search and news outlets such as CoinDesk, CoinTelegraph, and Medium.

Results: Health care companies are using blockchain to improve challenges in five key areas. For electronic health records, blockchain can help to mitigate interoperability and data sharing in the industry by creating an overarching mechanism to link disparate personal records and can stimulate data sharing by connecting owners and buyers directly. For the drug (and food) supply chain, blockchain can provide an auditable log of a product's provenance and transportation (including information on the conditions in which the product was transported), increasing transparency and eliminating counterfeit products in the supply chain. For health insurance, blockchain can facilitate the claims management process and help users to calculate medical and pharmaceutical benefits. For genomics, by connecting data buyers and owners directly, blockchain can offer a secure and auditable way of sharing genomic data, increasing their availability. For consent management, as all participants in a blockchain network view an immutable version of the truth, blockchain can provide an immutable and timestamped log of consent, increasing transparency in the consent management process.

Conclusions: Blockchain technology can improve several challenges faced by the health care industry. However, companies must evaluate how the features of blockchain can affect their systems (eg, the append-only nature of blockchain limits the deletion of data stored in the network, and distributed systems, although more secure, are less efficient). Although these trade-offs need to be considered when viewing blockchain solutions, the technology has the potential to optimize processes, minimize inefficiencies, and increase trust in all contexts covered in this viewpoint.

(JMIR Med Inform 2021;9(6):e20713) doi:10.2196/20713

KEYWORDS

health care; blockchain; EHR; health insurance; drug supply chain; genomics; consent; digital ledger; food supply chain

Introduction

Background

Global society is moving into an age of ubiquitous and smart technologies that monitor our health, such as smart devices, Internet of Things solutions, and ambient assisted living systems. These technologies allow continuous and effortless health data collection at a previously unseen scale [1,2], generating rich and massive data sets, known as big data [3].

The *age of big data* can lead to a change in the way health care is delivered. Generally, health care is reactive, in which individuals interact with health care services when there is something wrong [4,5] and usually to treat acute diseases, instead of proactive, in which real-time monitoring of health data from different sources leads to predictions and insights into individual and population health, as opposed to checkups with health services when a problem appears [4,5]. In this manner, a proactive and predictive health care model includes surveillance and monitoring of individuals through remote sensing technologies, such as smart bands and smart thermostats, generating large volumes of diverse and real-time data in a cost-effective manner. The use of such technologies in a community will also enable public health surveillance on a scale never seen before, allowing public health agencies to better understand the socioeconomic determinants of health and prevent disease outbreaks [5,6].

However, to achieve this model of health care, there are challenges that need to be overcome. For example, health records are stored by different providers in systems that lack interoperability [7,8]. This makes data sharing difficult and prevents doctors from having a complete view of a patient's health [7,8]. Interoperability issues and costs also affect the availability of genomic data and minimize their benefits [9]. In addition, increasingly advanced methods of data collection and analysis of personal, medical, and genomic data raise concerns regarding ownership, privacy, and regulations of health data [1,3].

One possible tool to overcome or mitigate these challenges is blockchain [6,10-13]. This technology can be seen as a distributed virtual ledger that records timestamped transactions [6,12,13]. Cryptography is used to ensure that when a block is added to the blockchain, it cannot be tampered with [12]. Hence, blockchain is a tamperproof digital ledger in which all participants view an immutable version of the truth, making it ideal to track an asset and enable trust among parties (eg, health data or user consent for data collection) [6,7,12,14].

In 2016 and 2018, IBM Corporation surveyed more than 400 health care and life sciences executives on the use of blockchain technology. Among their findings, more than half of the executives in both industries had plans to adopt it by 2020 [6,10,11]. Given the perceived potential of blockchain by industry experts from multiple areas [6,10-12] and to help guide the implementation of digital solutions that can solve pressing needs in health care systems, the aim of this study is to review current blockchain solutions being developed by the health care industry. This paper provides a comprehensive view of the blockchain health care industry, providing guidance to innovators about how to leverage this technology in daily operations and how to implement solutions that can help evolve health care delivery. The COVID-19 outbreak has created an increased demand for home-based digital health solutions such as telehealth and telemonitoring [15], increasing the importance of using technologies such as blockchain to increase the transparency of digital transactions and data provenance [16,17].

Related Work

McGhin et al [18] provide an overview of the main opportunities and challenges for blockchain in the health care field and describe some initiatives (both in academia and industry) focused on developing blockchain solutions. Vazirani et al [19] detailed a systematic review examining the feasibility of blockchain for electronic health record (EHR) systems, finding several trade-offs that need to be considered during the design and development of blockchain. Trade-offs were further explored by O'Donoghue et al [20].

Farouk et al [13] provided a similar review to this one on the use of blockchain in the health care industry but mostly focused on its integration with Internet of Things devices and record management. Hasselgren et al [21] conducted a scoping review of blockchain in health care and, while focusing on peer-reviewed publications rather than the industry, they found that both the number and quality of blockchain research is growing.

Chukwu and Garg [22] provide a systematic review of blockchain applications specifically for the use of EHRs and health data sharing and do not focus on industry applications.

Agbo et al [23] conducted a systematic review of blockchain applications in health care, also focusing on academic literature, although some studies mention companies working with blockchain. The use cases found in this work are very similar to the use cases explored in this paper (suggesting a convergence between academia and industry research), but Agbo et al [23] found a predominance of studies focusing on EHRs when compared with other areas.

Most of these did not have an industry focus; rather, they usually discussed the computer science aspects of the technology or evaluated mostly academic work. In addition, as found by Agbo et al [23], most reviews focused on EHRs and not on additional use cases. Therefore, this review contributes to preview work by providing an overview of blockchain applications in the health care industry, while identifying what challenges and use cases are the current focus of health care companies working with blockchain.

Methods

Overview

This narrative review [24] focuses on providing eHealth experts with a comprehensive narrative review of blockchain in health care. Blockchain is a novel technology that can provide increased transparency to data transactions in health care and public health [6,10,11]. Owing to its novelty and early stage implementation, significant development has been accomplished at the industry level, driving this review toward a combination of peer-reviewed academic literature and gray literature.

Our aim was to analyze blockchain in health care from an *industry* perspective, focusing on solutions developed by health and technology companies (although results from research and development initiatives and academia were used to complement knowledge when necessary).

Although not a scoping review, this paper followed the framework defined by Levac et al [25] for scoping reviews, ensuring that the findings were analyzed in a systematic manner. This framework consists of six stages: (1) identifying the research question (RQ); (2) identifying relevant studies; (3) selecting studies; (4) charting the data; (5) collating, summarizing, and reporting results; and (6) consultation (optional). In this narrative review, we leveraged phases 1-5.

Identifying the RQ

The primary objective is to identify how the health care industry views the potential of blockchain to solve current challenges. To fulfill this, two secondary goals need to be achieved: we must understand how blockchain works and the challenges facing the industry. Therefore, the following RQs were used to guide the reviews:

1. How do the blockchain systems work?
2. What are the current challenges faced by the health care industry today that can be addressed by blockchain technology?
3. For each of these challenges, which blockchain solutions are being developed by the health care industry?

Identifying Relevant Studies

Our review analyzes how the health care industry perceives the blockchain's potential to solve current challenges. To this end, we looked at gray literature in addition to traditional databases such as IEEE and PubMed, including search and news outlets such as Google Scholar, CoinDesk [26], CoinTelegraph [27], and Medium [28]. The keywords were a combination of "blockchain," "distributed ledger," "health," "industry," and "health care." Whenever possible, we looked at technical reports (usually available on companies' websites) in addition to news articles.

Study Selection

The primary exclusion criteria involved selecting solutions that address issues or challenges in health care. Blockchain solutions that only had applications in unrelated fields were not included. Additional restrictions included practical concerns regarding availability and language (only English references were included).

Charting the Data

To extract useful insights from the publications, we focused on two main types of information:

- What are the main health care challenges that the solution aims to improve?
- How is blockchain being used to improve the challenges?

More specifically, we looked at the main objective of the blockchain solution and the methods in which blockchain is being developed. Relevant bibliographical information, including title, authors, country, and year, was also extracted. This review focused on technical reports. If the technical report did not provide sufficient information, web articles were used to complement the results.

Collating, Summarizing, and Reporting the Results

Following the recommendations presented by Levac et al [25], the steps are as follows:

- Analysis: for each solution being presented, we mapped the challenges addressed and how blockchain is being used.
- Reporting results: after presenting additional information on blockchain, we will describe the challenge in question and its importance in health care, followed by a discussion on how blockchain is being used by the industry in this context.
- Implications for future research, practice, and policy: this final step will be addressed in the Discussion section, where we discuss the limitations of blockchain and additional concerns.

Results

Overview

We started this review by presenting relevant background information about blockchain, followed by an overview of the main challenges identified in our review: EHRs, supply chain, health insurance, genomics, and consent management. For each of these areas, we have also presented blockchain solutions developed by industry. Table 1 provides a summary of the results by describing each of the five identified challenges explaining how blockchain can offer a solution, along with examples.

Table 1. Results of the literature review.

Challenges	Description	Solutions
Electronic health records	Blockchain can provide an overarching framework that allows transparent and auditable access to disparate individuals' health records stored off-chain. Patients would control data sharing parameters and access. Some solutions also discuss integrating health data from less traditional sources (eg, connected devices) and the creation of a health data marketplace, in which patients can sell their data to buyers through crypto tokens	• MedRec [7,8,29], PatientTruth [30,31], CareX [32,33], MEDIS [34,35], GEM [36-40], MedicalChain [41,42], Humantiv and Medoplex [43-45]
Supply chain	Blockchain can establish an immutable record of a product's tracing throughout the supply chain. In the case of health care, there have been many solutions that implement a blockchain to track-and-trace drugs and food products. In addition, smart contracts can be used as monitoring and alert systems for proper transport conditions (eg, a certain temperature range)	• Drug supply chain: BlockVerify [46-48], Merck [49,50], Modum [51-54] • Food supply chain: IBM Food Trust [55-58], Alibaba and Ant Financial [59,60]
Health insurance	Smart contracts on the blockchain can potentially help to settle health insurance claims and manage payment in real time, making the process more efficient and transparent for payers, providers, and patients. Other potential use cases include pharmaceutical and medical benefits, checks, and payment risk calculation	• PokitDok and DokChain [61-67], GEM [39], Payspan [68,69]
Genomics	Much like with electronic health records, blockchain can provide a mechanism for controlling access to separate existing data banks of genetic information. In addition, blockchain can directly connect sellers of genomic data-to-data buyers, creating a genomic data marketplace. Data buyers could even provide rewards for individuals to sequence their genomes, creating their own data sets (eg, providing crypto tokens to individuals with a certain feature to be researched, in return for their genomic information)	• Nebula Genomics [9,70], LunaDNA [71-75], Shivom [76-79], Zenome [80,81], EncrypGen [82-85], Macrogen [86-88]
Consent management	Blockchain can provide an immutable and timestamped log of consent, allowing individuals to grant and revoke consent for different data types and periods. In the case of health studies, it can also help researchers to easily track, manage, and update user consent	• My31 app [89,90], Bitfury [91,92], HealthVerity Consent [93], Verifiable Audit Trail (tracking of events related to health data) [94-98], INSERM[a] and APHP[b] consent project [14], Queen's University BlockTrial [99], Patient Control and Consent Blockchain initiative [100-102], Ubiquitous Health Technology Lab [6,103]

[a]INSERM: Institut National de la Santé Et de la Recherche Médicale.

[b]APHP: Assistance Publique-Hôpitaux de Paris.

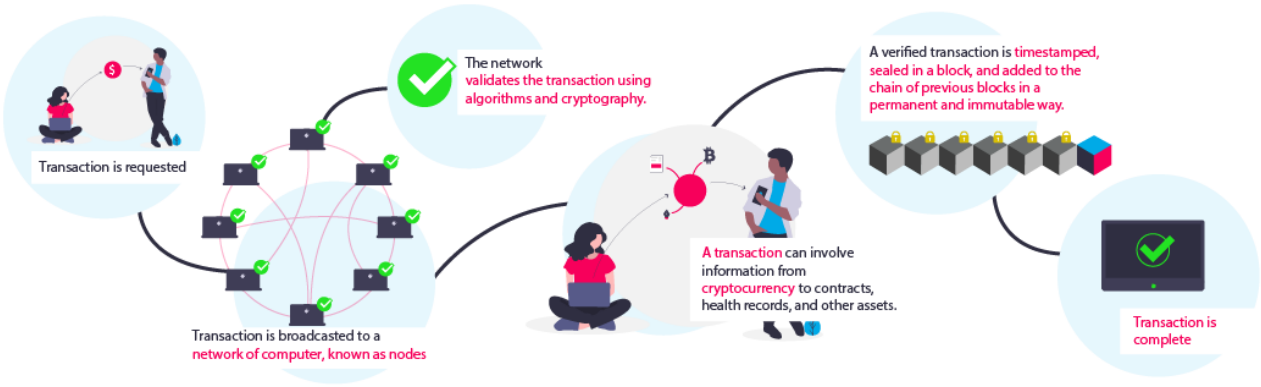
What Is Blockchain?

Blockchain is a virtual distributed ledger that records transactions among parties. It is operated by a network of computers in which each participant is called a node and possesses a copy of the ledger, regularly updated to ensure consistency. In other words, all nodes have access to the exact information [12,18].

When a user makes a transaction, this transaction is timestamped and *sealed* in a block [12]. Through a consensus mechanism, this block is linked to previous existing blocks—hence the name blockchain. Different blockchains (eg, Bitcoin and Ethereum) have different consensus mechanisms [12]. A typical consensus mechanism, called Proof of Work, requires the nodes in the network to randomly guess a number that solves a mathematical puzzle; the first node to discover it seals the block. This process is called mining [6,12,13,18].

The linkage between blocks is achieved through a method called hashing, in which new blocks point to the previous ones [12]. This technique converts data into a string of characters, called a hash. For example, a user may convert a text into the following hash: "f1abc234b79f6d6ay42a12c53468a1b13553r1r0fgr4039 rf08h958b5232b9n8." If a single character from this hash is changed, an entirely new string is generated. Although it is easy to generate a hash from a piece of information, it is impossible to discover the original information from a hash [12,22]. The Bitcoin blockchain hashes the nonce, alongside the transaction information and the hash of the preceding block. If a malicious party tries to tamper with information already stored in a block, the hash is altered, breaking the chain. This ensures that the blocks cannot be tampered with, and the information contained in the blockchain cannot be altered. Therefore, blockchain is a tamperproof digital ledger where all participants have access to an immutable version of the truth [12,18]. The flow of a transaction in the blockchain is shown in Figure 1.

Figure 1. Flow of a transaction in the blockchain.



Blockchain is a type of distributed ledger technology, in which a consistent ledger is shared among parties to store a record, creating a distributed database. It is a distributed ledger technology that uses cryptographic and consensus mechanisms to increase trust [22].

There are also different types of blockchains. Although the nomenclature varies, they are usually defined as follows [13,104,105]:

- Public blockchain: all participants can read and write new information to the ledger. Although new information can be added, no information can be deleted. Bitcoin is an example of a public blockchain.
- Permission (consortium and federated) blockchain: this is owned by a consortium of participants who define the permissions for joining and updating the network. For example, a consortium blockchain owned by health care providers can allow patients to change their information, but only providers may upload new information.
- Private blockchain: this is owned by a single entity that manages access, permission to read or write data, and even data deletion. Among the blockchain communities, some are of the opinion that private blockchains defeat the purpose of decentralized technologies by introducing a central authority.

From a health care perspective, one of the biggest concerns in capturing and coding patient information is privacy [106]. Several blockchain implementations allow the creation of smart, codified contracts that allow for the storage of immutable information. For example, Ethereum enabled the creation of smart contracts that codify contract agreements. When several parties agree to a transaction, they create mechanisms to ensure trust [6,12,107]. Smart contracts write the terms of a contract in code, which is executed on the blockchain, and has *the ability to be self-executing and self-enforcing* [12]. Therefore, smart contracts can minimize trust concerns among parties [12,18,107].

Blockchain's features and design make it a model for processes plagued by trust issues [6,12,108], and it is ideal for increasing trust in contexts involving parties that do not have reason to trust each other [6,12,108]. One such context is health care [6].

Blockchain in Health Care: Challenges and Solutions

The following subsections describe a challenge in health care and discuss blockchain solutions being developed by companies to address them.

Electronic Health Records

Challenge Description

EHRs digitally store patients' health data [6,7,109,110]. However, data are fragmented throughout EHR systems: patients often interact with different health care providers (usually the stewards of the data), creating challenges related to accessing past information [6,7,22]. In addition, providers have different EHR systems that may not be fully interoperable [6,7]. These factors contribute to difficulties in data sharing [6,7].

Patients' health data end up in silos and cannot be integrated with data from other providers or sources, such as connected devices. Ultimately, there is no easy way to obtain a holistic view of a patient's health, leading to errors, delays, and poorer health outcomes [6,7,18,110].

Use of Blockchain

Blockchain solutions can create an overarching hub, potentially on the cloud, to link all records of individual patients [7,8,18,29], without storing health data on the blockchain itself [7,8,29,111]. Rather, the blockchain infrastructure would act as a hub that points to the location of a patient's records off-chain [7,8,29]. Data access and changes to records can be tracked and displayed to the patient in real time. Furthermore, patients could control access to their records by giving permission to providers, researchers, and third parties to access their data. In this manner, an EHR-blockchain solution would allow for all health data from individuals to be accessed and controlled by the patient, facilitating a complete view of patients' health [7,8,18,29]. This solution would also give patients greater control and transparency over their health data [7,8,12,18,19,29].

For example, MedRec is a blockchain-enabled solution for EHRs [7,8,29]. It is a system developed by the Massachusetts Institute of Technology that provides a transparent view of medical history. MedRec uses smart contracts in Ethereum to encode metadata by referencing medical data from different sources, including information about ownership and permission. These references "create an accessible bread crumb trail for medical histor[ies]" [7]. Providers may append a new patient

record in MedRec, but patients are the ones who give permission for data to be accessed and shared. This increases transparency and allows patients to keep track of their records [7,8].

Similar solutions include PatientTruth [30,31], CareX [32,33], MEDIS [34,35], and MedicalChain [41,42]. Typically, EHR-blockchain solutions store references to off-chain files containing EHRs and also work with less traditional sources, such as data from connected devices. Patients control their records and with whom they wish to share their data (eg, health care professionals, hospitals, and insurance providers). Some of these solutions allow patients to sell deidentified records (eg, to health studies) with a crypto token from the platform. The financial component creates a form of health data marketplace in which patients own and are able to profit from their health data.

Another organization working with blockchain in an EHR context is CitizenHealth [33,43], which developed two solutions: Humantiv [44], which also combines data from EHRs and other sources with an added gamification component in which patients earn rewards according to their health indicators, and Medoplex [45], the company's marketplace component. GEM, a US-based start-up, is developing a solution that uses Ethereum to create a shared network where providers have real-time access to medical documents [18]. GEM is partnering with Nordic-based Tieto to create a blockchain platform that enables patient control over medical records and genomic data [40,112,113].

It is important to note that a blockchain infrastructure, as described above, could mitigate data sharing issues by providing an interoperable, auditable, and secure landscape of transactions controlled by data owners. This, in turn, would allow easy and transparent access to disparate health records. As stated by McGhin et al [18], when discussing blockchain cloud infrastructures, "the role of blockchain in cloud data infrastructure is facilitating the creation of a decentralized and trusted cloud data provenance architecture that allows tamperproof records, greater transparency of data accountability, and enhanced privacy and availability of the data." However, the blockchain itself does not impact the interoperability of the health data itself or the local systems in which it is stored. Rather, it acts as an overarching infrastructure with references to off-chain resources whose access is auditable, secure, and transparent to all authorized parties within the distributed network.

Drug Supply Chain

Challenge Description

One of the biggest challenges faced by pharmaceutical companies today is counterfeit drugs. In total, US $200 billion are lost to counterfeit drugs annually, and their use puts patients' lives at risk [46]. Manufacturers do not have a unified and interoperable system of supply chain management, lack incentives to share data and information, and are consequently siloed, making end-to-end traceability and drug provenance difficult [46,114].

In the United States, the Drug Supply Chain Security Act (DSCSA) established a set of requirements that must be implemented by pharmaceutical companies until 2023. These requirements include product tracing and verification [46,115,116].

Use of Blockchain

By storing transactional data from the supply chain on blockchain, it is possible to establish an *immutable record of provenance* [117]. Blockchain can provide a transparent ledger that traces products throughout the supply chain, from manufacturing to distribution. This will ensure compliance with the DSCSA and improve patient safety. Furthermore, blockchain can also track whether products are being transported and handled under appropriate conditions [111].

One of the companies working in this scenario was BlockVerify. The company is working on a DSCSA-compliant solution that traces products and identifies counterfeit drugs [46-48] A product is labeled with BlockVerify's tag and verified along the supply chain, with a permanent record on a private blockchain. Consumers and retail locations can use this record to ensure that the product is genuine. Merck has filed a patent to use blockchain to track drug information in the supply chain; while the company already has systems to prevent fraud, it is expected that blockchain could minimize existing inefficiencies [49,50]. Modum uses blockchain to ensure that medical products are being transported at the correct temperature [51-54]. Sensor devices are added to shipments, and smart contracts for each shipment are fixed with sensors, including the alarms. The sensors monitor the temperature during transportation and, when the shipment is received, the data are transferred to a blockchain and the smart contract evaluates whether conditions and regulations have been met [51-54].

Efforts have also been made to tackle the food supply chain. IBM Food Trust is a permissioned blockchain that allows stakeholders to view the supply chain history of a food item and complementary information (eg, certifications, testing, temperature, and location). For example, organizations can identify when a food item is contaminated and trace the contamination back to its source [55-58]. Another solution is being used in China: Ant Financial, an affiliate of the e-commerce enterprise Alibaba, is using a permissioned blockchain that tracks the production of rice in the city of Wuchang. Quick Response codes were added to rice packages, so that users scan these codes to obtain information (eg, location of the harvest, type of seed used, or transportation). Alibaba is also working on developing a food trust framework that, in its pilot phase, tracked shipments from China to Australia and New Zealand [59,60].

Health Insurance

Challenge Description

There is a lack of trust between payers, providers, and patients with a complete view and coordination of health care [6]. Patients often pay expensive premiums while dealing with a lack of transparency and the ability to compare prices, in addition to the risk of insurance fraud that affects all stakeholders [32,41,69]. Providers must go through complex and bureaucratic processes to submit a claim [41]. The challenging ecosystem can be exemplified through the claims management process: (1) first, a provider must be covered by

health insurers—meaning that the provider must also maintain benefits' databases and keep track of services delivered, adding additional expenses; (2) if patients receive services from the provider, the insurer checks the service against the patients' health plans to check their eligibility; (3) the process takes several weeks and involves multiple people checking agreements, leading to delays [41].

Use of Blockchain

In 2017, a survey revealed that 98% of payers with over 500,000 members are pursuing blockchain-enabled solutions [69]. Through the use of smart contracts, it is possible to codify terms of agreement between providers, payers, and patients, automating processes and minimizing inefficiencies. For example, PokitDok developed DokChain, a private blockchain that references off-chain file systems. DokChain contains several smart contracts that request and return data from health insurance providers and payers in real time [61-63], possibly enabling real-time status checks and mechanisms for error identification. The goal is for DokChain to make decisions on insurance claims in real time using smart contracts. As soon as a patient receives a service, it is recorded on DokChain and visualized by all stakeholders. Smart contracts can determine whose responsibility the claim falls under and how much is owed to each party, processing that amount in real time [62,64,65]. PokitDok also offers application programming interfaces to calculate payment risk for a patient [66] and to allow patients to schedule and pay for health services [67], automating payments and checking benefits in real time.

The start-up GEM created a blockchain prototype in 2017 that helped to settle a claim in less than 5 minutes [39]. Payspan, a 25-year-old health care reimbursement company, is also working on blockchain networks to connect providers and payers for claims management and payment processing [68,69].

Initiatives that focus on the creation of a health data marketplace (eg, Medoplex) want to allow health care buyers (eg, patients) and sellers (eg, providers) to connect without the need for insurance companies. Their goal is to create a marketplace platform where patients can search, select, and pay for health care services using a crypto token. CareX, MedicalChain, and BlockRx are other solutions that allow for the payment of health services with crypto tokens obtained from health record sharing [32,41,46,111,114]. Many companies that deal with EHRs seem to view health insurance as a complementary use case: by creating a health data marketplace platform that allows patients to share their data with health care providers and professionals, the idea seems to be that intermediaries (eg, insurance companies) will not be needed or their role will be diminished.

Genomics

Challenge Description

DNA is a molecule that encodes the genetic instructions of organisms [9], where genes are collections of DNA [118] and a genome is the collection of an organism's genes [9,119]. The human genome comprises more than 20,000 genes [9,81,120,121].

The study of genomics data in conjunction with social and environmental determinants of health gave rise to precision medicine [122]. This field of research and treatment can help individuals and researchers better understand the cause of diseases, contribute to the development of new drugs, and aid in the creation of personalized interventions for individuals with specific genetic traits, in addition to many other benefits [123]. However, for this potential to be fully realized, large volumes of genomic data are required [9,124].

There are several barriers to the availability of massive genomic data sets, including security and privacy concerns, prohibitive costs, and data sharing [9]. The latter is related to a lack of interoperability between systems that store genomic data and to data ownership [9]. In addition, a human genome generates over 200 GB of data, and it is estimated that more than 100 million genomes will be sequenced by 2025. Therefore, storage and network transfer speeds also limit data sharing [9,81,121].

Use of Blockchain

The business model for genomic data involves individuals hiring companies such as 23andMe and Ancestry to sequence their genome and receive results. These companies then sell the sequenced data to researchers. With a blockchain network that connects data sellers (eg, an individual who sequenced their genome) and data buyers (eg, a pharmaceutical or research company), without the need for personal genomic companies acting as intermediaries, the transparency of the process will increase while costs will decrease. In addition, data sharing problems can be improved by connecting disparate genomic records on the blockchain, similar to what MedRec and other blockchain-EHR solutions have been proposing [9,111]. Although this will not solve all interoperability issues in the industry, it will allow easier, secure, and transparent access to disparate records of data, facilitating data sharing.

One of the companies that hope to build a genomic and health data marketplace is Nebula Genomics, through the development of a *storage, sharing, and computing platform for biomedical big data* [70]. The company has a partnership with Veritas Genomics. Veritas' platform processes and stores large amounts of genetic information, and Nebula hopes to build on top of it. Individuals will be able to store their genetic information on Veritas' platform and, with blockchain, share and sell their data in the genomic data marketplace. Similar to the case with EHRs, users own their data and control permissions for data sharing, increasing transparency and minimizing security and privacy concerns. Their proposed solution has the potential to [9] (1) minimize costs, since data buyers will acquire data directly from owners who can receive sequencing subsidies from buyers to encourage the generation of genomic data sets (eg, offering sequencing subsidies to individuals with specific traits that a specific research group plans to study); (2) increase transparency, since owners will control access and data sharing will be protected through cryptography. Owners will remain anonymous, while buyers will have to provide information about their identity. Blockchain records all transactions in an immutable log that is easily auditable; and (3) increase the availability of data by integrating data from several sources. The network offers space-efficient data-encoding formats and

leverages computer resources to facilitate the transfer of information [9,70,121].

Other companies have similar solutions, with minor differences. LunaDNA allows anyone to join the blockchain network and win company shares based on shared data in the form of a crypto token. Researchers pay to conduct research on aggregated data, and proceeds earned from research are passed on to stakeholders as dividends. Unlike Nebula Genomics, LunaDNA currently does not provide genome sequencing services, but it accepts files from companies such as 23andMe and Ancestry. Blockchain is used to give individuals ownership of their data and for the generation of an immutable log of transactions [71-75]. Shivom [76-79], Zenome [80,81], EncrypGen [82-85], and Macrogen [86-88] are other similar companies that are using blockchain to empower patients by allowing them to own and share genomic data.

Consent Management

Challenge Description

In health studies, the process of obtaining consent from participants is fallible [125]. The Food and Drug Administration cited the main deficiencies related to consent: the failure to obtain informed consent, use of expired or incomplete forms, failure to provide copies of the forms to participants, missing documents, and changes made to documents without the approval of a review ethics board [14,125]. These problems are aggravated given that reconsent has to be sought in several cases (eg, when there is a revision of the study protocol, new risks are discovered, or there is a worsening of the medical condition of a participant) [126]. Oftentimes, consent needs to be obtained for cohabitants, caregivers, or legal guardians [125-128].

In addition to limitations in traditional consent methods, the global society is moving into an age of ubiquitous smart technologies that monitor our health, which increases the complexity of data collection points and, in turn, of consent management. The challenge of obtaining consent for increasingly advanced methods of data collection, use, and disclosure calls for new solutions to perfect consent procedures and the need to protect the safety of individuals.

Use of Blockchain

Blockchain can provide an immutable and timestamped log of consent, making the process more transparent [6,12,14,111,129]. In the case of health studies, participants will be able to monitor and manage consent, giving informed consent for certain types of data to be collected but not others and revoking their consent at any time. Through the use of cryptographic techniques of identity management, participants can make sure that they are reviewing the latest consent forms and that these were approved by the review ethics board. Researchers would also be benefited, as the measures taken to ensure ethical and legal requirements throughout the research would be clearly auditable. Moreover, it would be easier for them to obtain, track, and update patient consent during the study.

With a consent management platform built on blockchain, authorized parties in the network will have access to timestamped and tamperproof logs of user consent. However, this does not serve as a magic bullet to solve all consent management issues, and researchers must still be careful when collecting informed consent and ensure all ethical and legal requirements. For example, participants may revoke their consent at any time during the study (revocability), and researchers would need to stop collecting participant data at this point despite the fact that data may already have been collected (nonretractability). Similar to traditional health studies, consent forms should include a description of the protocols necessary for revoking consent according to data protection regulations and indicate to participants whether their data are being deleted after their consent is nullified. For example, Article 17 of the General Data Protection Regulation describes the “right to be forgotten” [130], in which personal data must be deleted after consent is revoked.

This would not present a problem for blockchain systems, as the collected data would not be stored on the blockchain, rather the consent of the individual is. If a participant revokes consent, the system will instantly be updated to reflect this and notify the researchers of the change. All consent status updates of participants cannot be tampered with on the blockchain and can be easily auditable in case of problems [129]. Furthermore, if the study uses connected devices for data collection, smart contracts can be developed to ensure that, as soon as a consent is revoked, data immediately stop being collected. What will happen to the data that were already collected depends on the study protocol designed by the researcher. If data deletion is needed, the researcher will have to delete the data manually from their own database (which is not related to blockchain in any way; as mentioned, blockchain only stores the information on consent and not the data as such). In this way, a blockchain-consent platform can facilitate the process of consent management and mitigate some trust issues, but it is still up to researchers to ensure correct and ethical protocols and guidelines are being followed.

Hu-manity.co developed a mobile app called My31, built on IBM blockchain, to help individuals manage their consent to the use of their personal and health information. Users can consent to sharing their data with third parties and researchers, and receive compensation for it [89,90]. Another blockchain-powered platform for managing consent is being developed by Bitfury, with the goal of managing consent for research. All updates related to user consent are timestamped and recorded on the blockchain for future auditing [91,92]. The solution can be used both as a new system and in conjunction with existing systems. HealthVerity is also developing a blockchain solution to manage consent. Unlike other solutions, HealthVerity’s consent platform seems to be more focused on consumer applications that collect health data than medical or clinical research [93].

Although not specifically on consent management, DeepMind (Google’s artificial intelligence conglomerate) is developing a project to allow hospitals and patients to track events related to health data in real time [94-98]. Any interaction with a patient’s health data is recorded on a distributed ledger, which will store information stating that the data were used and their purpose. This project, *Verifiable Audit Trail*, has the potential to increase

transparency in health research and minimize some of the trust issues between stakeholders [94-98].

There have also been several academic initiatives exploring the use of blockchain for consent management. For example, researchers from Institut National de la Santé Et de la Recherche Médicale and Assistance Publique-Hôpitaux de Paris in France have created a proof of concept to manage patient consent in clinical trials that use cryptographic signatures for e-signing. The timestamped consent for different form versions is recorded on the blockchain as a master document [14]. In Canada, Queen's University has also developed a similar solution for clinical trials, titled BlockTrial, where patients assign permission for data access and researchers can query off-chain data [99]. The Toronto-based University Health Network, in partnership with IBM and digital health agencies, is working on the patient control and consent blockchain initiative to allow the permissioned access of data, managed by patients through individual consent and built on blockchain to allow immutable storage of consent directives. Patients can access a mobile app, managing who has access to their data and for what reason. Currently, only University Health Network–produced data are available, but the goal is to enable the integration of data from multiple sources [100-102].

Velmovitsky et al [129] also developed a proof-of-concept blockchain to help patients manage consent, focusing on third-party consumer apps that collect health data such as smart devices. In this prototype, patients can grant and revoke consent for different data types and different periods (eg, users give consent for temperature to be collected but not movement, from September to October). This work uses Hyperledger Fabric to create a blockchain network and proposes a governance structure to collect data from smart devices.

Discussion

Blockchain in Health Care

The health care challenges in the areas of EHRs, supply chains, health insurance, genomics, and consent management are not mutually exclusive. For example, the use of blockchain to address challenges in genomics and EHRs shares many similar challenges regarding data ownership, data sharing, and the creation of a marketplace where owners and buyers can trade data. Consent management and EHRs can also be complementary, with blockchain-enabled solutions for consent acting as a sort of *access control*. Researchers and health care providers could potentially access individual health records stored in off-chain databases through blockchain infrastructure, provided that they have consent from the individual.

It is interesting to note that many of the solutions described seem to consider only an ideal workflow. In an emergency situation where the patient is unable to authorize data sharing, there should be mechanisms for doctors and nurses to access data [18].

Similar to any new technology, while blockchain seems to hold the potential to improve several existing challenges in health care and give more agency to patients, the actual impact of the technology is unclear. Many solutions want to eliminate intermediaries (eg, allowing patients and providers to connect without the need of insurance companies; however, it is impossible to determine whether this kind of disruption will actually happen and whether patients are ready for such complexity [6,12,18]). Although blockchain solutions can minimize inefficiencies, they are more likely to run in conjunction with existing systems from third parties. In other words, the solutions will incorporate third parties and make the process more transparent for the participants, thereby increasing trust [12]. This would also ease the adoption of blockchain in the industry, as several stakeholders would offer less resistance.

Most of the solutions we found, with a few exceptions, seem to be at the initial development or prototyping stage: their architecture and basic functionalities are planned out, and companies are now looking to raise funds and continue their implementation. Although this shows that the health care sector is positive about the potential of blockchain, it also makes it difficult to concretely evaluate the potential of the technology.

Immutability, Decentralization, and Trust

Blockchain is designed to be distributed, transparent, and immutable by design [6,12] and can improve trust among stakeholders [6,12,18,108]. This is true in the scenarios described above. By eliminating centralized decision making, automating processes, and increasing transparency in data collection and use, blockchain can increase trust in health care processes.

However, blockchain's embedded features could prove to be a challenge if misused. For example, the fact that every node in the network maintains a copy of the ledger is redundant. Decentralized systems are less efficient, scalable, and cost-effective compared with centralized systems [12]. By design, every participant node must possess a full copy of the distributed ledger; as the number of participants increases, so does the computational requirements such as storage and energy. Blockchain has been seen as a potential solution to address issues in the environment, such as creating a marketplace for energy trading. However, some critics point out that blockchain will do more harm than good because it expends a huge amount of energy [6,99]. These factors may limit the scalability of the blockchain solutions. They are not limitations of the technology per se, but design characteristics, and developers must consider a trade-off between these and the desired transparency and security provided by the distributed ledger. Trade-offs between different blockchain features must also be considered (eg, using a public or permissioned blockchain), as explored by O'Donoghue et al [20].

Similarly, blockchain's immutability means it is *append-only* for public implementation [12]. Companies dealing with General Data Protection Regulation must abide by a key principle of regulation, which is the right to data deletion [12,19,131]. However, this is not an option for public blockchain implementations [12]. This is one of the reasons why it is not recommended to store personal data on the blockchain [7,19]. Rather, it should be stored in off-chain databases with reference to the chain. A study conducted by Park et al [132] studied the feasibility of storing, sharing, and managing records on the blockchain. The study confirmed that it was possible to manage

records in a private blockchain, but several challenges need to be addressed first, such as data size and costs.

Although blockchain provides an immutable ledger, it should also be noted that the information stored in the ledger is only as accurate as its input. This means that developers should create mechanisms to ensure that the correct data are uploaded to the blockchain [18,99].

Key and Identity Management

Information stored on the blockchain is secured through cryptographic techniques that require the use of public and private keys [18]. Private keys act as a password that allows the user to access information. For example, in the Bitcoin blockchain, user A can send Bitcoins to user B using B's public key. However, user B will only be able to access the Bitcoin with their private key. Users have one key to access all their blocks (analogous to a password to access all user data). If the key is compromised, all user data may be leaked. In the case of health care, which deals with sensitive data, this issue should not be understated [18]. For blockchain-EHR solutions, if a malicious party obtains a patient's private key, there is a significant risk of identity theft, as they will have access to the patient's full medical history [18].

One possible alternative to key management is through identity verification using blockchain [18,62,131]; it can provide an immutable ledger that stores and maintains legal documents such as birth certificates and business contracts. A person would then be able to prove their identity by accessing the blockchain [18]. Many countries, such as Estonia, use blockchain to verify various attributes of its citizens [12,131]. By being able to say if a person is an Estonian resident through blockchain, the platform can provide additional web-based services (eg, allowing citizens to vote on the web). The country already has an EHR system, accessed by a unique ID that uses blockchain to ensure integrity [10,131]. PokitDok is studying the application of a protocol called the contextually relevant identity management protocol with DokChain, in which different aspects of an identity are used for validation. For example, to buy an alcoholic a drink, the only necessary information is age. DokChain's solution would allow users to decide which specific personal information to share based on the context. To validate an identity, the technology is planning to create a consensus mechanism in which an individual's private key is partitioned among third parties. The key can be regenerated by using a subset of partitions. This is initially used for key recovery, but the goal is to have this function as implicit identity verification every time an identity needs to be validated [62]. A similar solution that distributes a user's identity attributes is being developed by the Canadian company SecureKey [131].

Conclusions

Blockchain, an immutable ledger in which all participants view an immutable version of the truth, is a promising technology that can help to minimize several challenges currently experienced in health care. This paper focused on a literature review and market assessment to determine the main challenges that could be improved by blockchain in today's health care industry. The results showed the following five challenges:

- Health records are stored between different providers in systems that lack interoperability. This makes it difficult for stakeholders to share data and, therefore, to have a complete view of a patient's health.
- Provenance and counterfeit products in drug supply chains are among the biggest hurdles faced by pharmaceutical companies today.
- There is a lack of connectivity among payers, providers, and patients in terms of insurance, which prevents access and coordination of care.
- Genomics data, one of the greatest promises of precision health, are limited by high costs, lack of massive data sets, and interoperability issues.
- Increasingly advanced methods of data collection and analysis raise concerns regarding ethics, ownership, privacy, and regulations of health data.

Although solutions vary, blockchain can provide an immutable and tamperproof log of transactions. Blockchain's features of immutability, decentralization, distribution, and transparency can optimize processes, minimize inefficiencies, and increase trust in all contexts covered in this review. However, no silver bullet to solve every need in health care exists, and companies, developers, and decision makers need to be careful when considering a blockchain solution. Although the technology has potential, it also brings new concerns about data ownership and security.

Limitations

As the aim of this review was to provide an industry point of view of the use of blockchain in health care, when defining the protocols for inclusion and exclusion, we did not include a critical component to evaluate the solutions included. Furthermore, while we searched for white papers and reports of the solutions whenever possible, we included in our results solutions already in place, in early stages of development, or recently announced, as our goal was to provide a mindset of the health care industry pertaining to blockchain. Future work should focus on critically analyzing the solutions described here for feasibility and efficiency. An interesting consideration from this analysis would be to define which challenges can be further improved with the use of blockchain. In other words, this analysis could point to where blockchain would be the most useful. In addition, blockchain can be used in conjunction with new fields of computer science, including big data and artificial intelligence [6,117]. Future work should also focus on studying blockchain in the context of these emerging technologies. Finally, it is important to note that the ontology used here, dividing the topics into five major challenges, is not exhaustive. There are many use cases where blockchain can be applied that, while not directly related to health care, may affect it (eg, key and identity management). In addition, many challenges described in the literature (eg, interoperability and data sharing of health systems) are encompassed by this ontology. In exploring these five areas, our aim was to provide an overview of the solutions being developed by the health care industry and which areas they are directing their efforts, rather than providing a defining list of use cases in health care in which blockchain can be used.

Acknowledgments

This research was supported by the Canadian Standards Association Group and MITACS internship to the corresponding author.

Authors' Contributions

PEV wrote the manuscript and conducted a literature review. PEV, PPM, LXF, and FMB contributed to the conceptualization, design, and approach of the manuscript, as well as to the interpretation of the argument made in the manuscript. All authors contributed to the writing and revision of the manuscript. All authors provided the final approval of the manuscript and agree to be accountable for this manuscript.

Conflicts of Interest

None declared.

References

1. Prosperi M, Min JS, Bian J, Modave F. Big data hurdles in precision medicine and precision public health. BMC Med Inform Decis Mak 2018 Dec 29;18(1):139. [doi: 10.1186/s12911-018-0719-2] [Medline: 30594159]
2. de Arriba-Pérez F, Caeiro-Rodríguez M, Santos-Gago J. Collection and processing of data from wrist wearable devices in heterogeneous and multiple-user scenarios. Sensors (Basel) 2016 Sep 21;16(9):1538 [FREE Full text] [doi: 10.3390/s16091538] [Medline: 27657081]
3. Mooney SJ, Pejaver V. Big data in public health: terminology, machine learning, and privacy. Annu Rev Public Health 2018 Apr 01;39(1):95-112 [FREE Full text] [doi: 10.1146/annurev-publhealth-040617-014208] [Medline: 29261408]
4. Sakr S, Elgammal A. Towards a comprehensive data analytics framework for smart healthcare services. Big Data Res 2016 Jun;4:44-58. [doi: 10.1016/j.bdr.2016.05.002]
5. Barrett MA, Humblet O, Hiatt RA, Adler NE. Big data and disease prevention: from quantified self to quantified communities. Big Data 2013 Sep;1(3):168-175. [doi: 10.1089/big.2013.0027] [Medline: 27442198]
6. Bublitz FM, Oetomo A, Sahu KS, Kuang A, Fadrique LX, Velmovitsky PE, et al. Disruptive technologies for environment and health research: an overview of artificial intelligence, blockchain, and internet of things. Int J Environ Res Public Health 2019 Oct 11;16(20):3847 [FREE Full text] [doi: 10.3390/ijerph16203847] [Medline: 31614632]
7. Azaria A, Ekblaw A, Vieira T, Lippman A. MedRec: Using blockchain for medical data access and permission management. In: Proceedings of the 2nd International Conference on Open and Big Data (OBD). 2016 Presented at: 2nd International Conference on Open and Big Data (OBD); Aug. 22-24, 2016; Vienna, Austria p. 30. [doi: 10.1109/obd.2016.11]
8. Ekblaw AC. MedRec: blockchain for medical data access, permission management and trend analysis. Massachusetts Institute of Technology. 2017. URL: https://dspace.mit.edu/bitstream/handle/1721.1/109658/987247095-MIT.pdf?sequence=1%0Ahttps://dspace.mit.edu/handle/1721.1/109658 [accessed 2021-05-22]
9. Grishin D, Obbad K, Estep P, Cifric M, Zhao Y, Church G. Blockchain-enabled genomic data sharing and analysis platform. 2018. URL: https://arep.med.harvard.edu/pdf/Grishin_Church_v4.52_2018.pdf [accessed 2021-05-22]
10. Healthcare rallies for blockchains. IBM Institute for Business Value. 2017. URL: https://www.ibm.com/downloads/cas/BBRQK3WY [accessed 2021-05-22]
11. Team Medicine: how life sciences can win with blockchain. IBM Institute for Business Value. URL: https://www.ibm.com/downloads/cas/RYD0QA7G [accessed 2021-05-22]
12. Urban MC, Pineda D. Inside the black blocks. Toronto, ON: Mowat Centre; 2018. URL: https://munkschool.utoronto.ca/mowatcentre/wp-content/uploads/publications/168_inside_the_black_blocks.pdf [accessed 2021-05-25]
13. Farouk A, Alahmadi A, Ghose S, Mashatan A. Blockchain platform for industrial healthcare: vision and future opportunities. Comp Commun 2020 Mar;154:223-235. [doi: 10.1016/j.comcom.2020.02.058]
14. Benchoufi M, Ravaud P. Blockchain technology for improving clinical research quality. Trials 2017 Jul 19;18(1):335 [FREE Full text] [doi: 10.1186/s13063-017-2035-z] [Medline: 28724395]
15. Hollander JE, Carr BG. Virtually perfect? Telemedicine for Covid-19. N Engl J Med 2020 Apr 30;382(18):1679-1681. [doi: 10.1056/nejmp2003539]
16. Dragov R, Croce CL, Hefny M. How blockchain can help in the COVID-19 crisis and recovery. IDC UK Blog. 2020. URL: https://blog-idcuk.com/blockchain-help-in-the-covid-19-and-recovery/ [accessed 2020-05-20]
17. World Economic Forum: how blockchain could help with Covid-19 supply chain disruption. Ledger Insights. 2020. URL: https://www.ledgerinsights.com/world-economic-forum-how-blockchain-could-help-with-covid-19-supply-chain-disruption/ [accessed 2020-05-25]
18. McGhin T, Choo KR, Liu CZ, He D. Blockchain in healthcare applications: research challenges and opportunities. J Net Comp Appl 2019 Jun;135:62-75. [doi: 10.1016/j.jnca.2019.02.027]
19. Vazirani AA, O'Donoghue O, Brindley D, Meinert E. Implementing blockchains for efficient health care: systematic review. J Med Internet Res 2019 Feb 12;21(2):e12439 [FREE Full text] [doi: 10.2196/12439] [Medline: 30747714]

20. O'Donoghue O, Vazirani AA, Brindley D, Meinert E. Design choices and trade-offs in health care blockchain implementations: systematic review. J Med Internet Res 2019 May 10;21(5):e12426 [FREE Full text] [doi: 10.2196/12426] [Medline: 31094344]
21. Hasselgren A, Kralevska K, Gligoroski D, Pedersen SA, Faxvaag A. Blockchain in healthcare and health sciences-a scoping review. Int J Med Inform 2020 Feb;134:104040 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.104040] [Medline: 31865055]
22. Chukwu E, Garg L. A systematic review of blockchain in healthcare: frameworks, prototypes, and implementations. IEEE Access 2020;8:21196-21214. [doi: 10.1109/access.2020.2969881]
23. Agbo C, Mahmoud Q, Eklund J. Blockchain technology in healthcare: a systematic review. Healthcare (Basel) 2019 Apr 04;7(2):56 [FREE Full text] [doi: 10.3390/healthcare7020056] [Medline: 30987333]
24. Reviews: from systematic to narrative: narrative review. University of Alabama - Birmingham. URL: https://guides.library.uab.edu/c.php?g=63689&p=409774 [accessed 2020-05-25]
25. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5(1):69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
26. Leader in blockchain news. CoinDesk. URL: https://www.coindesk.com/ [accessed 2020-05-18]
27. Cointelegraph. URL: https://cointelegraph.com/ [accessed 2020-05-18]
28. Get smarter about what matters to you. Medium. URL: https://medium.com/ [accessed 2020-05-18]
29. MedRec. URL: https://medrec.media.mit.edu/ [accessed 2018-10-03]
30. PatientTruth. Embleema - Blockchain for Real-World Evidence. URL: https://embleema.com/2020/01/16/explore-real-world-evidence/ [accessed 2018-11-08]
31. Embleema blockchain network v.2. Embleema WhitePaper. 2018. URL: https://icocube.io/uploads/Embleema.pdf [accessed 2021-05-22]
32. Adams J. CareX Whitepaper. 2018. URL: https://coins.newbium.com/post/13327-carex-blockchain-healthcare-ecosystem [accessed 2021-05-22]
33. Stoffregen E. Blockchain healthcare ecosystem in 2018. Medium. 2018. URL: https://medium.com/@erikstoffregen/blockchain-healthcare-ecosystem-d21631024454 [accessed 2018-11-22]
34. Kovach A, Ronai G. MyMEDIS: a new medical data storage and access system. 2018. URL: https://mymedis.in/documents/MEDIS-White-Paper.pdf [accessed 2021-05-22]
35. Globally decentralized medical data store and blockchain-based ecosystem. MEDIS. URL: https://mymedis.in/ [accessed 2019-03-12]
36. Allison I. Gem shows off first blockchain application for health claims. Intenational Business Times. 2017. URL: https://www.ibtimes.co.uk/gem-shows-off-first-blockchain-application-health-claims-1622574 [accessed 2018-10-09]
37. Health. Gem. URL: https://enterprise.gem.co/health/ [accessed 2018-10-04]
38. Rizzo P. Gem partners with Philips for blockchain healthcare initiative. CoinDesk. 2016. URL: https://www.coindesk.com/gem-philips-blockchain-healthcare/ [accessed 2018-10-05]
39. Redman J. Gem health unveils medical management blockchain platform. Bitcoin. 2016. URL: https://news.bitcoin.com/gem-health-blockchain-medical-mgmt/ [accessed 2018-10-09]
40. Shieber J. Gem looks to CDC and European giant Tieto to take blockchain into healthcare. TechCrunch. 2017. URL: https://techcrunch.com/2017/09/25/gem-looks-to-cdc-and-european-giant-tieto-to-take-blockchain-into-healthcare/ [accessed 2018-10-09]
41. Albeyatti A. Whitepaper. MedicalChain. 2017. URL: https://medicalchain.com/en/whitepaper/ [accessed 2021-05-22]
42. Blockchain for electronic health records. MedicalChain. URL: https://medicalchain.com/en/ [accessed 2019-02-27]
43. Rebuilding healthcare for the next generation. Citizen Health. URL: https://citizenhealth.io/ [accessed 2018-11-22]
44. Humantiv. A Citizen Health Development. URL: https://citizenhealth.io/humantiv/ [accessed 2018-11-22]
45. Medoplex. Citizen Health. URL: https://citizenhealth.io/medoplex/ [accessed 2018-11-22]
46. Siwicki B. The next big thing in pharmacy supply chain: Blockchain. Healthcare IT News. 2017. URL: https://www.healthcareitnews.com/news/next-big-thing-pharmacy-supply-chain-blockchain [accessed 2018-10-09]
47. Block Verify. URL: https://true.global/startups/blockverify/#:~:text=Blockverify%20is%20a%20blockchain%2Dbased,%2C%20apparel%2C%20electronics%20and%20pharmaceuticals [accessed 2018-10-09]
48. Hulseapple C. Block verify uses blockchains to end counterfeiting and ‘make world more honest’. Cointelegraph. 2015. URL: https://cointelegraph.com/news/block-verify-uses-blockchains-to-end-counterfeiting-and-make-world-more-honest [accessed 2018-10-09]
49. De N. Pharma giant merck eyes blockchain for fighting counterfeit meds. CoinDesk. 2018. URL: https://www.coindesk.com/merck-proposes-blockchain-platform-for-combat-counterfeiters/ [accessed 2018-11-07]
50. Haring B. BlockTribune. URL: https://blocktribune.com/blockchain-patent-filed-by-pharmaceutical-giant-merck-co/ [accessed 2018-11-07]
51. Modum. URL: https://modum.io/ [accessed 2018-10-09]
52. Products. Modum. URL: https://www.modum.io/solutions [accessed 2018-10-09]

53. Data integrity for supply chain operations, powered by blockchain technology. Modum. 2017. URL: https://assets.modum.io/wp-content/uploads/2017/08/modum-whitepaper-v.-1.0.pdf [accessed 2021-05-22]
54. Uhlmann S. Reducing counterfeit products with blockchains. University of Zurich. 2017. URL: https://www.merlin.uzh.ch/contributionDocument/download/10024 [accessed 2021-05-22]
55. Stanley A. Ready to rumble: IBM launches food trust blockchain for commercial use. Forbes. 2018. URL: https://www.forbes.com/sites/astanley/2018/10/08/ready-to-rumble-ibm-launches-food-trust-blockchain-for-commercial-use/#68bf18817439 [accessed 2018-10-18]
56. IBM food trust expands blockchain network to foster a safer, more transparent and efficient global food system. IBM News Room. URL: https://newsroom.ibm.com/2018-10-08-IBM-Food-Trust-Expands-Blockchain-Network-to-Foster-a-Safer-More-Transparent-and-Efficient-Global-Food-System-1 [accessed 2018-10-18]
57. IBM Food Trust: a new era for the world's food supply. 2018. URL: https://www.ibm.com/blockchain/solutions/food-trust [accessed 2021-05-22]
58. Walmart's food safety solution using IBM food trust built on the IBM blockchain platform. IBM Blockchain. 2017. URL: https://www.youtube.com/watch?time_continue=173&v=SV0KXBxSoio [accessed 2018-09-26]
59. Huillet M. Alibaba's Ant Financial to launch blockchain backend-as-a-service platform. Cointelegraph. 2018. URL: https://cointelegraph.com/news/alibabas-ant-financial-to-launch-blockchain-backend-as-a-service-platform [accessed 2018-10-19]
60. Zhao W. Ant Financial is launching a blockchain app to tackle food fraud. CoinDesk. 2018. URL: https://www.coindesk.com/ant-financial-is-launching-a-blockchain-app-to-tackle-food-fraud/ [accessed 2018-09-26]
61. DokChain. PokitDok. URL: https://pokitdok.com/dokchain/ [accessed 2018-09-26]
62. Smith WB. DokChain: intelligent automation in healthcare transaction processing. PokitDok. 2018. URL: https://pokitdok.com/wp-content/uploads/2018/02/DokChain_Whitepaper.pdf [accessed 2021-05-22]
63. Brennan B. DokChain by PokitDok – blockchain for healthcare. Blockchain Healthcare Review. URL: https://blockchainhealthcarereview.com/dokchain-by-pokitdoc-blockchain-for-healthcare/ [accessed 2018-10-01]
64. Healthcare claims processing software. PokitDok. URL: https://pokitdok.com/business/claims-management/ [accessed 2018-10-03]
65. Autonomous Auto-Adjudication 101: blockchains in healthcare. Pokitdok. URL: https://techcrunch.com/2017/05/10/pokitdok-teams-with-intel-on-healthcare-blockchain-solution/ [accessed 2018-10-03]
66. Healthcare propensity to pay. PokitDok. URL: https://pokitdok.com/business/payment-risk/ [accessed 2018-10-02]
67. Patient access solutions. PokitDok. URL: https://pokitdok.com/business/patient-access-solutions/ [accessed 2018-10-03]
68. Healthcare reimbursement solutions - hospital payment systems. Payspan. URL: https://payspan.com/ [accessed 2018-11-08]
69. How blockchain can connect payers, providers and consumers. Payspan. URL: https://payspan.com/wp-content/uploads/2018/02/Payspan-white-paper-February-2018.pdf [accessed 2021-05-22]
70. Nebula Genomics. URL: https://www.nebula.org/ [accessed 2018-11-12]
71. LunaDNA : frequently asked questions. URL: https://support.lunadna.com/support/solutions/articles/43000038763-what-is-lunadna- [accessed 2018-11-12]
72. About LunaDNA - learn more about the LunaDNA team company. URL: https://www.lunadna.com/what-we-do/ [accessed 2018-11-12]
73. Interview with Luna DNA's Co-Founder and President Dawn Barry. SanDiegOmics. 2018. URL: https://sandiegomics.com/luna-dna-interview-with-co-founder-dawn-barry/ [accessed 2018-11-12]
74. Bigelow B. Luna DNA uses blockchain to share genomic data as a "Public Benefit". Xconomy. 2018. URL: https://xconomy.com/san-diego/2018/01/22/luna-dna-uses-blockchain-to-share-genomic-data-as-a-public-benefit/ [accessed 2018-11-12]
75. Farr C, Levy A. Luna Coin project: sell your genetic data for crypto tokens. CNBC Tech. 2017. URL: https://www.cnbc.com/2017/12/18/luna-coin-project-sell-your-genetic-data-for-crypto-tokens.html [accessed 2018-11-12]
76. Project SHIVOM (Official Video) - powering the next era of genomics through blockchain. Shivom. 2018. URL: https://www.youtube.com/watch?v=jce9vB5zbps [accessed 2018-11-14]
77. Shivom. URL: https://shivom.io/ [accessed 2018-11-14]
78. Thrill W. Shivom: the uncanny synergy of blockchain and genomics. Hacker Noon. 2018. URL: https://hackernoon.com/shivom-the-uncanny-synergy-of-blockchain-and-genomics-e1ca7f2a0173 [accessed 2018-11-15]
79. Shivom innovation council. Shivom. URL: https://www.youtube.com/watch?v=NuaOV82kCjc [accessed 2018-11-14]
80. Zenome. URL: https://zenome.io/ [accessed 2018-11-15]
81. Kulemin N, Popov S, Gorbachev A. The Zenome Project: whitepaper blockchain-based genomic ecosystem. Zenome.io 2017:A. [doi: 10.13140/RG.2.2.25865.13925]
82. Thrill W. EncrypGen uses blockchain technology to store and manage DNA profiles. Hacker Noon. URL: https://hackernoon.com/encrypgen-uses-blockchain-technology-to-store-and-manage-dna-profiles-a920e898b6a8 [accessed 2018-11-15]
83. Gene-chain DNA data marketplace. EncrypGen. URL: https://encrypgen.com/encrypgen-gene-chain-dna-data-marketplace/ [accessed 2018-11-15]
84. Marketplace partners. EncrypGen. URL: https://encrypgen.com/marketplace-partners/ [accessed 2018-11-15]
85. The DNA data marketplace. EncrypGen. URL: https://encrypgen.com/ [accessed 2018-11-15]

86. S Korea's Macrogen to leverage blockchain for genomic data. Cryptovest. 2018. URL: https://www.investing.com/news/cryptocurrency-news/s-koreas-macrogen-to-leverage-blockchain-for-genomic-data-1562879 [accessed 2018-10-10]
87. Say N. Macrogen develops blockchain platform to share genetic data. Blockonomi. 2018. URL: https://blockonomi.com/macrogen-blockchain/ [accessed 2018-10-10]
88. Ji-young S. Korea's Macrogen, Bigster to create blockchain-based medical data platform. The Korea Herald. 2018. URL: http://www.koreaherald.com/view.php?ud=20180806000646 [accessed 2018-10-10]
89. Hu-manity.co collaborates with IBM blockchain on consumer app to manage personal data property rights. IBM Announcements. 2018. URL: https://newsroom.ibm.com/2018-09-06-Hu-manity-co-Collaborates-with-IBM-Blockchain-on-Consumer-App-to-Manage-Personal-Data-Property-Rights [accessed 2020-05-15]
90. Takahashi D. Hu-manity.co uses IBM blockchain to give you the right to control your personal data. VentureBeat. URL: https://venturebeat.com/2018/09/06/hu-manity-co-uses-ibm-blockchain-to-give-you-the-right-to-control-your-personal-data/ [accessed 2020-05-15]
91. Alexandre A. New Bitfury joint project to manage medical data permissions with blockchain tech. Cointelegraph. 2019. URL: https://cointelegraph.com/news/new-bitfury-joint-project-to-manage-medical-data-permissions-with-blockchain-tech [accessed 2020-05-15]
92. Bitfury announces blockchain-based consent management system; partners with Hancom to distribute Crystal platform. TokenPost. 2019. URL: https://tokenpost.com/Bitfury-announces-blockchain-based-consent-management-system-partners-with-Hancom-to-distribute-Crystal-platform-1603 [accessed 2020-05-15]
93. HealthVerity consent. URL: https://healthverity.com/solutions/healthverity-consent/ [accessed 2020-05-21]
94. Hern A. Google's DeepMind plans bitcoin-style health record tracking for hospitals. The Guardian. 2017. URL: https://www.theguardian.com/technology/2017/mar/09/google-deepmind-health-records-tracking-blockchain-nhs-hospitals [accessed 2018-09-28]
95. Suleyman M, Laurie B. Trust, confidence and verifiable data audit. DeepMind. 2017. URL: https://deepmind.com/blog/trust-confidence-verifiable-data-audit/ [accessed 2021-05-22]
96. DeepMind. URL: https://deepmind.com/ [accessed 2021-05-22]
97. Metz C. Google DeepMind's untrendy play to make the blockchain actually useful. Wired. 2017. URL: https://www.wired.com/2017/03/google-deepminds-untrendy-blockchain-play-make-actually-useful/ [accessed 2018-09-28]
98. Powles J, Hodson H. Google DeepMind and healthcare in an age of algorithms. Health Technol (Berl) 2017 Mar 16;7(4):351-367 [FREE Full text] [doi: 10.1007/s12553-017-0179-1] [Medline: 29308344]
99. Maslove DM, Klein J, Brohman K, Martin P. Using blockchain technology to manage clinical trials data: a proof-of-concept study. JMIR Med Inform 2018 Dec 21;6(4):e11949 [FREE Full text] [doi: 10.2196/11949] [Medline: 30578196]
100. Wiljer D, Brudnicki S. Bringing blockchain to healthcare for a new view on data. IBM Think Blog. 2019. URL: https://www.ibm.com/blogs/think/2019/08/bringing-blockchain-to-healthcare-for-a-new-view-on-data/ [accessed 2020-05-25]
101. Canadian hospital collaborates with IBM for health consent blockchain. Ledger Insights. 2020. URL: https://www.ledgerinsights.com/health-consent-blockchain-university-health-network-uhn/ [accessed 2020-05-25]
102. Smith K. Clinician engagement, local impact awards, budget risk meetings and beyond. University Health Network. 2019. URL: https://www.uhn.ca/corporate/AboutUHN/Updates_from_CEO/Pages/Clinician_engagement_Local_Impact_Awards_budget_risk_meetings_and_beyond.aspx [accessed 2020-05-25]
103. Velmovitsky P, Morita P. Blockchain platform for consent management in ambient assisted living. AAL Forum Poster Presentations. 2019. URL: https://www.aalforum.eu/about/poster-presentations-aal-forum-2019/ [accessed 2020-05-26]
104. Guegan D. Public blockchain versus private blockhain. HAL Archives-Ouvertes. 2017. URL: https://halshs.archives-ouvertes.fr/halshs-01524440/document [accessed 2021-05-22]
105. Dias JP, Ferreira HS, Martins A. A blockchain-based scheme for access control in e-health scenarios. In: Proceedings of the Tenth International Conference on Soft Computing and Pattern Recognition (SoCPaR 2018). Switzerland: Springer; 2018:238-247.
106. Abouelmehdi K, Beni-Hssane A, Khaloufi H, Saadi M. Big data security and privacy in healthcare: a review. Procedia Comp Sci 2017;113:73-80. [doi: 10.1016/j.procs.2017.08.292]
107. Blockchain's healthcare technology impact. Cognizant. URL: https://soundcloud.com/cognizant-worldwide/impact-of-blockchain-technology-on-healthcare-codex2937 [accessed 2018-11-09]
108. Gorenflo C, Golab L, Keshav S. Mitigating trust issues in electric vehicle charging using a blockchain. In: Proceedings of the Tenth ACM International Conference on Future Energy Systems. 2019 Presented at: e-Energy '19: The Tenth ACM International Conference on Future Energy Systems; June, 2019; Phoenix AZ USA p. 160-164. [doi: 10.1145/3307772.3328283]
109. Electronic health records. Canada Health Infoway. URL: https://www.infoway-inforoute.ca/en/solutions/digital-health-foundation/electronic-health-records [accessed 2018-09-27]
110. Sharma R. Blockchain: the magic pill to alleviate the pain points of the healthcare industry? France Canada Chamber of Commerce Ontario. 2018. URL: https://www.fccco.org/post/blockchain-in-healthcare [accessed 2021-05-22]

111. Velmovitsky PE, Miranda PA, Fadrique LX, Morita PP. Blockchain in health care. CSA Group Report. 2021. URL: https://www.csagroup.org/article/research/blockchain-in-health-care/ [accessed 2021-04-06]
112. Stanley A. Better off abroad? Blockchain health firms gain ground outside the US. CoinDesk. URL: https://www.coindesk.com/better-off-abroad-blockchain-health-firms-are-gaining-ground-outside-the-us/ [accessed 2018-10-09]
113. Tieto establishes a blockchain pilot program in the Nordics – introduces a global identity network for secure digital interactions. Tietoevry Newsroom. URL: https://www.tietoevry.com/en/newsroom/all-news-and-releases/press-releases/2017/11/tieto-establishes-a-blockchain-pilot-program-in-the-nordics--introduces-a-global-identity-network-for-secure-digital-in/ [accessed 2018-10-09]
114. ICO alert report: BlockRx. URL: https://blog.icoalert.com/ico-alert-report-blockrx [accessed 2018-11-07]
115. Drug Supply Chain Security Act (DSCSA). US Food & Drug Administration. URL: https://www.fda.gov/drugs/drug-supply-chain-integrity/drug-supply-chain-security-act-dscsa [accessed 2020-02-01]
116. The Drug Supply Chain Security Act and Blockchain : a white paper for stakeholders in the pharmaceutical supply chain. Center for Supply Chain Studies. 2018. URL: https://static1.squarespace.com/static/563240cae4b056714fc21c26/t/5b3426b088251b230ba9e6e5/1530144436146/C4SCS+White+Paper_+DSCSA+and+Blockchain+Study_FINAL3.pdf [accessed 2021-05-22]
117. Building block(chain)s for a better planet. PwC Global. 2018. URL: https://www.pwc.com/gx/en/services/sustainability/building-blockchains-for-the-earth.html [accessed 2021-05-22]
118. What is a gene? MedlinePlus. URL: https://ghr.nlm.nih.gov/primer/basics/gene [accessed 2020-05-12]
119. Ferguson GK. Primer. The Human Genome: Poems on the Book of Life. URL: http://www.thehumangenome.co.uk/THE_HUMAN_GENOME/Primer.html [accessed 2020-05-12]
120. Barry D. There is nothing more personal than your genome. Tedx Talks. 2016. URL: https://www.youtube.com/watch?v=M3SLHhWYxiY [accessed 2018-11-12]
121. Nebula genomics - platform overview. Vimeo. URL: https://vimeo.com/287152656 [accessed 2018-11-12]
122. Adams SA, Petersen C. Precision medicine: opportunities, possibilities, and challenges for patients and providers. J Am Med Inform Assoc 2016 Jul;23(4):787-790. [doi: 10.1093/jamia/ocv215] [Medline: 26977101]
123. Khoury MJ. The shift from personalized medicine to precision medicine and precision public health: words matter!. Centers for Disease Control and Prevention. 2016. URL: https://blogs.cdc.gov/genomics/2016/04/21/shift/ [accessed 2019-08-08]
124. Benke K, Benke G. Artificial Intelligence and Big Data in Public Health. Int J Environ Res Public Health 2018 Dec 10;15(12):2796 [FREE Full text] [doi: 10.3390/ijerph15122796] [Medline: 30544648]
125. Barney JR, Antisdel M. Common problems in informed consent. Yale University. 2013. URL: https://your.yale.edu/policies-procedures/other/common-problems-informed-consent [accessed 2021-05-19]
126. Novitzky P, Smeaton AF, Chen C, Irving K, Jacquemard T, O'Brolcháin F, et al. A review of contemporary work on the ethics of ambient assisted living technologies for people with dementia. Sci Eng Ethics 2015 Jun 19;21(3):707-765. [doi: 10.1007/s11948-014-9552-x] [Medline: 24942810]
127. Gupta U. Informed consent in clinical research: revisiting few concepts and areas. Perspect Clin Res 2013 Jan;4(1):26-32 [FREE Full text] [doi: 10.4103/2229-3485.106373] [Medline: 23533976]
128. Morgan-Linnell SK, Stewart DJ, Kurzrock R. U.S. Food and Drug Administration Inspections of Clinical Investigators: overview of results from 1977 to 2009. Clin Cancer Res 2014 Apr 15;20(13):3364-3370. [doi: 10.1158/1078-0432.ccr-13-3206]
129. Velmovitsky PE, Miranda PA, Vaillancourt H, Donovska T, Teague J, Morita PP. A blockchain-based consent platform for active assisted living: modeling study and conceptual framework. J Med Internet Res 2020 Dec 04;22(12):e20832 [FREE Full text] [doi: 10.2196/20832] [Medline: 33275111]
130. Art. 17 GDPR : right to erasure ('right to be forgotten'). General Data Protection Regulation (GDPR). URL: https://gdpr-info.eu/art-17-gdpr/ [accessed 2021-04-07]
131. Compert C, Luinetti M, Portier B. Blockchain and GDPR: how blockchain could address five areas associated with GDPR compliance. IBM Security. 2018. URL: https://iapp.org/media/pdf/resource_center/blockchain_and_gdpr.pdf [accessed 2021-05-15]
132. Park YR, Lee E, Na W, Park S, Lee Y, Lee J. Is blockchain technology suitable for managing personal health records? Mixed-methods study to test feasibility. J Med Internet Res 2019 Feb 08;21(2):e12533 [FREE Full text] [doi: 10.2196/12533] [Medline: 30735142]

Abbreviations

DSCSA: Drug Supply Chain Security Act
EHR: electronic health record
RQ: research question

Edited by C Lovis; submitted 26.05.20; peer-reviewed by M Raghavendra, A Hasselgren; comments to author 23.06.20; revised version received 21.07.20; accepted 24.04.21; published 08.06.21.

Please cite as:
Velmovitsky PE, Bublitz FM, Fadrique LX, Morita PP
Blockchain Applications in Health Care and Public Health: Increased Transparency
JMIR Med Inform 2021;9(6):e20713
URL: https://medinform.jmir.org/2021/6/e20713
doi:10.2196/20713
PMID:34100768

Review

Hyperpolarized Magnetic Resonance and Artificial Intelligence: Frontiers of Imaging in Pancreatic Cancer

José S Enriquez[1,2*], MS; Yan Chu[3*], MS; Shivanand Pudakalakatti[1], DPhil; Kang Lin Hsieh[3], DPhil; Duncan Salmon[4]; Prasanta Dutta[1], DPhil; Niki Zacharias Millward[2,5], DPhil; Eugene Lurie[6], DPhil; Steven Millward[1,2], DPhil; Florencia McAllister[2,7], MD; Anirban Maitra[2,8], MD; Subrata Sen[2,6], DPhil; Ann Killary[2,6], DPhil; Jian Zhang[9], DPhil; Xiaoqian Jiang[3], DPhil; Pratip K Bhattacharya[1,2], DPhil; Shayan Shams[3], DPhil

[1]Department of Cancer Systems Imaging, University of Texas MD Anderson Cancer Center, Houston, TX, United States
[2]Graduate School of Biomedical Sciences, University of Texas MD Anderson Cancer Center, Houston, TX, United States
[3]School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX, United States
[4]Department of Electrical and Computer Engineering, Rice University, Houston, TX, United States
[5]Department of Urology, University of Texas MD Anderson Cancer Center, Houston, TX, United States
[6]Department of Translational Molecular Pathology, University of Texas MD Anderson Cancer Center, Houston, TX, United States
[7]Department of Clinical Cancer Prevention, University of Texas MD Anderson Cancer Center, Houston, TX, United States
[8]Department of Pathology, University of Texas MD Anderson Cancer Center, Houston, TX, United States
[9]Division of Computer Science and Engineering, Louisiana State University, Baton Rouge, LA, United States
[*]these authors contributed equally

Corresponding Author:
Shayan Shams, DPhil
School of Biomedical Informatics
University of Texas Health Science Center at Houston
7000 Fannin St
Houston, TX, 77030
United States
Phone: 1 713 500 3940
Fax: 1 713 500 3765
Email: shayan.shams@uth.tmc.edu

Abstract

Background: There is an unmet need for noninvasive imaging markers that can help identify the aggressive subtype(s) of pancreatic ductal adenocarcinoma (PDAC) at diagnosis and at an earlier time point, and evaluate the efficacy of therapy prior to tumor reduction. In the past few years, there have been two major developments with potential for a significant impact in establishing imaging biomarkers for PDAC and pancreatic cancer premalignancy: (1) hyperpolarized metabolic (HP)-magnetic resonance (MR), which increases the sensitivity of conventional MR by over 10,000-fold, enabling real-time metabolic measurements; and (2) applications of artificial intelligence (AI).

Objective: Our objective of this review was to discuss these two exciting but independent developments (HP-MR and AI) in the realm of PDAC imaging and detection from the available literature to date.

Methods: A systematic review following the PRISMA extension for Scoping Reviews (PRISMA-ScR) guidelines was performed. Studies addressing the utilization of HP-MR and/or AI for early detection, assessment of aggressiveness, and interrogating the early efficacy of therapy in patients with PDAC cited in recent clinical guidelines were extracted from the PubMed and Google Scholar databases. The studies were reviewed following predefined exclusion and inclusion criteria, and grouped based on the utilization of HP-MR and/or AI in PDAC diagnosis.

Results: Part of the goal of this review was to highlight the knowledge gap of early detection in pancreatic cancer by any imaging modality, and to emphasize how AI and HP-MR can address this critical gap. We reviewed every paper published on HP-MR applications in PDAC, including six preclinical studies and one clinical trial. We also reviewed several HP-MR–related articles describing new probes with many functional applications in PDAC. On the AI side, we reviewed all existing papers that met our inclusion criteria on AI applications for evaluating computed tomography (CT) and MR images in PDAC. With the emergence of AI and its unique capability to learn across multimodal data, along with sensitive metabolic imaging using HP-MR, this

knowledge gap in PDAC can be adequately addressed. CT is an accessible and widespread imaging modality worldwide as it is affordable; because of this reason alone, most of the data discussed are based on CT imaging datasets. Although there were relatively few MR-related papers included in this review, we believe that with rapid adoption of MR imaging and HP-MR, more clinical data on pancreatic cancer imaging will be available in the near future.

Conclusions: Integration of AI, HP-MR, and multimodal imaging information in pancreatic cancer may lead to the development of real-time biomarkers of early detection, assessing aggressiveness, and interrogating early efficacy of therapy in PDAC.

(JMIR Med Inform 2021;9(6):e26601) doi:10.2196/26601

KEYWORDS

artificial intelligence; deep learning; hyperpolarization; metabolic imaging; MRI; 13C; HP-MR; pancreatic ductal adenocarcinoma; pancreatic cancer; early detection; assessment of treatment response; probes; cancer; marker; imaging; treatment; review; detection; efficacy

Introduction

There is an unmet need for noninvasive surrogate markers that can help to identify the aggressive subtype(s) in a pancreatic lesion at an early time point [1]. In contrast to the declines in cancer-related deaths from other malignancies, progress in the management of pancreatic ductal adenocarcinoma (PDAC) has been slow, and the incidence of cancer-related deaths due to PDAC continues to rise [2]. PDAC develops relatively symptom-free, and is one of the leading causes of cancer-related deaths in the United States. In 2020 alone, it was estimated that approximately 57,600 people (30,400 men and 27,200 women) would be diagnosed with PDAC, and approximately 47,050 people (24,640 men and 22,410 women) were projected to die of the disease [3]. Early detection of PDAC is unusual and typically incidental, with the majority (~85%) presenting with locally advanced or metastatic disease when surgery, the only curative modality, is not an option. Overall, PDAC is associated with a dire prognosis and a 5-year survival rate of only 8% [3]. The absence of early symptoms and lack of a reliable screening test have created a critical need for identifying and developing new noninvasive biomarkers for the early detection of PDAC [1].

Hyperpolarization (HP)-based magnetic resonance (MR) has become a major new imaging modality by providing valuable information on previously inaccessible aspects of biological processes owing to its ability for detecting endogenous, nontoxic ^{13}C-labeled probes that can monitor enzymatic conversions through key biochemical pathways [4-6]. Clinical trials with this modality are ongoing at several centers worldwide [7]. HP-MR provides an exciting opportunity to identify and understand early metabolic aberrations, enabling the detection of advanced pancreatic preneoplastic lesions and PDAC at the smallest size for which no methods of detection currently exist. In general, cancer, and PDAC in particular, is considered a paradigm of genetically defined metabolic abnormalities. Genetic mutations can trigger specific signaling pathways that are associated with metabolic transformations, which can potentially be detected by HP methods with a high degree of sensitivity.

In conventional MR, the signal measured is generated from the abundance of hydrogen in the body, specifically water [8]. Organic molecules at high concentration in the body with a high abundance of hydrogens such as choline, lipids, and lactate can also be measured using MR. Other nuclei such ^{13}C and ^{15}N can also be measured using MR, but their utility in living systems is low due to their low abundance in nature (the natural abundance of ^{13}C is 1%) and their smaller gyromagnetic ratio compared to that of hydrogen [9]. HP enables these nuclei to be observed in vivo.

HP allows for >10,000-fold sensitivity enhancement relative to conventional MR, and is a nontoxic, nonradioactive method for assessing tissue metabolism and other physiological properties [10-13]. There are four established methods for producing HP probes: (i) dynamic nuclear polarization (DNP) [10,13], (ii) optical pumping of noble gases [14], (iii) the brute force approach [15], and (iv) parahydrogen-induced polarization [16]. The detailed physics of these HP methods can be found elsewhere [17]. The most common and widely used method for HP is DNP, in which magnetization is transferred from the unpaired electrons (usually from added radicals) to the isotopically labeled probe [17]. This transfer of magnetization occurs under microwave irradiation at a low temperature of 1.5 K and a high magnetic field of 3 T. Development of the dissolution DNP technique in 2003 [4] opened a new avenue to monitor in vivo metabolism, enabling the detection and tracking of the fate of metabolites containing low-abundance nuclei such as ^{13}C [18]. The routine dissolution DNP instrument employed, which carries out HP in the preclinical setting, is HyperSense (Oxford Instruments, UK), as shown in Figure S1 in Multimedia Appendix 1. A clinical polarizer is available for performing real-time metabolic profiling in humans (SPINLab, GE Healthcare) and over 20 such polarizers have been installed worldwide [19].

The most commonly used HP probes to track the pathways of interest are ^{13}C-enriched probes, which are either uniformly or selectively enriched. The other reason to employ ^{13}C-enriched molecules is the comparatively longer longitudinal relaxation time (T_1) of the ^{13}C nucleus compared to that of other nuclei. The high ^{13}C signal of HP probes and the fact that an HP signal is carried over in the products of biochemical transformation allow investigators to interrogate biochemical reactions in real time. These probes are usually part of essential biochemical reactions such as glycolysis (glucose and pyruvate) and the

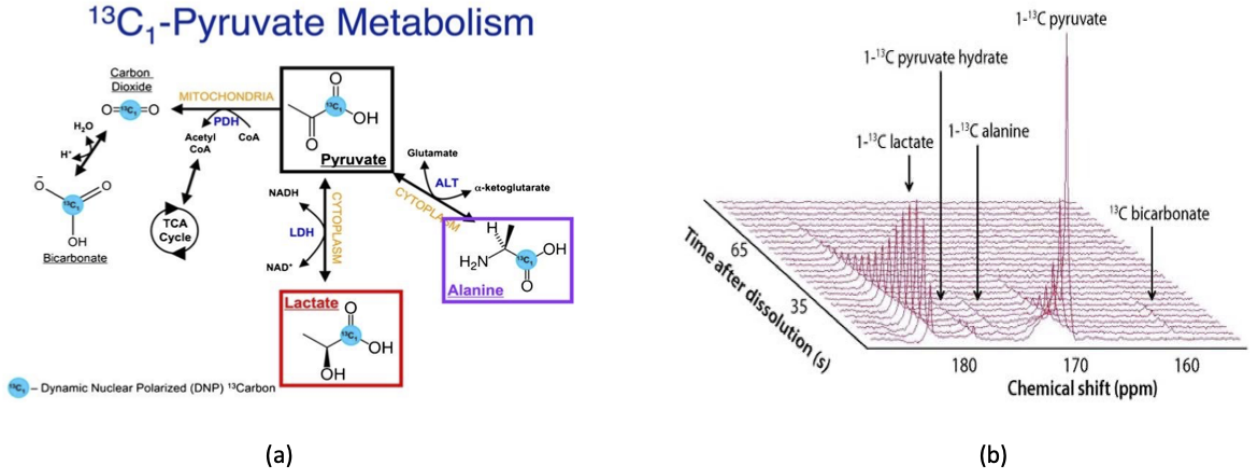
tricarboxylic acid (TCA) cycle (succinate, fumarate, and glutamine).

HP-MR experiments have been performed mostly in preclinical models to date, and HP-MR is not currently routinely used in clinical settings. However, several clinical trials have been performed or are ongoing [5]. HP-MR in the preclinical setting involves injecting the HP probe dissolved in a biocompatible solvent into the tail vein of rodents. The probe diffuses through the blood to populate in well-perfused body tissues. After entering the extracellular fluid, the molecule is taken up into the cells with the help of endogenous transporters. All of these processes must occur before the HP signal decays, which is determined by the decay time (ie, T_1) of the HP probe. For most probes, T_1 ranges from 15-20 seconds to approximately 1 minute. Hence, it is important to dissolve the probe in the solvent immediately and inject into the animals quickly to avoid loss of the HP signal due to relaxation. A specially designed proton volume coil and ^{13}C surface coil are used to receive the signal from the enriched HP ^{13}C probe in vivo.

The utility of HP-MR is not only simply tracking the probe diffusing inside the body but also its ability to visualize downstream metabolic products of injected probes converted by endogenous enzymes [5]. HP-MR can be used to quantify in vivo metabolic flux in real time. However, all processes must be completed within the time frame of T_1 of the HP probe. Therefore, only relatively fast biochemical reactions can be visualized.

Glycolysis (the breakdown of glucose) is a multistep process that eventually yields pyruvate in the cytosol. Pyruvate is the final breakdown product of glucose in glycolysis and is preferably converted to lactate. The high dependence of cancer cells on glucose and glycolysis is often referred to as the Warburg effect after the initial discovery of this dependence by Dr. Otto Warburg [20]. Therefore, HP [1-^{13}C]-pyruvate is the most common HP probe for determining glycolytic flux in cancer. Another key point is that pyruvate is taken up rapidly by monocarboxylate transporters [21]. In the cytosol, the HP pyruvate has four important fates [22]: (i) conversion to lactate; (ii) conversion to alanine; (iii) transport into the mitochondria and conversion to carbon dioxide; and (iv) conversion to acetyl-coenzyme A to be utilized in the TCA cycle, which can be tracked by labeling the first carbon of pyruvate (Figure 1a). When HP-pyruvate is injected into an animal, the signal is recorded from an anatomical imaging slice placed in the tissue of interest. An example of a metabolic HP-MR spectrum is shown in Figure 1b. The flux from pyruvate to a downstream metabolite can be visualized and evaluated using either TopSpin (Bruker BioSpin GmbH) or MestReNova (Mestrelab Research) in either of the two following ways: by measuring the ratio of signals integrated over time (eg, lactate-to-pyruvate ratio, alanine-to-lactate ratio) [23] or by calculating the Kp value (according to the Bolch equation): K_{PL} (pyruvate to lactate) and K_{PA} (pyruvate to alanine) [24].

Figure 1. (a) Schematic showing pyruvate metabolism inside a cell. The [1-13C] pyruvate can be converted to 13C-lactate, 13C-alanine, and 13C-bicarbonate in the presence of enzymes lactate dehydrogenase-A (LDHA), alanine transferase (ALT), and pyruvate decarboxylase, respectively. (b) Downstream products of pyruvate metabolism such as lactate and alanine can be imaged using hyperpolarized magnetic resonance. A 3D, real-time readout of the signals, as shown here, can be created using standard software such as Chenomx.



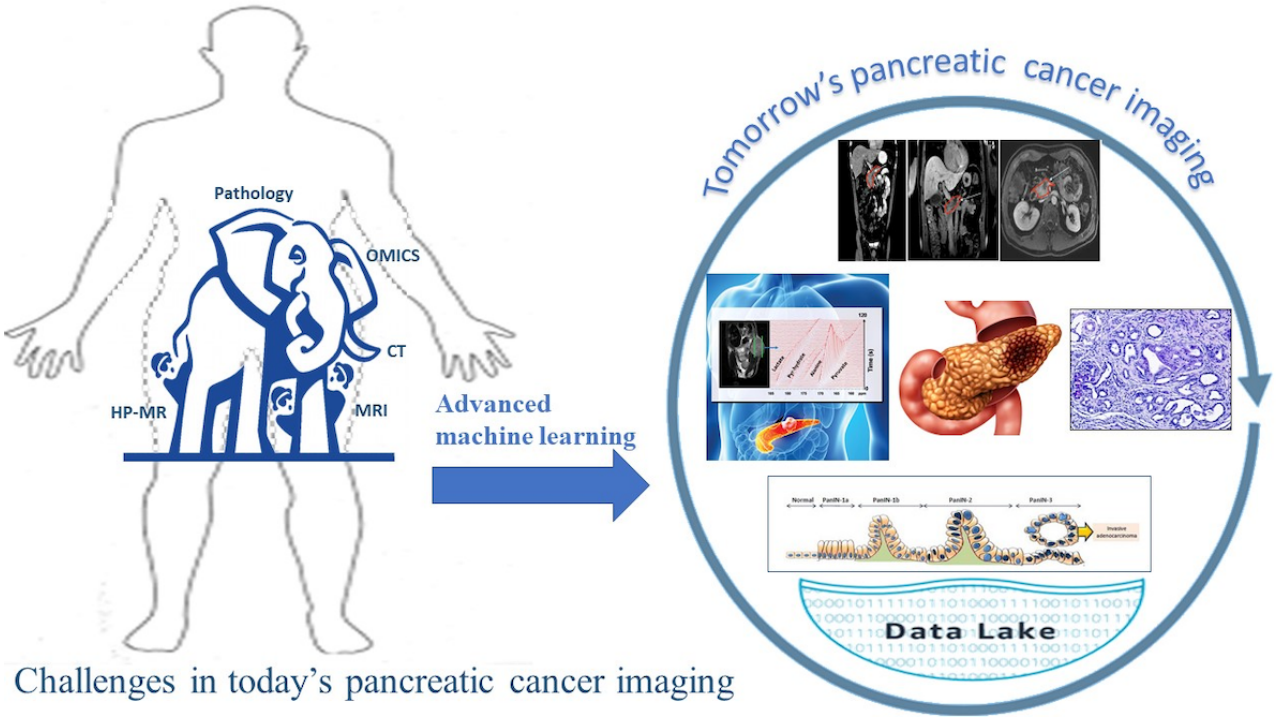
In summary, HP-MR provides a unique opportunity to measure real-time metabolic signals arising in the tissue of interest with over 10,000-fold sensitivity enhancement that cannot be interrogated by other imaging techniques. The provided outcome is the spectroscopic signatures of the metabolites of interest that are recorded as resonances at different and unique chemical shifts (Figure 1b). The HP-pyruvate signal undergoes decay once it is hyperpolarized with a characteristic decay constant (T_1 ~50 seconds) as well as the downstream products of the metabolism (eg, alanine and lactate). Overall, there is a time window of $3\times T_1$ (~150 seconds) to accomplish this real-time metabolic imaging, and this short time frame is a major limitation of HP-MR. Fast MR sequence design along with powerful and rapid imaging gradients can help in acquiring more sensitive and informative spectra in the future to mitigate this limitation. Several MR imaging (MRI) companies such as GE Healthcare, Siemens, and Bruker have devoted considerable research investment on this matter.

Artificial intelligence (AI) is a fast-developing research field in which machines are utilized to learn from observations to mimic human intelligence. Kaplan et al [25] define AI as a system's ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation. Over the last decade, deep learning has dramatically reshaped AI research. With the development of deep learning, a subfield of AI, and recognition of its potential in feature extraction and flexibility, it has increasingly been applied to numerous medical scenarios such as diagnosis, health care delivery optimization, genomics, and drug discovery [26-31]. Machine learning has been utilized for online health care management [32], disease prevention [33], clinical note processing [34], and management of chronic diseases [35]. AI has been leveraged for diagnosis and localization of regions of interest (ROIs) using a vast array of medical images such as optical images, MRI, X-rays, and computed tomography (CT) [36-41]. As a result, there is a great opportunity to utilize AI for the early detection of cancer such as PDAC.

Deep-learning algorithms rely on neural networks, which mimic the process of information transformation by neurons in the biological brain [42]. Neural networks adaptively learn features from observations during training and translate the input data to high-dimensional representations suitable for classification or regression tasks. The success of deep-learning algorithms is rooted in their multiple stacked layers and efficient feature extraction, often explained as a powerful representation learning method. Each layer consists of multiple neurons transforming the information nonlinearly by an activation function. This architecture allows for high-level interactions between transformed features coming from the previous layers to contribute to the output. Hence, deep-learning algorithms could automatically optimize the parameters and learn a high-level representation of input data aligned with the target task.

As shown in Figure 2, we believe that the knowledge gap of "early diagnosis of pancreatic cancer with noninvasive imaging" is an elephant in the dark that cannot be accomplished with a single modality. Pancreatic cancer at the very early stages is completely asymptomatic. Conventional anatomical imaging cannot detect any of these early stages of premalignancy of this deadly disease when therapeutic or early surgical interventions can be most effective. Conventional MRI can detect intraductal papillary mucinous neoplasms (IPMNs) where epithelial pancreatic cystic tumors of mucin-producing cells arise from the pancreatic ducts [43]. Although IPMNs are benign tumors, they can progress to pancreatic cancer in some cases [43]. However, MRI as well other imaging modalities fail to detect any other premalignant lesions such as pancreatic intraepithelial neoplasia (PanIN), which is a more commonly accepted mechanistic pathway of the tumorigenesis of PDAC [44]. It is important to recognize that an individual with even stage I (localized) pancreatic cancer has a 5-year survival rate of only 39% [45]. This emphasizes the point that early detection in pancreatic cancer must occur at stages earlier than clinical stage I.

Figure 2. Cartoon showing the challenges of imaging pancreatic cancer at early stages and how artificial intelligence can interface with hyperpolarized magnetic resonance (HP-MR), anatomical magnetic resonance imaging (MRI), and pathology data toward developing biomarkers of pancreatic cancer premalignancy. This approach may become the standard of care in the clinic of the future. CT: computed tomography.



HP-MR can detect metabolic changes at very early stages of lesion formation in the pancreas; however, this is more of an MR spectroscopic technique than an MRI modality. Moreover, the signal from HP compounds lasts no more than a few minutes that allow for a rapid acquisition of dynamic metabolic flux measurements in the organ of interest. This review will focus on the introduction of AI approaches to CT and MRI datasets, and the applications of HP-MR in pancreatic cancer. In the

Results section, we summarize the strengths and weaknesses of each technique, and discuss our solution to leverage the unique strengths of AI to learn biomarkers from both HP-MR and MRI modalities, in addition to the available pathology and immunohistochemistry data to bridge this crucial knowledge gap. Our laboratories are currently pursuing an AI approach using an HP-MR dataset as applied to PDAC, the results of which will be published in the near future. In addition, we discuss the broad range of HP probes used to interrogate physiological functions such as metabolism and pH, which may expand the scope of applying AI to the functional imaging of PDAC.

Methods

A systematic review was performed following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) extension for scoping reviews (PRISMA-ScR) guidelines. Studies addressing the utilization of HP-MR and/or AI for early detection, assessment of aggressiveness, and interrogation of the early efficacy of therapy in patients with PDAC cited in recent clinical guidelines were extracted from the PubMed and Google Scholar databases. The studies were reviewed following predefined exclusion and inclusion criteria, which were grouped based on the utilization of HP-MR and AI in PDAC diagnosis.

Application of the HP-MR technique in pancreatic cancer is still nascent. We have reviewed every paper published in this broad area up to November 2020. Taken together, we have summarized our review in two tables. Table 1 summarizes all ^{13}C-labeled HP probes employed in interrogating different metabolic pathways in pancreatic cancer systems, and Table 2 summarizes all published applications of HP-MR in preclinical models of PDAC. In all, we have classified all of the physiological applications of HP-MR in pancreatic cancer under seven categories. The details of the deep-learning methods and HP-MR in different PDAC applications are discussed in the Introduction section above and in the relevant subsections of the Results.

Table 1. Review of 13C-labeled probes employed in interrogating different metabolic pathways in pancreatic cancer systems.

HP[a] probe	Biochemical reaction	T_1[b] of HP probe (seconds)	Quantification	Biological significance	References
[1-^{13}C] Pyruvate	Pyruvate to lactate (catalyzed by LDH[c]); pyruvate to alanine (catalyzed by ALT[d])	44-67	Rate constant of pyruvate to lactate (or alanine) or time-integrated ratio of lactate (or alanine)-to-pyruvate signals	Increased pyruvate-to-lactate flux is an indicator of the Warburg effect; total flux from pyruvate to (lactate+alanine) could be a measure of anaerobic glucose metabolism	Viale et al [17], Rao et al [22], Halbrook and Lyssiotis [49], Dutta et al [50]
[5-^{13}C] or [5-^{13}C-4-^{2}H2] glutamine	Glutamine to glutamate (catalyzed by glutaminase)	16-30	Time-integrated ratio of glutamate-to-glutamine signals	Indicator of glutamine addiction as a characteristic of certain cancers; also a measure of α-ketoglutarate metabolism (glutamate converts to α- ketoglutarate and can feed the TCA[e] cycle).	Son et al [51]
[H$^{13}CO_3^-$] bicarbonate	Bicarbonate to carbon dioxide	10-20	Using the relative concentrations of bicarbonate and carbon dioxide, apply the Henderson-Hasselbalch equation to calculate the tissue pH	The bicarbonate buffer system controls tissue pH; greater acidity of the tumor microenvironment has been linked to treatment resistance	Cruz-Monserrate et al [52], Gallagher et al [53]
[1,5-$^{13}C_2$] zymonic acid	N/A[f]	43-51	Chemical shift difference based on pH measurement	This is an organic moiety with no significant biological importance	Rao et al [21]
[1,4-$^{13}C_2$] fumarate	Fumarate to malate (cytosolic washout after cell necrosis)	~30	Malate signal is proportional to the amount of cell death	Fumarase (FH) enzyme is present in the cytosol and mitochondria of viable cells. Since cells cannot uptake fumarate, any HP malate production is a direct result of injected HP fumarate interacting with FH in the extracellular space, which has leaked out of necrotic cells; thus, it can be used to differentiate necrotic from viable cells	Silvers et al [54], Lee et al [55]
[1-^{13}C] dehydroascorbate (DHA)	DHA/ascorbate cycle, GSH[g]/GSSG[h] cycle, and NAPDH[i] to NADP+	>50	Ratio of time-integrated ascorbate-to-DHA signal	Greater flux from DHA to ascorbate indicates less redox stress inside the cell; this is also an indirect measure of the GSSG-to-GSH ratio and NADPH metabolism	Lai et al [56], Salamanca-Cardona et al [57], Keshari et al [58-60]
[1-^{13}C] α-keto isocaproate (α-KIC)	α-KIC to leucine (catalyzed by BCAT[j])	100	Ratio of time-integrated leucine-to-α-KIC signals	Indicator of BCAT level, which is upregulated in certain cancers	Wilson et al [61]

[a]HP: hyperpolarization.

[b]T_1: longitudinal relaxation time.

[c]LDH: lactate dehydrogenase.

[d]ALT: alanine transaminase.

[e]TCA: tricarboxylic acid cycle.

[f]N/A: not applicable.

[g]GSH: reduced glutathione.

[h]GSSG: glutathione disulfide.

[i]NAPDH: nicotinamide adenine dinucleotide phosphate.

[J]BCAT: branched-chain aminotransferase.

Table 2. Review of published applications of hyperpolarized magnetic resonance (HP-MR) in preclinical pancreatic ductal adenocarcinoma (PDAC) models.

Purpose of study	Mouse model/cell line/site of injection	HP-MR probe and downstream reaction	Results	Implications for HP-MR	Reference
To investigate whether pancreatic preneoplasia can be detected prior to the development of invasive cancers in GEM[a] models of PDAC using HP-MR.	I. For early-onset PDAC: GEM (K-Ras and p53 mutations); cell line II. For late-onset PDAC: GEM (only K-Ras mutation); cell line III. Wild-type mice; pancreatitis induced using caerulein injection	[1-^{13}C] pyruvate Pyruvate to lactate and pyruvate to alanine	I. The alanine-to-lactate signal ratio decreases progressively from the normal pancreas to pancreatitis to low-grade PanIN[b] to high-grade PanIN to PDAC, using HP-MR II. Holds true for individual mice with time as well as upon comparing the three groups, considering their genetic proximity to PDAC (I>II>III) III. Caused by increasing LDH[c] activity and decreasing ALT[d] activity	Clinical potential for early detection of advanced pancreatic preneoplasia in high-risk patients using the alanine-to-lactate signal ratio as a biomarker. Diseased areas can be monitored over time. Kinetic rate constants (k_{PA} and k_{PL}) can be used as metabolic imaging biomarkers of pancreatic premalignant lesions	Düwel et al [22], Dutta et al [50]
I. To determine if HP-MR can inform the sensitivity of pancreatic tumors to the hypoxia-activated prodrug TH-302 II. To test whether an adjuvant injection of pyruvate would enhance TH-302 efficacy	I. In female SCID mice: (i) highly sensitive to TH-302: SC[e] injection of the PDX[f] Hs766t; (ii) moderately sensitive to TH-302: SC injection of the PDX MIAPaCa-2; (iii) resistant to TH-302: SC injection of the PDX SU.86.86 II. Treatment groups: (i) Control, (ii) TH-302, (iii) TH-302+pyruvate	[1-^{13}C] Pyruvate Pyruvate to lactate	I. Higher lactate-to-pyruvate ratio observed in Hs766t and MIAPaCa groups; lower lactate-to-pyruvate ratio in SU.86.86 group II. Treatment with only TH-302: response of Hs766t (highly sensitive)> MIAPaCa-2> SU.86.86 (resistant). Treatment with TH-302+pyruvate: Hs766t and MIAPaCa-2 respond to a greater extent; SU.86.86 still resistant III. Exogenous pyruvate would be a successful adjuvant to enhance TH-302 efficacy because it stimulates oxygen consumption in glycolytic cells and decreases tumor pO_2 transiently	HP-MR can be used to predict treatment response to hypoxia-activated prodrugs, and thus provide a prognostic biomarker	Stødkilde-Jørgensen et al [63]

Purpose of study	Mouse model/cell line/site of injection	HP-MR probe and downstream reaction	Results	Implications for HP-MR	Reference
I. To determine a genetic biomarker of the response to the LDH-A inhibitor FX11 II. To test the response of HP-MR output to FX11 in PDAC murine models	I. In male nu/nu athymic mice: SC injection of PDX of PDAC with (i) wild-type TP53 or (ii) mutant TP53 II. Treatment groups: (i) Control, (ii) FX11	[1-^{13}C] Pyruvate Pyruvate to lactate	I. Mice injected with mutant TP53 PDAC responded to FX11; those injected with wild-type TP53 did not respond to FX11 by the end of 4 weeks II. The TP53 target gene *TIGAR* was responsible for the lack of response in wild-type TP53 PDAC. TIGAR lowers glycolytic flux and diverts glucose-6-phosphate into the PPP[g], reducing the dependence on glucose. III. Prior to FX11 treatment, the lactate-to-pyruvate ratio was increased in wild-type TP53 PDAC; following FX11 treatment, the lactate-to-pyruvate ratio decreased in mutant TP53 PDAC	I. HP-MR can be used to confirm the desired effect of metabolic therapies in tumors in early stages of drug development II. The lactate-to-pyruvate ratio can serve as a biomarker for response to metabolic therapies early in the treatment regimen	Wojtkowiak et al [64]
To determine if treating a PDAC cell line with β-lapachone, a chemotherapeutic agent activated by the enzyme NQ01 (upregulated in PDAC), will lead to the breakdown of energetic metabolic pathways such as glycolysis and the tricarboxylic acid cycle (due to depletion of NAD+ and ATP).	I. In vitro model: MIA-PaCa2 (NQO1+) pancreatic cancer cells (sensitive to β-lapachone) II. Treatment groups: (i) β-lapachone, (ii) no treatment	[1-^{13}C] Pyruvate Pyruvate to lactate	HP [1-^{13}C] pyruvate conversion to lactate was lower in cells treated with β-lapachone, suggesting that the activity of LDH is compromised from treatment	HP-MR can noninvasively detect the metabolic response of β-lapachone-treated cells. Thus, it can be used as a direct readout of treatment efficacy in PDAC patients with NQ01 upregulation	Rajeshkumar et al [65]

Purpose of study	Mouse model/cell line/site of injection	HP-MR probe and downstream reaction	Results	Implications for HP-MR	Reference
To determine whether measurement of the apparent diffusion coefficient (ADC) and conversion of injected copolarized ^{13}C-labeled pyruvic acid and fumaric acid can detect changes in lactate export and necrosis, respectively	In vitro model: (i) human breast cancer cell line MCF-7 (do not upregulate MCT1 or MCT4 under hypoxic conditions); (ii) mouse PDAC cell line 8932	Mixture of [1-^{13}C] pyruvic acid and [1,4-$^{13}C_2$] fumarate Pyruvate to lactate Fumarate to malate	I. The ADC_{lac}-to-ADC_{pyr} ratio is significantly greater in PDAC cells compared to that in MCF-7 cells II. This is corroborated by greater extracellular concentrations from the PDAC line III. Fumarate to malate conversion is detectable only in necrotic cells lysed with Triton X-100; no lactate formation was observed due to dilution of LDH and NADH[h].	I. Diffusion and conversion of HP pyruvate can provide information about the lactate efflux using the ADC_{lac}-to-ADC_{pyr} ratio, which is linked to the relative distribution of lactate in the intra- and extracellular compartments II. Diffusion MR and conversion of HP fumarate can inform necrosis; the rationale is that intracellular ADC<extracellular ADC due to restricted diffusion inside the cell III. Together, the cell's viability can be assessed. This may be used (1) to localize necrotic areas and (2) to assess the therapeutic response, especially for antiangiogenic agents such as bevacizumab	Silvers et al [54]
To determine whether mice injected with cancer cells (transfected with luciferase) in the peritoneum could be imaged using HP-MR and D_2O radicals	BALB/cA nu/nu mice: (i) peritoneal metastasis; (ii) intraperitoneal injection of human pancreatic carcinoma (SUIT-2) cells	Free radical (Oxo 63, CmP, nitroxyl)-D_2O probe	The image intensity correlated positively with the density of malignant ascites in the peritoneum	Radical-D_2O and HP-MR can be used to selectively visualize H_2O in the peritoneal cavity of mice and hence detect peritoneal metastasis early; this may then also be used to evaluate drug efficacy	Karlsson et al [66]

[a]GEM: genetically engineered mouse.

[b]PanIN: pancreatic intraepithelial neoplasia.

[c]LDH: lactate dehydrogenase.

[d]ALT: alanine transaminase.

[e]SC: subcutaneous.

[f]PDX: pancreatic ductal adenocarcinoma xenograft.

[g]PPP: pentose phosphate pathway.

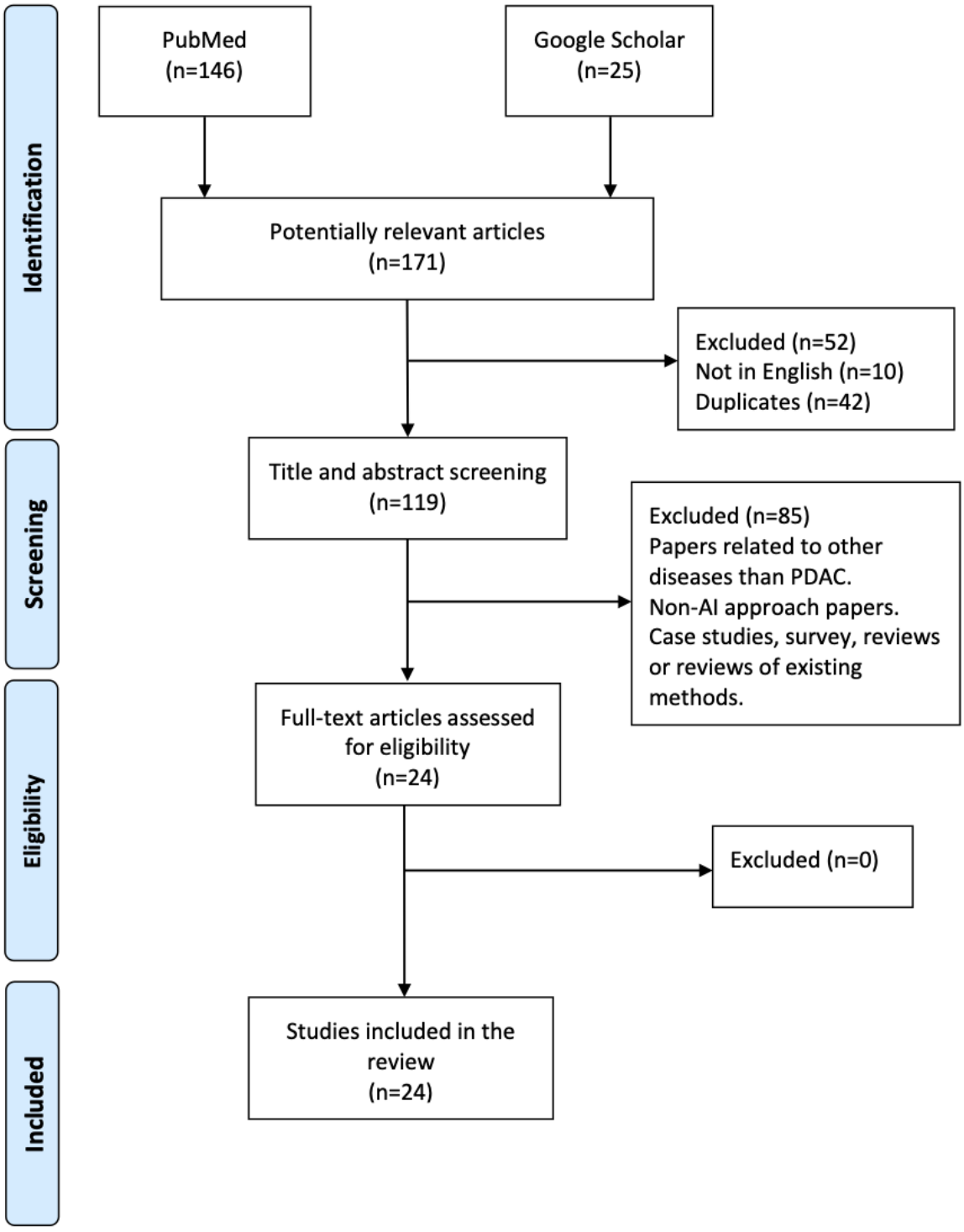
[h]NADH: nicotinamide adenine dinucleotide hydrogen.

Results

Characteristics of Retrieved Articles

For AI applications in pancreatic cancer, we retrieved 112 articles from the two sources, including 87 articles from PubMed and 25 articles from Google Scholar. An article was included if it satisfied our inclusion criteria: (1) written in English; (2) utilized AI/machine learning/deep learning for prediction, diagnosis, or classification; and (3) proposed a novel method of employing AI for PDAC (Figure 3). Review, evaluation, and comparison papers were therefore not included. Among the retrieved papers, a total of 17 met the inclusion criteria (Figure 3, Table 2, and Table S1 in Multimedia Appendix 1). The selected papers were grouped into six categories based on how AI was utilized in the context of PDAC to recognize the gaps in the previous studies and to discuss the novel approaches that fill the current gaps in detecting PDAC by imaging modalities.

Figure 3. PRISMA flow chart showing the selection criteria of the publications to include in this review. AI: artificial intelligence; PDAC: pancreatic ductal adenocarcinoma.



For HP-MR, we retrieved and reviewed all papers published in this broad area up to November 2020, which included six preclinical studies and one clinical study. We also reviewed several HP-MR–related articles (52 articles) that described new probes that can be applied in many functional future applications in PDAC. These references are not included in the PRISMA flow chart in Figure 3, as they have not yet been demonstrated in PDAC imaging and spectroscopic applications.

HP Metabolic Imaging Applications in PDAC

Context for Application of HP-MR in Pancreatic Cancer

PDAC tumors can be removed by surgery if detected early [23]. There is unequivocal evidence that diagnosis of PDAC at earlier, resectable stages has a profoundly favorable impact on prognosis [1]. The 5-year survival of patients with resected PDAC can reach up to ~25%-30% in major treatment centers, increasing

to 30%-60% for tumors <2 cm, and as high as 75% for minute lesions under 10 mm in size [46,47]. Unfortunately, most tumors are diagnosed at a late stage, once advanced into the local blood vessels and other body organs, and can no longer be excised. Thus, there is an urgent call to develop noninvasive imaging modalities for the early detection of PDAC, especially in high-risk patients (eg, those with a familial predisposition, long-standing diabetes, or chronic pancreatitis) [48]. Unlike other cancers such as breast or prostate cancer that have close to 100% survival if detected at early stages, PDAC is associated with a survival rate of only 39% even when detected at stage I [45]. Therefore, there is an urgency to develop novel methods for the detection of preneoplastic lesions in the pancreas.

Grading of PDAC

The type of treatment administered is often dependent on the tumor grade; therefore, there is a need for noninvasive methods to determine tumor grade. HP-MR uses metabolic changes to determine a grade [49]. Inside a PDAC tumor, the malignant cells become dependent on glycolysis for energy generation (Warburg effect). Dutta et al [50] recently reported that the aggressiveness of PDAC is directly correlated to pyruvate-to-lactate conversion measured using HP-MR (Table 1) and ex vivo ^{1}H nuclear magnetic resonance (NMR) spectroscopy in a panel of well-annotated patient-derived PDAC xenograft (PDX) mouse models. The ex vivo ^{1}H NMR spectroscopy results were also in good agreement with in vivo pyruvate-to-lactate conversion, showing a higher abundance of lactate in aggressive tumors. The expression levels of lactate dehydrogenase (LDH)-A and hypoxia-inducing factor-1α were also found to be elevated in aggressive tumors compared to those in less aggressive tumors in PDX mouse tumors. This study demonstrated that the aggressiveness of PDAC could be interrogated noninvasively by employing [1-^{13}C] pyruvate with HP-MR [50] to track cellular metabolic activity.

An interesting work by Serrao et al [23] (summarized in Table 2) demonstrated a method for the early detection of PDAC in murine models when the disease is in the early PanIN precursor stage employing HP [1-^{13}C] pyruvate. They used genetically engineered mice with K-Ras and p53 mutations or with K-Ras mutation only, which developed PanIN spontaneously. In addition, wild-type mice treated with caerulein injections to induce acute pancreatitis that developed into PanIN over time were included. The mice were imaged using HP [1-^{13}C] pyruvate at different stages of development from PanIN to PDAC precursor lesions, and the metabolic fluxes from [1-^{13}C] pyruvate to lactate and alanine were measured [23]. The results from individual mice showed a decreasing alanine-to-lactate ratio with disease progression from normal tissue to pancreatitis to low-grade PanIN to high-grade PanIN and finally to PDAC. Mice from all three groups followed this disease progression course, although with disparate timelines. The observed metabolic flux pattern correlated with increasing LDH activity and decreasing alanine transaminase activity. The metabolic flux from pyruvate to lactate and alanine is minimal in normal pancreatic tissue and progressively increases with disease progression. This technique can be used to create 3D metabolic maps of the pancreas to identify the extent of cancerous growth. This work was extended by Dutta et al [62] to demonstrate that real-time conversion kinetic rate constants (k_{PA} and k_{PL}) can be used as metabolic imaging biomarkers of premalignant pancreatic lesions. However, the translational potential of this approach can only be ascertained through clinical trials, which is feasible as this emerging technology can be translated to the clinic for the detection of premalignant pancreatic lesions in high-risk populations. Recently, a pilot study reported the feasibility of HP [1-^{13}C] pyruvate MRI in PDAC patients, and no adverse effect was observed after bolus injection of pyruvate [63]. These studies reveal the potential for the conversion of HP-pyruvate to lactate in the early detection of PDAC.

In addition, the HP pyruvate-to-lactate ratio may be used for staging tumors in the context of their aggression, although how this paradigm would fit in with the existing standards of staging is debatable (stage I or II: surgically resectable; stage III: locally advanced, unresectable; stage IV: metastatic) [48]. A very promising use of pyruvate-to-lactate flux is to identify PDAC advancing toward stage IV (metastasis) because these tumors show higher pyruvate-to-lactate conversion compared to that of less aggressive pancreatic cancer [50].

Early Assessment of Treatment Response

One of the promising utilities of HP-MR is its ability to assess treatment response early during the regimen; this has been established for solid tumors characterized by “aggression correlated with increased glycolysis.” This technique can thus complement the standard fluorodeoxyglucose-positron emission tomography imaging, which can only detect changes in tumor size (rather than intracellular metabolic changes) once it shrinks in response to a long-term regimen of chemotherapy or radiation therapy. Table 2 summarizes four published studies that show how HP-MR can be employed to predict responders (prognostic biomarkers) or assess treatment response early in PDAC tumors or cells [54,63-65]. The treatment efficacy of drugs (hypoxia-activated prodrugs, β-lapachone, and LDH-A inhibitors) evaluated using HP [1-^{13}C] pyruvate has only been studied in preclinical models to date; however, the preclinical data illustrate the ability of HP-MR to assist clinical trials by providing a framework for personalized medicine. HP-MR can provide information about the efficacy of drugs at an early stage that can lead to changes in clinical management, enabling the clinician to change the drug for a nonresponding patient to a more effective drug at an early stage.

Wojtkowiak et al [64] (Table 2) screened a hypoxia-activated prodrug (TH-302) as a monotherapy and in combination with pyruvate (not to be confused with the HP probe, [1-^{13}C] pyruvate) on three subcutaneous (Hs766t, MIAPaCa-2, and SU.86.86 cells) patient-derived xenografts of PDAC in mice. HP-MR using [1-^{13}C] pyruvate was employed to evaluate the metabolic phenotypes of Hs766t, MIAPaCa-2, and SU.86.86 PDAC cell line xenografts. The Hs766t and MIAPaCa-2 xenografts showed higher lactate-to-pyruvate ratios and more hypoxia. However, the SU.86.86 xenograft was resistant to the TH302 hypoxic prodrug because it was less hypoxic. The mice were treated for 2 weeks at a rate of five times a week and tumor

sizes were measured at regular intervals with calipers to determine the treatment efficacy. The Hs766t and MIAPaCa-2 groups showed an excellent response with TH302 compared to the SU.86.86 group [64].

Rajeshkumar et al [65] (Table 2) tested the treatment efficacy of the drug FX11, which inhibits LDH-A, on 15 patient-derived PDAC mouse models [65]. LDH-A converts pyruvate to lactate in the presence of its cofactor nicotinamide adenine dinucleotide hydrogen (NADH). Inhibition of LDH-A is a metabolic vulnerability that can be exploited for cancer treatment, and hence FX11 was evaluated in PDAC animal models. The drug was injected once daily for 4 weeks using PDX mouse models with tumors in their flank. The drug efficacy was tested using HP [1-^{13}C] pyruvate, which was injected into the mice prior to the start of treatment and 7 days after treatment, prior to any changes in tumor volume. Mice responding to the treatment showed a decreased lactate-to-pyruvate ratio after FX11 administration, whereas nonresponders showed an increased HP lactate-to-pyruvate ratio after the treatment. This result demonstrates the strength of the noninvasive HP-MR modality to predict treatment efficacy prior to tumor size reduction.

The β-lapachone chemotherapeutic drug acts on the quinone oxidoreductase 1 (NQO1)-mediated redox cycle, resulting in elevated superoxide and peroxide formation and in turn nicotinamide adenine dinucleotide (NAD+) depletion due to DNA damage and hyperactivation of poly(ADP-ribose) polymerase. Silvers et al [54] (Table 2) screened β-lapachone on patient-derived MIAPaCa2 cells (which were NQO1+, and hence sensitive to β-lapachone) in vitro to understand the effects of the drug on energy metabolism due to NAD+ depletion. Using metabolic imaging with HP pyruvate, this study showed a decrease in glycolytic flux upon treatment, thus validating the use of HP-MR as a direct readout of the treatment efficacy of β-lapachone in patients with PDAC with upregulated NQO1 expression.

Feuerecker et al [67] (Table 2) took an interesting in vitro approach to understand cancer tumor characteristics such as necrosis and lactate export, which are important parameters to determine cancer aggressiveness. They injected copolarized pyruvate and fumarate to measure the lactate export and necrosis in PDAC and MCF-7 breast carcinoma cells. Increased lactate export and cell necrosis are indicators of tumor aggressiveness, which can be determined using pyruvate-to-lactate flux and fumarate-to-malate flux, respectively. This study measured the apparent diffusion coefficient (ADC) and used HP-MR to examine the necrosis grade. The ADC of intracellular metabolites depends on the intactness of the plasma membrane. A greater $ADC_{lactate}$-to-$ADC_{pyruvate}$ ratio was observed in viable PDAC compared to MCF-7 breast carcinoma cells. The ADC measurements of metabolites could complement the HP lactate-to-pyruvate and HP fumarate-to-malate ratios to determine cell necrosis. This technique can be extended to in vivo measurements to determine the necrotic areas and evaluate the therapeutic response in PDAC patients.

Response to Radiation Therapy

Several studies have shown that early responses to radiation therapy can be assessed using molecular imaging. Ionizing radiation generates reactive oxygen species in tumor tissues [68]. Determining oxidative stress noninvasively could measure the extent of oxidative damage. HP pyruvate-to-lactate conversion predicted the response of solid tumors to radiation therapy in animal models [56]. This is an indirect approach and exploits the fact that pyruvate-to-lactate conversion requires reducing equivalents [56]. More direct measurement of redox stress inside cells is provided by HP dehydroascorbate–based MR, as summarized in Table 1 [57-61].

Collateral Lethality

Collateral lethality is a novel therapeutic approach that exploits the deletion of passenger genes alongside neighboring (deleted) tumor suppressor genes, thus conferring cancer-specific vulnerabilities [69]. One such instance is the deletion of both copies of malic enzyme 2 (*ME2*) with homozygous deletion of the neighboring *SMAD4* in many cases of PDAC. This makes ME3 inhibition a useful drug target because ME2 and ME3 are paralogous isoforms involved in NADPH regeneration and thus redox balance. The downstream effect of ME3 inhibition entails a reduction in the levels of branched-chain amino acid aminotransferase (BCAT) (encoded by *BCAT2*) via AMP-activated protein kinase–mediated mechanisms [69]. An HP α-keto isocaproate probe (Table 1), which can detect BCAT levels in vivo, could potentially be used for prognosis in the near future [66].

Imaging Peritoneal Metastasis

An interesting investigation by Eto et al [70] (Table 2) illustrates a method for the selective imaging of malignant ascites in a mouse model of peritoneal metastasis using HP-MR and bioluminescence studies [70]. In vivo HP images obtained using H_2O and D_2O as a radical in SUIT-2 peritoneal metastasis mice showed increasing intensity with time (0, 7, 14, and 21 days after tumor cell administration). This correlated with the increased density of bioluminescence as the density of PDAC ascites increased, thus providing the capability to monitor peritoneal metastasis as well as to evaluate the efficacy of antimetastatic drugs using these two techniques.

Metabolic Imaging Employing HP 13C Glutamine

Another possible approach for the early detection of PDAC is using HP ^{13}C glutamine. Son et al [51] described a noncanonical metabolic pathway for glutamine observed in PDAC cells. Normal cells convert glutamine-derived glutamate to α-ketoglutarate, which then feeds into the TCA cycle, whereas PDAC cells convert glutamine-derived glutamate into aspartate inside the mitochondria. This aspartate migrates to the cytosol and undergoes further biochemical reactions, which ultimately contribute to redox balance. This study also stated that the pathway is dispensable in normal cells (inhibiting the enzymes of this pathway is easily tolerated by normal cells), but is crucial to the survival of PDAC cells. However, it is not clear whether the pathway of glutamine to aspartate via glutamate is more pronounced in PDAC as compared to the normal tissue. If the glutamine to aspartate via glutamate pathway is upregulated by

several fold compared to that in normal cells, this metabolic pathway can be exploited to diagnose and grade PDAC tumors employing HP-MR with HP [^{13}C] glutamine. The feasibility of this approach depends on several factors. First, the decay time for HP ^{13}C glutamine must be considerably longer than the uptake of glutamine by PDAC cells, and the time of conversion to glutamate and then to aspartate. Additionally, there needs to be preferential uptake in PDAC cells compared to the cells of the normal pancreas. HP glutamine has already been used to study cancer cells from other tumor types [71] (Table 1).

Interstitial pH Mapping

Pancreatitis (inflammation of the pancreas) and PDAC are characterized by acidic microenvironments. The interstitial pH of the pancreas is reduced in patients with chronic pancreatitis [72-74]. The use of pH imaging to differentiate the acidic microenvironment of pancreatic tumors from that of PanIN lesions in mice has been elucidated by Cruz-Monserrate et al [52]. Several HP probes such as bicarbonate and zymonic acid can be potentially employed to image extracellular pH in tissue, which are summarized in Table 1 [22,53,55,67,75].

AI Applications in PDAC

Overview of AI and Deep Learning for PDAC

Deep learning has shown robust and extraordinary performance in medical image analysis. Many previous studies have explored the applications of AI, especially deep learning, for diagnosing and detecting various diseases, including pancreatic cancer, from different imaging modalities [76,77]. Leveraging HP-MR with deep learning is a promising approach to interrogate the early diagnosis and early efficacy of therapy for pancreatic cancer.

Most of the innovative applications of deep learning in biomedical imaging were triggered by convolutional neural networks (CNNs) [78], a powerful method for representation learning in images and structured data. As discussed above, neural networks, inspired by information transformation in the biological brain, require connections of all nodes of one layer to the next, which is insufficient for image analysis and fails to make use of spatial information. To overcome these issues, CNN introduces convolutional layers and pooling operations. In addition, many innovative modifications have been proposed to boost the performance of CNN, including dropout [79], batch normalization [80], and residual learning [81]. Essentially, the input to CNN is in a grid structure to preserve the spatial information, and then multiple convolutional layers and activation layers, interspersed with pooling layers, are utilized to process the data and learn structure in each level. Furthermore, a fully connected layer computes the final outputs for image analysis tasks.

A convolutional layer includes a set of filters with learnable parameters. Each filter is slid across the width and height of the input, and the dot product of the filter and input at every special position is calculated and goes through an activation function. A nonlinear activation function, typically rectified linear units (ReLUs), expands the potential in approximation of any nonlinear function [82]. The output of a convolution layer is a stack of activation maps of all filters. For pooling layers, it takes small regions in the feature map and produces a single number as the output to extract the most significant information learned from convolutional layers.

Several variants of CNNs with innovative architectures have been proposed to achieve better performance on specific tasks or types of data. VGG [83] introduced smaller filter kernels and constructed a deeper network compared with AlexNet [84], which first utilized ReLUs, dropout, and GPU accelerations. ResNet [81] proposed residual learning by using skip connections, which not only reduces the number of parameters but also makes the network deeper at up to 152 layers without a vanishing gradient. For biomedical images, U-Net [85] constructed downstreaming and upstreaming paths for biomedical images processing, connected by a skip connection, which concatenates features to the upstreaming path. V-Net [86] extended U-Net to 3D datasets using 3D convolutional layers and achieved extraordinary performance.

To review the previous studies on using AI for PDAC, we grouped the 17 selected papers (Table 3 and Table S1 in Multimedia Appendix 1) meeting our inclusion criteria into six categories based on how AI was utilized in the context of PDAC to help recognize the gaps in the previous studies and to discuss the novel approaches that can fill the current gaps in detecting PDAC by imaging modalities.

Table 3. Review of published applications of artificial intelligence for pancreatic ductal adenocarcinoma (PDAC).

Reference	Task	Method	Dataset	Performance
Liu et al [87]	A patient-specific tumor growth model based on longitudinal multimodal imaging data, including dual-phase CT[a] and FDG-PET[b]	A coupled PDE[c] system to develop a reaction-diffusion model enabling the incorporation of the cell metabolic rate and calculate ICVF[d]	Average ICVF difference (AICVFD) of tumor surface and tumor relative volume difference (RVD) on six patients with pathologically confirmed pancreatic neuroendocrine tumors	The ASD[e] between the predicted tumor and the reference tumor was 2.4 mm (SD 0.5), the RMSD[f] was 4.3% (SD 0.4), the AICVFD was 2.6% (SD 0.6), and the RVD was 7.7% (SD 1.3)
Fu et al [88]	CT pancreas segmentation (edge detection)	Proposed model includes 13 convolutional layers and 4 pooling layers; introduced multilayer upsampling structure	CT images from the General Surgery Department of Peking Union Medical College Hospital; 59 patients, including 15 with nonpancreas diseases and 44 with pancreas-related diseases	76.36% DSC[g]
Gibson et al [89]	Multiorgan segmentation on abdominal CT	Modified V-net proposed by replacing the convolutional layers in the encoder path by DenseNet consisting of stacks of dense blocks combined with bilinear upsampling in the decoder path	Two publicly available datasets: 43 subjects from the Cancer Imaging Archive Pancreas CT dataset with pancreas segmentations and 47 subjects from the Beyond the Cranial Vault segmentation challenge with segmentations of all organs except the duodenum	DSC of 78% for the pancreas, 90% for the stomach, and 76% for the esophagus
Luo et al [90]	Preoperative prediction of pancreatic neuroendocrine neoplasms (pNENs) grading by CECT[h]	The proposed 3D CNN[i] composed of 1 CNN layer with 1 rectifier linear unit layer, a max pooling layer, 12 IdentityBlock, 4 ConvBlock, 1 global average pooling layer, and 1 fully connected layer	CT images of 93 patients from Sun Yat-Sen University and 19 patients from The Cancer Center of Sun Yat-Sen University with pathologically confirmed pNENs	AUC[j] of 0.81
Liu et al [91]	Diagnosis of pancreatic cancer using CNN	Pretrained VGG16 serves as a feature extraction network, and Faster R-CNN is used for diagnosis	6084 enhanced CT horizontal images from 338 pancreatic cancer patients	AUC of 0.96
Boers et al [92]	Segmentation of the pancreas	U-net model was changed by adding one interactive layer that takes feedback from the annotator while freezing other layers to do retraining	Public dataset (Gibson et al [89]), which contains 90 late venous-phase abdominal CT images	DSC of 78.1% (SD 8.7)
Liu et al [93]	Cone-beam CT (CBCT) quality and HU[k] accuracy improvement	A self-attention cycle generative adversarial network (cycleGAN) was used to generate CBCT from synthetic CT	Thirty patients previously treated with pancreas SBRT[l] at Emory University	Mean absolute error between CT and synthetic CT of 56.89 (SD 13.84) HU and 1.06 (SD 15.86) HU between CT and the raw CBCT
Park et al [94]	CT data collection for deep learning	Two U-Net models were linked by an organ-attention module	From 575 participants, a total of 1150 CT images	Mean DSC of 89.4% and mean surface distance of 1.29 mm
Liu et al [95]	Multiorgan segmentation for pancreatic CT	3D U‐Net with an attention strategy is proposed	100 patients with CT simulation scanned	DSC of 91% (SD 3), 89% (SD 6), 86% (SD 6), 95% (SD 2), 95% (SD 2), 96% (SD 1), 87% (SD 5), and 93% (SD 3) for the large bowel, small bowel, duodenum, left kidney, right kidney, liver, spinal cord, and stomach, respectively.

Reference	Task	Method	Dataset	Performance
Mu et al [96]	Prediction of clinically relevant postoperative pancreatic fistula using CECT	One convblock, 8 residual blocks, and one fully connected layer	A group of 513 patients underwent pancreaticoenteric anastomosis after PD[m] at three institutions between 2006 and 2019	AUC of 0.89
Chu et al [77]	Deep-learning models for abdominal organs segmentation using CT	Three networks with different voxel sizes. Each network follows an encoder-decoder topology and includes a series of CNN layers max pooling and deconvolutional layers	Dual-phase CT from 575 control subjects and 750 patients with PDAC from 2005 to 2017	Accuracy of 87.8% (SD 3.1)
Suman et al [97]	Deep-learning models for pancreas segmentation using CT	NVIDIA 3D Slicer segmentation module	347 CECT scans based on a statement of a negative or unremarkable pancreas in the original radiologist's report	DSC of 63% (SD 15)
Ma et al [98]	Pancreatic cancer diagnosis using CT	The model consisted of three convolutional layers and a fully connected layer	3494 CT images from 222 patients with pathologically confirmed pancreatic cancer and 3751 CT images from 190 patients with normal pancreas from June 2017 to June 2018	Accuracy of 82.06%, 79.06%, and 78.80% on plain phase, arterial phase, and venous phase
Zhang et al [99]	Tumor detection framework for pancreatic cancer via CECT	Feature pyramid networks with Faster R-CNN	2890 CT images from Qingdao University	AUC of 0.9455
Corral et al [100]	Intraductal papillary mucinous neoplasms (IPMN) classification using MRI[n]	Integration of CNN and SVM[o]	171 patients, 39 MRIs with no pancreatic lesions served, and 132 confirmed IPMN	AUC of 0.77
Hussein et al [101]	IPMN classification using MRI	VGG network and SVM	171 MRIs for 38 subjects with normal pancreas, and the remaining 133 from subjects diagnosed with IPMN	Accuracy of 84.22%
Liang et al [102]	MRI pancreas segmentation	SVM with recursively retraining samples	MRIs from four patients with locally advanced pancreatic cancer	DSC of 86%
Zheng et al [103]	MRI pancreas segmentation	2D Unet	20 patients with PDAC	DSC of 73.88%

[a]CT: computed tomography.
[b]FDG-PET: fluorodeoxyglucose-positron emission tomography.
[c]PDE: partial differential equation.
[d]ICVF: intracellular volume fraction.
[e]ASD: average surface distance.
[f]RMSD: root mean square deviation.
[g]DSC: Dice similarity coefficient.
[h]CECT: contrast-enhanced computed tomography.
[i]CNN: convolutional neural network.
[j]AUC: area under the receiver operating characteristic curve.
[k]HU: Hounsfield unit.
[l]SBRT: stereotactic body radiotherapy.
[m]PD: pancreatoduodenectomy.
[n]MRI: magnetic resonance imaging.
[o]SVM: support vector machine.

Tumor Growth Model

Tumor growth, especially for pancreatic neuroendocrine tumors, is related to cancer cell properties and relies on the dynamic interaction between cells and the microenvironment. Swanson et al [104] proposed a reaction-diffusion model by assuming infiltrative growth of the tumor cells but did not consider the cell metabolic rate. Liu et al [87] introduced dual-phase

CT-measured intracellular volume fraction (ICVF) to the reaction-diffusion model. Cell metabolic rate was considered in the prediction of pancreatic neuroendocrine tumor growth. They evaluated the model by comparing predictions with sequential observations regarding average surface distance, root mean square deviation (RMSD) of the ICVF map, and average ICVF difference in six patients with pancreatic neuroendocrine tumors. Although the RMSD was around 4.3%, the limited number of patients involved might have undermined the final findings.

Organ/Multiorgan Segmentation and Edge Detection in Medical Images

Fu et al [88] discussed the application of a CNN consisting of 13 convolution layers and 4 pooling layers with a multilayer upsampling structure in pancreas segmentation from CT images. The proposed model was evaluated using real PDAC CT images from a dataset created by the General Surgery Department of Peking Union Medical College Hospital. The 59 patients consisted of 15 patients with nonpancreas diseases and 44 patients with pancreas-related diseases. A Dice similarity coefficient (DSC) of 76.36% was achieved. The introduced fusion layer provided good visualization for decision-making and multilayer upsampling improved the performance. However, due to the limited number of CT images for training and validation, its performance suffered from the risk of overfitting as the reported SD from precision 5-fold cross-validation was very high (mean SD of 18.08 across all classes). Moreover, the reported precision and recall for the healthy cohort (80.95 and 86.53, respectively) was much higher than that for the IPMN (75.39 and 67.37) or pancreatic neuroendocrine tumor (70.44 and 74.86) cohort.

Alternatively, to implement multiorgan segmentation, especially on abdominal CT in the pancreas, Gibson et al [89] modified V-net by replacing the convolutional layers in the encoder path by DenseNet consisting of stacks of dense blocks combined with bilinear upsampling in the decoder path. They applied this on two public datasets: one including 43 subjects from the Cancer Imaging Archive Pancreas CT data with pancreas segmentation, and the other including 47 subjects from the Beyond the Cranial Vault segmentation challenge with segmentations of all organs except the duodenum. They achieved a DSC of 78%. The introduced dense feature stack considerably reduced the number of parameters for medical image classification tasks. However, this approach is only appropriate for relatively small datasets because of overfitting issues.

As another example of an attempt to improve pancreas segmentation performance in CT scans, Boers et al [92] developed an interactive version of U-net (iUnet) by adding one interactive layer after the last fully connected layer takes feedback from annotators while freezing other layers to do the retraining. This was applied to a public CT dataset used in Gibson et al [89], which contains 90 late venous-phase abdominal CT images and a respective reference segmentation. A DSC of 78.1% (SD 8.7%) was achieved from the interactive version of iUnet, which outperformed previous methods using the same dataset. However, this approach may also suffer from overfitting issues since interactive processes may introduce external information, which limits its scalability.

Liu et al [93] presented a deep-attention U-net approach to solve multiorgan segmentation for pancreatic cancer CT images. This method achieved state-of-the-art performance, but its performance in pancreas segmentation is unclear.

Besides CT images, investigators using T1-MRI proposed several innovative approaches to segment the pancreas. Liang et al [102] introduced a top-down and bottom-up approach. In the top-down path, the initial planning contours derived from simulation MR images are transferred to daily images, and in the bottom-up path, the probabilistic support vector machine (SVM) is used with recursively retraining samples. The final result is obtained by fusing both paths and the final reported DSC was 86%. Zheng et al [103] proposed a 2D U-Net approach with shadow sets for MRI and CT pancreas segmentation. The usage of shadow sets reduced uncertainty and achieved a DSC of 84.37% on the NIH-CT-82 public dataset [105] and 73.88% on an MRI dataset collected from Changhai Hospital, including 20 patients with PDAC.

Prediction of PDAC and Risk Evaluation

To implement preoperative prediction of pancreatic neuroendocrine neoplasms (pNENs) grading by CT, Luo et al [90] applied a CNN model with identity blocks and convolution blocks to a CT imaging dataset consisting of 93 patients from the hospital. An arterial model employed for the pathological grading of pNENs achieved an area under the receiver operating characteristic curve (AUC) of 0.81. Due to the limitations of the dataset, simple deep-learning models may undermine the feature extraction ability and lead to suboptimal performance. In addition, the limited observations may lead to a lack of an independent evaluation dataset and invalidation of n-fold cross-validation, which constrains the scalability and generalizability of the proposed model.

For the prediction of clinically relevant postoperative pancreatic fistula using CT images, Mu et al [96] utilized a Resnet18 model with fewer filters on 513 patients imaged between 2006 and 2019. All patients underwent pancreatico-enteric anastomosis after the diagnosis of pancreas disease at three institutions. Compared to the commonly used Fistula Risk Score (FRS), the proposed model improved the prediction AUC from 0.73 to 0.89. This study illustrated that deep learning might overcome intermediate risk score issues in FRS with greater predictability.

Hussein et al [101] utilized clustering and SVM with initial label estimation for risk stratification of pancreatic tumors. The model outperformed other unsupervised methods, achieving 58% accuracy out of 171 scans, of which 38 subjects were normal and the remaining 133 were diagnosed with IPMNs. In addition, Corral et al [100] performed similar experiments using MRI with CNN and SVM as the final classifier, and achieved sensitivity and specificity of 0.92 and 0.52, respectively, for the detection of dysplasia. In this task, the deep-learning protocol barely outperformed radiologists due to unbalanced data issues and the complexity of IPMN.

Diagnosis of PDAC

To implement diagnosis of PDAC from CT images, Liu et al [91] utilized a pretrained VGG16 model as a feature extraction network in conjunction with a faster recurrent CNN (R-CNN) as a decision-maker. Their CT dataset consisted of 6084 enhanced CT horizontal images from the abdomen of 338 PDAC patients. This achieved an AUC of 0.9632 for the prediction of PDAC. R-CNN usage for sequential information extraction greatly improved the diagnostic performance, and its dynamic feature extraction provided model interpretability and scalability.

Ma et al [98] utilized a regular CNN with four hidden layers on 3494 CT images from 222 patients with pathologically confirmed PDAC and 3751 CT images from 190 patients with a normal pancreas as controls from the First Affiliated Hospital, Zhejiang University School of Medicine. The overall diagnostic accuracy of trained binary classifiers was over 95%. However, it failed to beat human performance. With a similar data size, Liu et al [91] achieved better performance using a more complicated, faster R-CNN, implying the complexity of pancreatic cancer detection and its need for an appropriate model architecture design and parameter fine-tuning.

Zhang et al [99] proposed a tumor detection framework for PDAC using CT. The framework utilized feature pyramid networks with the faster R-CNN model. The Affiliated Hospital of Qingdao University provided a dataset containing 2890 CT images, and a classification AUC of 0.9455 was achieved. This framework outperformed the state-of-the-art methods, but still suffered from input uncertainty inherent in closed-source datasets. The comparison would be more effective if a public dataset was used in the experiment.

Improvement of Image Quality

Stereotactic body radiotherapy (SBRT) has shown more success in patients with locally advanced pancreatic cancer compared to conventional radiotherapy. To overcome interference due to motion in the breathing cycle and patient weight loss [106], cone-beam CT (CBCT) is commonly used for target position verification and setup displacement correction to avoid suboptimal target coverage and excessive doses to organs at risk. However, raw CBCT data cannot be used for SBRT dosage calculation due to considerable artifacts such as streaking and shading [107-109] caused by scatter contamination, resulting in different Hounsfield unit (HU) values from CT scans [110]. To improve the HU fidelity of CBCT, Liu et al [95] utilized self-attention cycleGan-based CBCT to synthetic CT (sCT) models on a dataset consisting of 30 patients previously treated with pancreas SBRT at Emory University. The mean absolute error of the proposed framework between CT and sCT was 56.89 (SD 13.84) HU compared to 81.06 (SD 15.86) HU between CT and the raw CBCT.

Criteria to Evaluate Annotation Accuracy in Medical Images

To test the CT data collection quality, Park et al [94] proposed two U-net–linked networks, linked by an organ-attention module, to test the performance of a well-annotated dataset, including a total of 575 participants and 1150 CT images. After appropriate management of the annotation process, an average DSC of 89.4% was achieved. This study innovatively employed a deep-learning model to test CT image annotation performance, and improved the annotation quality for further analysis and research. However, this approach still suffers from uncertainty introduced by model training and simulation.

Suman et al [97] used CNNs to train technologists in labeling pancreas segmentation CT datasets. DSC was improved through interactions between model output and expert correction, which implied that annotation quality was enhanced.

Discussion

In this review, we have discussed how two different techniques, HP-MR and AI, are revealing exciting information about PDAC and PanINs that was not accessible by diagnostic imaging even a few years ago. Deep-learning models eliminate the requirement of domain knowledge for feature engineering that is necessary for conventional machine learning models by learning from raw data. Deep-learning models are capable of learning features from the raw data and apply nonlinear transformations to map the input data to high-dimensional representations trivializing classification or regression. These models are uniquely able to transform multiple modalities into common latent space to synthesize features across all modalities to improve classification performance. However, there is no free lunch, and the flexibility and high accuracy resulting from millions of parameters comes with a requirement of a huge training dataset in comparison with other machine-learning techniques. Moreover, these models suffer from lack of interpretability and uncertainty measurement. In machine-learning algorithms, there is a tradeoff between interpretability and accuracy. When the prediction accuracy grows with more complex (increase in the number of trainable parameters) deep-learning models, the interpretability decreases. For instance, ResNet contains 5×10^7 parameters requiring 10^{10} floating point operations for a single classification task, making it almost impossible to be traced or explained by humans [81,111]. Lastly, deep-learning models do not provide any uncertainty measurement to measure how certain the model is with its prediction. These models are blindly used with the assumption of “good accuracy,” whereas previous experience has shown that these models are susceptible to overconfident decision-making, especially when the new data are far from the training data distribution (corner case). The lack of interpretability and uncertainty estimation is even more serious in clinical decision-making tasks since it is needed for building trust in the model’s prediction.

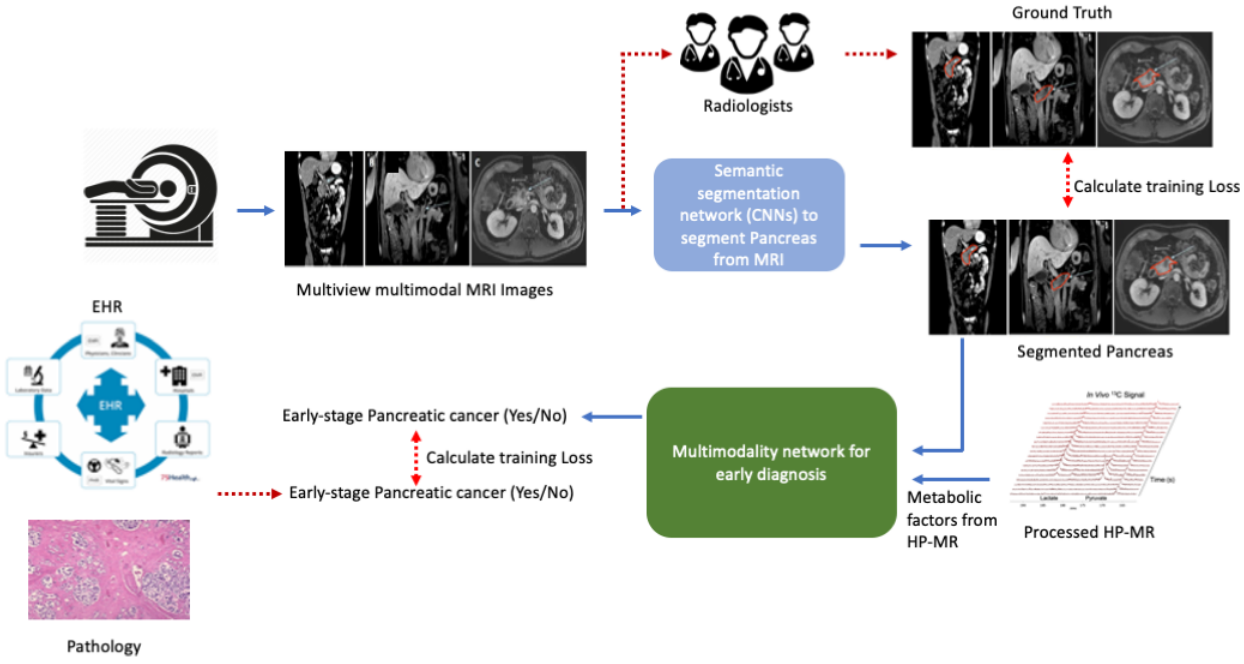
Studies on HP-MR have demonstrated that this modality can detect metabolic changes at very early stages of lesion formation in the pancreas (eg, PanIN 1 and 2); however, this is more of a spectroscopic technique than an imaging modality. Furthermore, the signal from HP compounds lasts no more than a few minutes depending on the T_1 that allows for rapid acquisition of dynamic metabolic flux in the organ of interest. Table 4 summarizes the strengths and weaknesses of AI, MRI, and HP-MR.

Table 4. Strengths and weaknesses of artificial intelligence (AI), magnetic resonance imaging (MRI), and hyperpolarized magnetic resonance (HP-MR).

Technique	Strengths	Weaknesses
MRI	Rapid acquisition of anatomical images. Well-established and widely distributed imaging modality.	Poor signal-to-noise ratio and contrast-to-noise ratio. Cannot detect pancreatic cancer at early stages.
HP-MR	Real-time metabolic flux measurements at the organ of interest. Can detect premalignant stages of pancreatic cancer.	Short time window of imaging (~2 minutes). Expensive initial investment in the infrastructure. Slow adoption in the clinical setting.
AI	No feature engineering, ability to learn features from raw data. Ability to learn features from and across multiple modalities. High accuracy result.	Intensive data requirement. High uncertainty on corner cases. Lack of interpretability.

To take advantage of the strengths of AI, MRI, and HP-MR, and mitigate their weaknesses, we propose the following pipeline as illustrated in Figure 4. Our pipeline leverages the unique capability of AI to learn features from each and across both HP-MR and MRI as complementary modalities to investigate the early detection of PDAC by overlaying the anatomical imaging for localized spectroscopic information of real-time metabolic flux in the pancreas. Additionally, we utilize Grad-CAM [112,113] and concrete dropout to provide a visual explanation, and introduce Bayesian inference to estimate uncertainty in the model's decision.

Figure 4. Schematic illustrating the concept of leveraging anatomical magnetic resonance imaging (MRI), hyperpolarized magnetic resonance (HP-MR), and artificial intelligence as complementary modalities toward developing actionable biomarkers of pancreatic ductal adenocarcinoma. CNNs: convolutional neural networks; EHR: electronic health record.



The training process of our pipeline is as follows: axial, sagittal, and coronal MR images in the T1 and T2 modalities are annotated to highlight the pancreas area by radiologists to train a deep-learning semantic segmentation network developed by our team. We extract the ROIs from MR images (ie, the pancreas). The extracted ROIs with metabolic information from HP-MR are the inputs for our multimodal deep-learning model to predict pancreatic cancer status. The appropriate combination of MRI and HP-MR as complementary modalities improves the classification performance. Therefore, the ground truth for our second deep-learning model is the presence of early stages of PDAC established by pathology reports and electronic health records of the patients. The training path is shown with the dashed lines and the inference path is shown with the solid lines in Figure 4. It has been estimated that there is a window of opportunity of ~10 years from the moment in which a pancreatic epithelial cell undergoes an oncogenic hit and the time of diagnosis of, often fatal, pancreatic cancer [46,114]. Together, AI, HP-MR, and conventional MRI as complementary modalities can address this knowledge gap in diagnostic imaging within this crucial time window of opportunity to save lives.

Leveraging AI and HP-MR applications together may lead to the development of real-time actionable biomarkers of early detection, assessing aggressiveness, and interrogating the early efficacy of therapy in PDAC. For example, multimodal AI can learn features from both HP-MR, as well as anatomical MRI and CT imaging modalities, to yield "hybrid biomarkers" and reduce the time required to detect PDAC evolution in three key areas of tumor progression: initial development of the tumor, its regression following therapy, and the eventual recurrence of the tumor. This innovative synthesis of these techniques may result in a more sensitive readout of tumor progression that can be readily translated and significantly impact how PDAC patients, as well as patients at high risk of developing this deadly disease, are currently managed in the clinic.

Acknowledgments

This research was funded in part by a grant from Pancreatic Cancer Action Network (PANCAN; 16-65-BHAT to PB and FM); Cancer Prevention and Research Institute of Texas (CPRIT; RP180164 to PB); Institutional Research Grants and a Startup grant (to PB) from MD Anderson Cancer Center; grants from the US National Cancer Institute (U01 CA214263 to SS and AK; U54 CA151668 and R21 CA185536 to PB; R01 CA218004 to AM; and 1P50 CA221707-01). XJ is a CPRIT Scholar in Cancer Research (RR180012). He was supported in part by Christopher Sarofim Family Professorship, UT Stars Award, UTHealth Startup, the National Institutes of Health (NIH) under award number U01 TR002062. SS is supported in part by RR180012 and RP200526 Computational Cancer Biology Training Program Fellowship from Gulf Coast Consortia (CPRIT) grants. KH is supported by a CPRIT grant (RP170593). This work also was supported by the NIH/NCI Cancer Center Support Grant under award number P30 CA016672 to MD Anderson Cancer Center.

Conflicts of Interest

AM receives royalties for a pancreatic cancer biomarker test from Cosmos Wisdom Biotechnology, and is listed as an inventor on a patent that has been licensed by Johns Hopkins University to ThriveEarlier Detection.

Multimedia Appendix 1
Supplemental material.
[DOCX File , 298 KB - medinform_v9i6e26601_app1.docx]

References

1. Blackford A, Canto M, Klein A, Hruban R, Goggins M. Recent trends in the incidence and survival of stage 1A pancreatic cancer: a surveillance, epidemiology, and end results analysis. J Natl Cancer Inst 2020 Nov 01;112(11):1162-1169 [FREE Full text] [doi: 10.1093/jnci/djaa004] [Medline: 31958122]
2. Rahib L, Smith BD, Aizenberg R, Rosenzweig AB, Fleshman JM, Matrisian LM. Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States. Cancer Res 2014 Jun 01;74(11):2913-2921 [FREE Full text] [doi: 10.1158/0008-5472.CAN-14-0155] [Medline: 24840647]
3. Siegel R, Miller K, Jemal A. Cancer statistics, 2020. CA Cancer J Clin 2020 Jan;70(1):7-30. [doi: 10.3322/caac.21590] [Medline: 31912902]
4. Ardenkjaer-Larsen JH, Fridlund B, Gram A, Hansson G, Hansson L, Lerche MH, et al. Increase in signal-to-noise ratio of > 10,000 times in liquid-state NMR. Proc Natl Acad Sci U S A 2003 Sep 02;100(18):10158-10163 [FREE Full text] [doi: 10.1073/pnas.1733835100] [Medline: 12930897]
5. Dutta P, Salzillo TC, Pudakalakatti S, Gammon ST, Kaipparettu BA, McAllister F, et al. Assessing therapeutic efficacy in real-time by hyperpolarized magnetic resonance metabolic imaging. Cells 2019 Apr 11;8(4):340 [FREE Full text] [doi: 10.3390/cells8040340] [Medline: 30978984]
6. Bhattacharya P, Ross BD, Bünger R. Cardiovascular applications of hyperpolarized contrast media and metabolic tracers. Exp Biol Med (Maywood) 2009 Dec;234(12):1395-1416. [doi: 10.3181/0904-MR-135] [Medline: 19934362]
7. Kurhanewicz J, Vigneron DB, Ardenkjaer-Larsen JH, Bankson JA, Brindle K, Cunningham CH, et al. Hyperpolarized C MRI: Path to clinical translation in oncology. Neoplasia 2019 Jan;21(1):1-16 [FREE Full text] [doi: 10.1016/j.neo.2018.09.006] [Medline: 30472500]
8. McRobbie D, Moore E, Graves M, Prince M. MRI from Picture to Proton. 3rd edition. Cambridge: Cambridge University Press; Mar 13, 2017.
9. Barker P, Bizzi A, De SN, Gullapalli R, Lin D. Clinical MR Spectroscopy: Techniques and Applications. Cambridge: Cambridge University Press; Nov 12, 2009:9780521868983.
10. Overhauser AW. Polarization of nuclei in metals. Phys Rev 1953 Oct 15;92(2):411-415. [doi: 10.1103/PhysRev.92.411]
11. Carver TR, Slichter CP. Polarization of nuclear spins in metals. Phys Rev 1953 Oct 1;92(1):212-213. [doi: 10.1103/PhysRev.92.212.2]
12. Abragam A, Goldman M. Principles of dynamic nuclear polarisation. Rep Prog Phys 2001 Feb 09;41(3):395-467. [doi: 10.1088/0034-4885/41/3/002]

13. Nikolaou P, Goodson BM, Chekmenev EY. NMR hyperpolarization techniques for biomedicine. Chemistry 2015 Feb 16;21(8):3156-3166 [FREE Full text] [doi: 10.1002/chem.201405253] [Medline: 25470566]
14. Walker TG, Happer W. Spin-exchange optical pumping of noble-gas nuclei. Rev Mod Phys 1997 Apr 1;69(2):629-642. [doi: 10.1103/RevModPhys.69.629]
15. Hirsch ML, Kalechofsky N, Belzer A, Rosay M, Kempf JG. Brute-force hyperpolarization for NMR and MRI. J Am Chem Soc 2015 Jul 08;137(26):8428-8434. [doi: 10.1021/jacs.5b01252] [Medline: 26098752]
16. Hövener JB, Pravdivtsev AN, Kidd B, Bowers CR, Glöggler S, Kovtunov KV, et al. Parahydrogen-based hyperpolarization for biomedicine. Angew Chem Int Ed Engl 2018 Aug 27;57(35):11140-11162 [FREE Full text] [doi: 10.1002/anie.201711842] [Medline: 29484795]
17. Viale A, Reineri F, Santelia D, Cerutti E, Ellena S, Gobetto R, et al. Hyperpolarized agents for advanced MRI investigations. Q J Nucl Med Mol Imaging 2009 Dec;53(6):604-617 [FREE Full text] [Medline: 20016452]
18. Golman K, Zandt RI, Lerche M, Pehrson R, Ardenkjaer-Larsen JH. Metabolic imaging by hyperpolarized 13C magnetic resonance imaging for in vivo tumor diagnosis. Cancer Res 2006 Nov 15;66(22):10855-10860 [FREE Full text] [doi: 10.1158/0008-5472.CAN-06-2564] [Medline: 17108122]
19. Lin C, Salzillo T, Bader D, Wilkenfeld S, Awad D, Pulliam T, et al. Prostate cancer energetics and biosynthesis. Adv Exp Med Biol 2019;1210:185-237. [doi: 10.1007/978-3-030-32656-2_10] [Medline: 31900911]
20. Warburg O. On the origin of cancer cells. Science 1956 Feb 24;123(3191):309-314. [doi: 10.1126/science.123.3191.309] [Medline: 13298683]
21. Rao Y, Gammon S, Zacharias NM, Liu T, Salzillo T, Xi Y, et al. Hyperpolarized [1-C]pyruvate-to-[1-C]lactate conversion is rate-limited by monocarboxylate transporter-1 in the plasma membrane. Proc Natl Acad Sci U S A 2020 Sep 08;117(36):22378-22389 [FREE Full text] [doi: 10.1073/pnas.2003537117] [Medline: 32839325]
22. Düwel S, Hundshammer C, Gersch M, Feuerecker B, Steiger K, Buck A, et al. Imaging of pH in vivo using hyperpolarized C-labelled zymonic acid. Nat Commun 2017 May 11;8:15126. [doi: 10.1038/ncomms15126] [Medline: 28492229]
23. Serrao EM, Kettunen MI, Rodrigues TB, Dzien P, Wright AJ, Gopinathan A, et al. MRI with hyperpolarised [1-13C]pyruvate detects advanced pancreatic preneoplasia prior to invasive disease in a mouse model. Gut 2016 Mar;65(3):465-475 [FREE Full text] [doi: 10.1136/gutjnl-2015-310114] [Medline: 26347531]
24. Zierhut M, Yen Y, Chen A, Bok R, Albers M, Zhang V, et al. Kinetic modeling of hyperpolarized 13C1-pyruvate metabolism in normal rats and TRAMP mice. J Magn Reson 2010 Jan;202(1):85-92 [FREE Full text] [doi: 10.1016/j.jmr.2009.10.003] [Medline: 19884027]
25. Kaplan A, Haenlein M. Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Business Horiz 2019 Jan;62(1):15-25. [doi: 10.1016/j.bushor.2018.08.004]
26. Chen Y, Lorenzi N, Sandberg W, Wolgast K, Malin B. Identifying collaborative care teams through electronic medical record utilization patterns. J Am Med Inform Assoc 2017 Apr 01;24(e1):e111-e120 [FREE Full text] [doi: 10.1093/jamia/ocw124] [Medline: 27570217]
27. Du J, Jia P, Dai Y, Tao C, Zhao Z, Zhi D. Gene2vec: distributed representation of genes based on co-expression. BMC Genomics 2019 Feb 04;20(Suppl 1):82 [FREE Full text] [doi: 10.1186/s12864-018-5370-x] [Medline: 30712510]
28. Hsieh K, Wang Y, Chen L, Zhao Z, Savitz S, Jiang X, et al. Drug repurposing for COVID-19 using graph neural network with genetic, mechanistic, and epidemiological validation. ArXiv 2020 Sep 23:arxiv:2009.10931v1. [Medline: 32995367]
29. Ianevski A, Giri AK, Gautam P, Kononov A, Potdar S, Saarela J, et al. Prediction of drug combination effects with a minimal set of experiments. Nat Mach Intell 2019 Dec;1(12):568-577 [FREE Full text] [doi: 10.1038/s42256-019-0122-4] [Medline: 32368721]
30. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform 2018 Nov 27;19(6):1236-1246 [FREE Full text] [doi: 10.1093/bib/bbx044] [Medline: 28481991]
31. Zheng T, Xie W, Xu L, He X, Zhang Y, You M, et al. A machine learning-based framework to identify type 2 diabetes through electronic health records. Int J Med Inform 2017 Jan;97:120-127 [FREE Full text] [doi: 10.1016/j.ijmedinf.2016.09.014] [Medline: 27919371]
32. Yu C, Lin Y, Lin C, Lin S, Wu JL, Chang S. Development of an online health care assessment for preventive medicine: a machine learning approach. J Med Internet Res 2020 Jun 05;22(6):e18585 [FREE Full text] [doi: 10.2196/18585] [Medline: 32501272]
33. Birnbaum ML, Kulkarni PP, Van Meter A, Chen V, Rizvi AF, Arenare E, et al. Utilizing machine learning on internet search activity to support the diagnostic process and relapse detection in young individuals with early psychosis: feasibility study. JMIR Ment Health 2020 Sep 01;7(9):e19348 [FREE Full text] [doi: 10.2196/19348] [Medline: 32870161]
34. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. JMIR Med Inform 2019 Apr 27;7(2):e12239 [FREE Full text] [doi: 10.2196/12239] [Medline: 31066697]
35. Contreras I, Vehi J. Artificial intelligence for diabetes management and decision support: literature review. J Med Internet Res 2018 May 30;20(5):e10775 [FREE Full text] [doi: 10.2196/10775] [Medline: 29848472]

36. Shan H, Padole A, Homayounieh F, Kruger U, Khera RD, Nitiwarangkul C, et al. Competitive performance of a modularized deep neural network compared to commercial algorithms for low-dose CT image reconstruction. Nat Mach Intell 2019 Jun;1(6):269-276 [FREE Full text] [doi: 10.1038/s42256-019-0057-9] [Medline: 33244514]
37. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med 2019 Jun;25(6):954-961. [doi: 10.1038/s41591-019-0447-x] [Medline: 31110349]
38. Shams S, Platania R, Zhang J, Kim J, Lee K, Park S. Deep generative breast cancer screening and diagnosis. 2018 Sep 16 Presented at: Medical Image Computing and Computer Assisted Intervention – MICCAI; September 16-20, 2018; Granada, Spain p. 859-867.
39. Platania R, Shams S, Yang S, Zhang J, Lee K, Park SJ. Automated breast cancer diagnosis using deep learning and region of interest detection (BC-DROID). : Association for Computing Machinery; 2017 Aug 20 Presented at: 8th ACM international conference on bioinformatics, computational biology, and health informatics; August 20-23, 2017; Boston, MA p. 536-543. [doi: 10.1145/3107411.3107484]
40. Arcadu F, Benmansour F, Maunz A, Willis J, Haskova Z, Prunotto M. Deep learning algorithm predicts diabetic retinopathy progression in individual patients. NPJ Digit Med 2019;2:92. [doi: 10.1038/s41746-019-0172-3] [Medline: 31552296]
41. Lundervold A, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. Z Med Phys 2019 May;29(2):102-127 [FREE Full text] [doi: 10.1016/j.zemedi.2018.11.002] [Medline: 30553609]
42. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 1943 Dec;5(4):115-133. [doi: 10.1007/bf02478259]
43. Machado NO, Al Qadhi H, Al Wahibi K. Intraductal papillary mucinous neoplasm of pancreas. N Am J Med Sci 2015 May;7(5):160-175 [FREE Full text] [doi: 10.4103/1947-2714.157477] [Medline: 26110127]
44. Felsenstein M, Hruban RH, Wood LD. New developments in the molecular mechanisms of pancreatic tumorigenesis. Adv Anat Pathol 2018 Mar;25(2):131-142 [FREE Full text] [doi: 10.1097/PAP.0000000000000172] [Medline: 28914620]
45. Cancer facts and figures 2021. American Cancer Society. URL: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2021/cancer-facts-and-figures-2021.pdf [accessed 2021-05-26]
46. Chari S, Kelly K, Hollingsworth M, Thayer S, Ahlquist D, Andersen D, et al. Early detection of sporadic pancreatic cancer: summative review. Pancreas 2015 Jul;44(5):693-712 [FREE Full text] [doi: 10.1097/MPA.0000000000000368] [Medline: 25931254]
47. Lennon AM, Wolfgang CL, Canto MI, Klein AP, Herman JM, Goggins M, et al. The early detection of pancreatic cancer: what will it take to diagnose and treat curable pancreatic neoplasia? Cancer Res 2014 Jul 01;74(13):3381-3389 [FREE Full text] [doi: 10.1158/0008-5472.CAN-14-0734] [Medline: 24924775]
48. Ryan DP, Hong TS, Bardeesy N. Pancreatic adenocarcinoma. N Engl J Med 2014 Sep 11;371(11):1039-1049. [doi: 10.1056/NEJMra1404198] [Medline: 25207767]
49. Halbrook CJ, Lyssiotis CA. Employing metabolism to improve the diagnosis and treatment of pancreatic cancer. Cancer Cell 2017 Jan 09;31(1):5-19 [FREE Full text] [doi: 10.1016/j.ccell.2016.12.006] [Medline: 28073003]
50. Dutta P, Perez M, Lee J, Kang Y, Pratt M, Salzillo T, et al. Combining hyperpolarized real-time metabolic imaging and NMR spectroscopy to identify metabolic biomarkers in pancreatic cancer. J Proteome Res 2019 Jul 05;18(7):2826-2834. [doi: 10.1021/acs.jproteome.9b00132] [Medline: 31120258]
51. Son J, Lyssiotis CA, Ying H, Wang X, Hua S, Ligorio M, et al. Glutamine supports pancreatic cancer growth through a KRAS-regulated metabolic pathway. Nature 2013 Apr 04;496(7443):101-105 [FREE Full text] [doi: 10.1038/nature12040] [Medline: 23535601]
52. Cruz-Monserrate Z, Roland CL, Deng D, Arumugam T, Moshnikova A, Andreev OA, et al. Targeting pancreatic ductal adenocarcinoma acidic microenvironment. Sci Rep 2014 Mar 19;4:4410. [doi: 10.1038/srep04410] [Medline: 24642931]
53. Gallagher FA, Kettunen MI, Day SE, Hu D, Ardenkjaer-Larsen JH, Zandt R, et al. Magnetic resonance imaging of pH in vivo using hyperpolarized 13C-labelled bicarbonate. Nature 2008 Jun 12;453(7197):940-943. [doi: 10.1038/nature07017] [Medline: 18509335]
54. Silvers MA, Deja S, Singh N, Egnatchik RA, Sudderth J, Luo X, et al. The NQO1 bioactivatable drug, β-lapachone, alters the redox state of NQO1+ pancreatic cancer cells, causing perturbation in central carbon metabolism. J Biol Chem 2017 Nov 03;292(44):18203-18216 [FREE Full text] [doi: 10.1074/jbc.M117.813923] [Medline: 28916726]
55. Lee Y, Zacharias NM, Piwnica-Worms D, Bhattacharya PK. Chemical reaction-induced multi-molecular polarization (CRIMP). Chem Commun (Camb) 2014 Nov 07;50(86):13030-13033 [FREE Full text] [doi: 10.1039/c4cc06199c] [Medline: 25224323]
56. Lai SY, Fuller CD, Bhattacharya PK, Frank SJ. Metabolic imaging as a biomarker of early radiation response in tumors. Clin Cancer Res 2015 Jul 31;21(22):4996-4998. [doi: 10.1158/1078-0432.ccr-15-1214]
57. Salamanca-Cardona L, Keshari KR. (13)C-labeled biochemical probes for the study of cancer metabolism with dynamic nuclear polarization-enhanced magnetic resonance imaging. Cancer Metab 2015;3:9 [FREE Full text] [doi: 10.1186/s40170-015-0136-2] [Medline: 26380082]

58. Keshari KR, Wilson DM, Sai V, Bok R, Jen K, Larson P, et al. Noninvasive in vivo imaging of diabetes-induced renal oxidative stress and response to therapy using hyperpolarized 13C dehydroascorbate magnetic resonance. Diabetes 2015 Feb;64(2):344-352 [FREE Full text] [doi: 10.2337/db13-1829] [Medline: 25187363]
59. Keshari KR, Sai V, Wang ZJ, Vanbrocklin HF, Kurhanewicz J, Wilson DM. Hyperpolarized [1-13C]dehydroascorbate MR spectroscopy in a murine model of prostate cancer: comparison with 18F-FDG PET. J Nucl Med 2013 Jun 10;54(6):922-928 [FREE Full text] [doi: 10.2967/jnumed.112.115402] [Medline: 23575993]
60. Keshari KR, Kurhanewicz J, Bok R, Larson PEZ, Vigneron DB, Wilson DM. Hyperpolarized 13C dehydroascorbate as an endogenous redox sensor for in vivo metabolic imaging. Proc Natl Acad Sci U S A 2011 Nov 15;108(46):18606-18611 [FREE Full text] [doi: 10.1073/pnas.1106920108] [Medline: 22042839]
61. Wilson DM, Di Gialleonardo V, Wang ZJ, Carroll V, Von Morze C, Taylor A, et al. Hyperpolarized C spectroscopic evaluation of oxidative stress in a rodent model of steatohepatitis. Sci Rep 2017 Apr 20;7:46014. [doi: 10.1038/srep46014] [Medline: 28425467]
62. Dutta P, Pando SC, Mascaro M, Riquelme E, Zoltan M, Zacharias NM, et al. Early detection of pancreatic intraepithelial neoplasias (PanINs) in transgenic mouse model by hyperpolarized C metabolic magnetic resonance spectroscopy. Int J Mol Sci 2020 May 25;21(10):3722 [FREE Full text] [doi: 10.3390/ijms21103722] [Medline: 32466260]
63. Stødkilde-Jørgensen H, Laustsen C, Hansen E, Schulte R, Ardenkjaer-Larsen J, Comment A, et al. Pilot study experiences with hyperpolarized [1- C]pyruvate MRI in pancreatic cancer patients. J Magn Reson Imaging 2020 Mar;51(3):961-963. [doi: 10.1002/jmri.26888] [Medline: 31368215]
64. Wojtkowiak JW, Cornnell HC, Matsumoto S, Saito K, Takakusagi Y, Dutta P, et al. Pyruvate sensitizes pancreatic tumors to hypoxia-activated prodrug TH-302. Cancer Metab 2015;3(1):2 [FREE Full text] [doi: 10.1186/s40170-014-0026-z] [Medline: 25635223]
65. Rajeshkumar N, Dutta P, Yabuuchi S, de Wilde RF, Martinez GV, Le A, et al. Therapeutic targeting of the Warburg effect in pancreatic cancer relies on an absence of p53 function. Cancer Res 2015 Aug 15;75(16):3355-3364 [FREE Full text] [doi: 10.1158/0008-5472.CAN-15-0108] [Medline: 26113084]
66. Karlsson M, Jensen PR, in 't Zandt R, Gisselsson A, Hansson G, Duus J, et al. Imaging of branched chain amino acid metabolism in tumors with hyperpolarized 13C ketoisocaproate. Int J Cancer 2010 Aug 01;127(3):729-736. [doi: 10.1002/ijc.25072] [Medline: 19960440]
67. Feuerecker B, Durst M, Michalik M, Schneider G, Saur D, Menzel M, et al. Hyperpolarized C diffusion MRS of co-polarized pyruvate and fumarate to measure lactate export and necrosis. J Cancer 2017;8(15):3078-3085 [FREE Full text] [doi: 10.7150/jca.20250] [Medline: 28928899]
68. Riley P. Free radicals in biology: oxidative stress and the effects of ionizing radiation. Int J Radiat Biol 1994 Jan 03;65(1):27-33. [doi: 10.1080/09553009414550041] [Medline: 7905906]
69. Dey P, Baddour J, Muller F, Wu CC, Wang H, Liao W, et al. Genomic deletion of malic enzyme 2 confers collateral lethality in pancreatic cancer. Nature 2017 Feb 02;542(7639):119-123 [FREE Full text] [doi: 10.1038/nature21052] [Medline: 28099419]
70. Eto H, Hyodo F, Nakano K, Utsumi H. Selective imaging of malignant ascites in a mouse model of peritoneal metastasis using in vivo dynamic nuclear polarization-magnetic resonance imaging. Anal Chem 2016 Feb 16;88(4):2021-2027. [doi: 10.1021/acs.analchem.5b04821] [Medline: 26796949]
71. Qu W, Zha Z, Lieberman B, Mancuso A, Stetz M, Rizzi R, et al. Facile synthesis [5-(13)C-4-(2)H(2)]-L-glutamine for hyperpolarized MRS imaging of cancer cell metabolism. Acad Radiol 2011 Aug;18(8):932-939. [doi: 10.1016/j.acra.2011.05.002] [Medline: 21658976]
72. Patel A, Toyama M, Alvarez C, Nguyen T, Reber P, Ashley S, et al. Pancreatic interstitial pH in human and feline chronic pancreatitis. Gastroenterology 1995 Nov;109(5):1639-1645. [doi: 10.1016/0016-5085(95)90654-1] [Medline: 7557149]
73. Toyama MT, Patel AG, Nguyen T, Ashley SW, Reber HA. Effect of ethanol on pancreatic interstitial pH and blood flow in cats with chronic pancreatitis. Ann Surg 1997 Feb;225(2):223-228. [doi: 10.1097/00000658-199702000-00011] [Medline: 9065300]
74. Reber PU, Patel AG, Toyama MT, Ashley SW, Reber HA. Feline model of chronic obstructive pancreatitis: effects of acute pancreatic duct decompression on blood flow and interstitial pH. Scand J Gastroenterol 1999 Apr;34(4):439-444. [doi: 10.1080/003655299750026489] [Medline: 10365907]
75. Gallagher FA, Kettunen MI, Hu D, Jensen PR, Zandt RIT, Karlsson M, et al. Production of hyperpolarized [1,4-13C2]malate from [1,4-13C2]fumarate is a marker of cell necrosis and treatment response in tumors. Proc Natl Acad Sci U S A 2009 Nov 24;106(47):19801-19806 [FREE Full text] [doi: 10.1073/pnas.0911447106] [Medline: 19903889]
76. Chu LC, Goggins MG, Fishman EK. Diagnosis and detection of pancreatic cancer. Cancer J 2017;23(6):333-342. [doi: 10.1097/ppo.0000000000000290]
77. Chu LC, Park S, Kawamoto S, Wang Y, Zhou Y, Shen W, et al. Application of deep learning to pancreatic cancer detection: lessons learned from our initial experience. J Am Coll Radiol 2019 Sep;16(9 Pt B):1338-1342. [doi: 10.1016/j.jacr.2019.05.034] [Medline: 31492412]
78. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE 1998 Nov 01;86(11):2278-2324. [doi: 10.1109/5.726791]

79. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J Machine Learn Res 2014;15(1):1929-1958.
80. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. 2015 Jun 01 Presented at: Proceedings of the 32nd Annual Conference on Machine Learning; July 7-9, 2015; Lille, France p. 448-456.
81. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 Presented at: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 27-30, 2016; Las Vegas, NV p. 770-778. [doi: 10.1109/cvpr.2016.90]
82. Sonoda S, Murata N. Neural network with unbounded activation functions is universal approximator. Appl Comput Harmon Anal 2017 Sep;43(2):233-268. [doi: 10.1016/j.acha.2015.12.005]
83. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv. 2014. URL: https://arxiv.org/abs/1409.1556 [accessed 2021-06-02]
84. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017 May 24;60(6):84-90. [doi: 10.1145/3065386]
85. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. Cham: Springer; 2015 Presented at: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015; October 5-9, 2015; Munich, Germany. [doi: 10.1007/978-3-319-24574-4_28]
86. Milletari F, Navab N, Ahmadi S. V-net: Fully convolutional neural networks for volumetric medical image segmentation. 2016 Oct 25 Presented at: 2016 Fourth International Conference on 3D Vision (3DV); October 25-28, 2016; Stanford, CA p. 565-571. [doi: 10.1109/3dv.2016.79]
87. Liu Y, Sadowski SM, Weisbrod AB, Kebebew E, Summers RM, Yao J. Patient specific tumor growth prediction using multimodal images. Med Image Anal 2014 Apr;18(3):555-566 [FREE Full text] [doi: 10.1016/j.media.2014.02.005] [Medline: 24607911]
88. Fu M, Wu W, Hong X, Liu Q, Jiang J, Ou Y, et al. Hierarchical combinatorial deep learning architecture for pancreas segmentation of medical computed tomography cancer images. BMC Syst Biol 2018 Apr 24;12(Suppl 4):56 [FREE Full text] [doi: 10.1186/s12918-018-0572-z] [Medline: 29745840]
89. Gibson E, Giganti F, Hu Y, Bonmati E, Bandula S, Gurusamy K, et al. Automatic multi-organ segmentation on abdominal CT with dense V-networks. IEEE Trans Med Imaging 2018 Aug;37(8):1822-1834 [FREE Full text] [doi: 10.1109/TMI.2018.2806309] [Medline: 29994628]
90. Luo Y, Chen X, Chen J, Song C, Shen J, Xiao H, et al. Preoperative prediction of pancreatic neuroendocrine neoplasms grading based on enhanced computed tomography imaging: validation of deep learning with a convolutional neural network. Neuroendocrinology 2020;110(5):338-350 [FREE Full text] [doi: 10.1159/000503291] [Medline: 31525737]
91. Liu S, Li S, Guo Y, Zhou Y, Zhang Z, Li S, et al. Establishment and application of an artificial intelligence diagnosis system for pancreatic cancer with a faster region-based convolutional neural network. Chin Med J (Engl) 2019 Dec 05;132(23):2795-2803 [FREE Full text] [doi: 10.1097/CM9.0000000000000544] [Medline: 31856050]
92. Boers TGW, Hu Y, Gibson E, Barratt DC, Bonmati E, Krdzalic J, et al. Interactive 3D U-net for the segmentation of the pancreas in computed tomography scans. Phys Med Biol 2020 Mar 11;65(6):065002. [doi: 10.1088/1361-6560/ab6f99] [Medline: 31978921]
93. Liu Y, Lei Y, Fu Y, Wang T, Tang X, Jiang X, et al. CT-based multi-organ segmentation using a 3D self-attention U-net network for pancreatic radiotherapy. Med Phys 2020 Sep;47(9):4316-4324. [doi: 10.1002/mp.14386] [Medline: 32654153]
94. Park S, Chu LC, Fishman EK, Yuille AL, Vogelstein B, Kinzler KW, et al. Annotated normal CT data of the abdomen for deep learning: Challenges and strategies for implementation. Diagn Interv Imaging 2020 Jan;101(1):35-44 [FREE Full text] [doi: 10.1016/j.diii.2019.05.008] [Medline: 31358460]
95. Liu Y, Lei Y, Wang T, Fu Y, Tang X, Curran W, et al. CBCT-based synthetic CT generation using deep-attention cycleGAN for pancreatic adaptive radiotherapy. Med Phys 2020 Jun;47(6):2472-2483 [FREE Full text] [doi: 10.1002/mp.14121] [Medline: 32141618]
96. Mu W, Liu C, Gao F, Qi Y, Lu H, Liu Z, et al. Prediction of clinically relevant pancreatico-enteric anastomotic fistulas after pancreatoduodenectomy using deep learning of preoperative computed tomography. Theranostics 2020;10(21):9779-9788 [FREE Full text] [doi: 10.7150/thno.49671] [Medline: 32863959]
97. Suman G, Panda A, Korfiatis P, Edwards ME, Garg S, Blezek DJ, et al. Development of a volumetric pancreas segmentation CT dataset for AI applications through trained technologists: a study during the COVID 19 containment phase. Abdom Radiol (NY) 2020 Dec;45(12):4302-4310 [FREE Full text] [doi: 10.1007/s00261-020-02741-x] [Medline: 32939632]
98. Ma H, Liu Z, Zhang J, Wu F, Xu C, Shen Z, et al. Construction of a convolutional neural network classifier developed by computed tomography images for pancreatic cancer diagnosis. World J Gastroenterol 2020 Sep 14;26(34):5156-5168 [FREE Full text] [doi: 10.3748/wjg.v26.i34.5156] [Medline: 32982116]
99. Zhang Z, Li S, Wang Z, Lu Y. A novel and efficient tumor detection framework for pancreatic cancer via CT images. 2020 Jul 20 Presented at: 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); July 20-24, 2020; Virtual p. 1160-1164.

100. Corral JE, Hussein S, Kandel P, Bolan CW, Bagci U, Wallace MB. Deep learning to classify intraductal papillary mucinous neoplasms using magnetic resonance imaging. Pancreas 2019 Jul;48(6):805-810. [doi: 10.1097/MPA.0000000000001327] [Medline: 31210661]
101. Hussein S, Kandel P, Bolan CW, Wallace MB, Bagci U. Lung and pancreatic tumor characterization in the deep learning era: novel supervised and unsupervised learning approaches. IEEE Trans Med Imaging 2019 Aug;38(8):1777-1787. [doi: 10.1109/TMI.2019.2894349] [Medline: 30676950]
102. Liang F, Qian P, Su K, Baydoun A, Leisser A, Van Hedent S, et al. Abdominal, multi-organ, auto-contouring method for online adaptive magnetic resonance guided radiotherapy: An intelligent, multi-level fusion approach. Artif Intell Med 2018 Aug;90:34-41. [doi: 10.1016/j.artmed.2018.07.001] [Medline: 30054121]
103. Zheng H, Chen Y, Yue X, Ma C, Liu X, Yang P, et al. Deep pancreas segmentation with uncertain regions of shadowed sets. Magn Reson Imaging 2020 May;68:45-52. [doi: 10.1016/j.mri.2020.01.008] [Medline: 31987903]
104. Swanson KR, Alvord EC, Murray JD. A quantitative model for differential motility of gliomas in grey and white matter. Cell Prolif 2000 Oct;33(5):317-329 [FREE Full text] [doi: 10.1046/j.1365-2184.2000.00177.x] [Medline: 11063134]
105. NIH Clinical Center releases dataset of 32,000 CT images. NIH Clinical Center. 2018 Jul 20. URL: https://www.nih.gov/news-events/news-releases/nih-clinical-center-releases-dataset-32000-ct-images [accessed 2021-05-26]
106. Langen K, Jones D. Organ motion and its management. Int J Radiat Oncol Biol Phys 2001 May 01;50(1):265-278. [doi: 10.1016/s0360-3016(01)01453-5] [Medline: 11316572]
107. Schulze R, Heil U, Gross D, Bruellmann D, Dranischnikow E, Schwanecke U, et al. Artefacts in CBCT: a review. Dentomaxillofac Radiol 2011 Jul;40(5):265-273 [FREE Full text] [doi: 10.1259/dmfr/30642039] [Medline: 21697151]
108. Cho PS, Johnson RH, Griffin TW. Cone-beam CT for radiotherapy applications. Phys Med Biol 1995 Nov;40(11):1863-1883. [doi: 10.1088/0031-9155/40/11/007] [Medline: 8587937]
109. Barrett JF, Keat N. Artifacts in CT: recognition and avoidance. Radiographics 2004;24(6):1679-1691. [doi: 10.1148/rg.246045065] [Medline: 15537976]
110. Abe T, Tateoka K, Saito Y, Nakazawa T, Yano M, Nakata K, et al. Method for converting cone-beam CT values into Hounsfield units for radiation treatment planning. Int J Med Phys Clin Eng Radiat Oncol 2017;06(04):361-375. [doi: 10.4236/ijmpcero.2017.64032]
111. Buhrmester V, Münch D, Arens M. Analysis of explainers of black box deep neural networks for computer vision: A survey. arXiv. 2019 Nov 27. URL: https://arxiv.org/abs/1911.12116 [accessed 2021-06-02]
112. Selvaraju R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. 2017 Presented at: Proceedings of the IEEE international conference on computer vision; October 22-29, 2017; Venice, Italy p. 618-626. [doi: 10.1109/iccv.2017.74]
113. Gal Y, Hron J, Kendall A. Concrete dropout. 2017 Dec Presented at: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach, CA p. 3584-3593.
114. McAllister F, Bailey J, Alsina J, Nirschl C, Sharma R, Fan H, et al. Oncogenic Kras activates a hematopoietic-to-epithelial IL-17 signaling axis in preinvasive pancreatic neoplasia. Cancer Cell 2014 May 12;25(5):621-637 [FREE Full text] [doi: 10.1016/j.ccr.2014.03.014] [Medline: 24823639]

Abbreviations

ADC: apparent diffusion coefficient
AI: artificial intelligence
AUC: area under the receiver operating characteristic curve
BCAT: branched-chain amino acid aminotransferase
CBCT: cone-beam computed tomography
CNN: convolutional neural network
CT: computed tomography
DNP: dynamic nuclear polarization
DSC: Dice similarity coefficient
FRS: Fistula Risk Score
HP: hyperpolarized
HU: Hounsfield unit
ICVF: intracellular volume fraction
IPMN: intraductal papillary mucinous neoplasms
LDH: lactate dehydrogenase
MR: magnetic resonance
MRI: magnetic resonance imaging
NAD+: nicotinamide adenine dinucleotide
NADH: nicotinamide adenine dinucleotide hydrogen
NQO1: quinone oxidoreductase 1

PanIN: pancreatic intraepithelial neoplasia
PDAC: pancreatic ductal adenocarcinoma
PDX: patient-derived xenograft
pNEN: pancreatic neuroendocrine neoplasm
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-analysis Extension for Scoping Reviews
R-CNN: recurrent convolutional neural network
ReLU: rectified linear unit
ROI: region of interest
SBRT: stereotactic body radiation therapy
sCT: synthetic computed tomography
SVM: support vector machine
T1: longitudinal relaxation time
TCA: tricarboxylic acid

Edited by C Lovis; submitted 18.12.20; peer-reviewed by 哲 張, Z Su, H Zhang; comments to author 12.01.21; revised version received 24.02.21; accepted 03.04.21; published 17.06.21.

Please cite as:
Enriquez JS, Chu Y, Pudakalakatti S, Hsieh KL, Salmon D, Dutta P, Millward NZ, Lurie E, Millward S, McAllister F, Maitra A, Sen S, Killary A, Zhang J, Jiang X, Bhattacharya PK, Shams S
Hyperpolarized Magnetic Resonance and Artificial Intelligence: Frontiers of Imaging in Pancreatic Cancer
JMIR Med Inform 2021;9(6):e26601
URL: https://medinform.jmir.org/2021/6/e26601
doi:10.2196/26601
PMID:34137725

Viewpoint

Ethical Applications of Artificial Intelligence: Evidence From Health Research on Veterans

Christos Makridis[1,2,3*], PhD; Seth Hurley[1*], PhD; Mary Klote[4], MD; Gil Alterovitz[1,5*], PhD

[1]National Artificial Intelligence Institute, Department of Veterans Affairs, Washington, DC, United States

[2]Stanford Digital Economy Lab, Stanford University, Stanford, CA, United States

[3]WP Carey School of Business, Arizona State University, Tempe, AZ, United States

[4]Office of Research & Development, Department of Veterans Affairs, Washington, DC, United States

[5]Boston Children's Hospital, Harvard Medical School, Boston, MA, United States

[*]these authors contributed equally

Corresponding Author:
Christos Makridis, PhD
National Artificial Intelligence Institute
Department of Veterans Affairs
810 Vermont Avenue NW
Washington, DC, 20420
United States
Phone: 1 2022977787
Email: christos.makridis@va.gov

Abstract

Background: Despite widespread agreement that artificial intelligence (AI) offers significant benefits for individuals and society at large, there are also serious challenges to overcome with respect to its governance. Recent policymaking has focused on establishing principles for the trustworthy use of AI. Adhering to these principles is especially important for ensuring that the development and application of AI raises economic and social welfare, including among vulnerable groups and veterans.

Objective: We explore the newly developed principles around trustworthy AI and how they can be readily applied at scale to vulnerable groups that are potentially less likely to benefit from technological advances.

Methods: Using the US Department of Veterans Affairs as a case study, we explore the principles of trustworthy AI that are of particular interest for vulnerable groups and veterans.

Results: We focus on three principles: (1) designing, developing, acquiring, and using AI so that the benefits of its use significantly outweigh the risks and the risks are assessed and managed; (2) ensuring that the application of AI occurs in well-defined domains and is accurate, effective, and fit for the intended purposes; and (3) ensuring that the operations and outcomes of AI applications are sufficiently interpretable and understandable by all subject matter experts, users, and others.

Conclusions: These principles and applications apply more generally to vulnerable groups, and adherence to them can allow the VA and other organizations to continue modernizing their technology governance, leveraging the gains of AI while simultaneously managing its risks.

(JMIR Med Inform 2021;9(6):e28921) doi:10.2196/28921

KEYWORDS

artificial intelligence; ethics; veterans; health data; technology; Veterans Affairs; health technology; data

Ethical Applications of Artificial Intelligence in Veterans' Health Research

There is increasing recognition that artificial intelligence (AI) offers significant potential to help or harm the world. Much like other technologies, ranging from the internet to computers, AI is neither bad nor good: the impact of AI depends on how its users wield it. Already, there is an emerging body of AI use cases in health care [1,2], including for vulnerable groups and veterans [3], that are increasingly originating from populations the federal government considers to be "potentially vulnerable patient populations" [4]. These groups can be especially sensitive to adoption of technology; therefore, additional

scrutiny is required around the ethical underpinnings and likely causal effects on these groups. In this sense, the question is not whether the federal government should engage AI for broader social benefit, but how it can do so using a values-based framework to guide AI applications and their continued research and development.

At least since the publication of the Belmont Report [5], there has been general recognition in the federal government of three principles that guide the introduction of new technologies to this day. First, *respect for persons* details that individual autonomy and privacy must be protected. Second, *beneficence* states that technologies should be designed to maximize the potential net benefits to society, safeguarding against potential harms and long-term consequences. Third, *justice* ensures that there are equitable benefits from research. That is, when individual data is collected, it must be used to benefit those individuals. Although the Belmont Report focused on biomedical technologies, they exhibit many similarities with AI, particularly in terms of their ethical implications and long-lasting impacts.

The primary contribution of this commentary is to explore the ethical applications of AI by building on the Belmont Report and relating it with the principles established in the recent executive order on trustworthy AI. Although there has been a recognition of data ethics and privacy within the US federal government, a new challenge has emerged: how can the federal government balance between the competing priorities of stewarding sensitive data and using AI to analyze it to drive veteran outcomes?

To answer this question, we apply the perspective of the Veterans Health Administration, within the Department of Veterans Affairs (VA), which has the largest integrated health care system in the United States and has pioneered several technological aspects now widely seen in this country, such as electronic health records (EHRs). Additionally, more than half of physicians training within the United States receive some training at a VA medical center. By evaluating uses for AI and implications in health care, veteran input and priorities can be proactively developed to enhance care. For example, the VA is already using AI to facilitate early detection of cancer [3], detection of acute kidney injury [6], and prediction of loneliness and declines in mental health [7,8]. These examples all highlight the ways that AI can be used to advance patient outcomes; however, they also point toward data privacy and trust considerations.

The recent executive order, "Promoting the Use of Trustworthy Artificial Intelligence in Government" [9], provides a framework for the VA to move forward with using AI to improve veteran health on a larger and more systematic scale. We focus on three principles that are especially relevant to the advancement of the health and well-being of veterans: (1) purposeful and performance-driven; (2) accurate, reliable, and effective; and (3) understandable.

1. Purposeful and Performance Driven

> *…seek opportunities for designing, developing, acquiring, and using AI, where the benefits of use significantly outweigh the risks and the risks are assessed and managed* [9]

The VA is working to employ AI in high priority areas where there is robust opportunity to advance veteran health outcomes. Recent work indicates that difficulty in transitioning to civilian life is a critical factor underlying negative mental health outcomes in veterans. For example, Makridis and Hirsch [10] have documented a deterioration in labor market outcomes among veterans over the past decade, showing that veterans are increasingly concentrated in metropolitan areas with lower wage and employment growth. Moreover, Makridis et al [11] show that socioeconomic factors are the largest predictors of mental health outcomes among veterans, dwarfing the contribution of location and demographic-specific features. Intuitively, because a significant amount of time is allocated toward work activities, the absence of purpose and self-efficacy in the workplace, especially after coming from a mission-driven environment in the armed services, will impact veterans' mental health.

AI can be part of the solution. To the extent that veteran records from combat are combined with self-assessments of skills and career preferences, and these data could be comprehensively gathered and harmonized, researchers could use methods from AI to provide veterans with personalized recommendations regarding not only potential job fits but also counseling over the course of their careers. One of the sources of low engagement among employees is a feeling of plateauing and helplessness; therefore, AI-driven recommendations regarding how to optimize career mobility and human capital development would provide veterans with actionable steps to continuously acquire and apply new skills at work.

Another prime example involves personalizing feedback to veterans about how to live healthier lives. End-of-life care is one of the largest sources of health care expenditures. For example, Riley and Lubitz [12] estimate that a quarter of all Medicare spending goes toward care for people during their last year of life. These resources could be more impactful if they were allocated more toward preventative care earlier in life. Using deep learning methods, Ahadi et al [13] illustrate how biological data can be used for longitudinal profiling. Implementing this algorithm, combined with EHRs at the VA, offers the potential to provide practical advice about how to live more productive and happier lives, raising both economic and social well-being.

Veterans in rural areas face challenges accessing care due to a paucity of rural treatment facilities. AI, implemented along with smart devices (eg, smart wearables), could allow for remote monitoring of rural veterans' health and enable smart devices to alert veterans of health concerns. Recent evidence indicates that AI may be able to predict a person's mental state, including the likelihood of suicide, raising the likelihood that smart devices could be used for predicting and intervening in veteran suicide [7,8].

However, the benefits of AI depend on ethical implementation. Risks associated with AI implementation need to be thoroughly assessed and managed. If, for example, privacy is disrespected, public trust and confidence, particularly among those who have already sacrificed so much for their country, would be undermined. This is extremely important at the VA, where sensitive data, which is under continuous reassessment and review, is routinely collected from veterans. Moreover, researchers must be cognizant of the potential for replicating sources of bias when training their AI algorithms. That is, researchers must investigate the data and model to, at least qualitatively, assess whether there are potential biases that could lead to error replication through the AI-driven recommendations. For example, one possibility is that samples are not representative of the entire population of veterans [14], particularly those who do not feel comfortable using technology. Researchers must also ensure that AI-driven insights are derived from representative samples that reflect the diversity of experiences, attitudes, ethnic, and gender composition among veterans. Recent evidence, for example, highlights the lack of diversity in many health care databases as a major limitation [15].

2. Accurate, Reliable, and Effective

> *…ensure that their application of AI occurs in well-defined domains, and is accurate, reliable, effective, and fit for intended purposes* [9]

The VA is well-equipped to ensure the accuracy, reliability, and effectiveness of AI applications in health and well-being. The VA has collected and catalogued over two petabytes of data, including data on veteran health, prescription data, and inpatient and outpatient services, among others. Further, the VA established the Million Veteran Program, which characterizes, through a consented cohort of subjects, the confluence of genes, lifestyle, and military exposure on veteran health outcomes. This breadth of data paves the way for algorithms that promote personalized medicine based upon life experience and genetic factors. In particular, the plethora of data at the VA can be leveraged to train high-quality algorithms to serve veteran needs.

Concerns have been raised over whether AI algorithms will be effective and generalize beyond the training set originally used to develop machine learning (ML) algorithms [14]. Importantly, the VA's data sets are generated from VA centers across the country and, in principle, data should accurately capture the diverse spectrum of veterans. Therefore, AI algorithms trained on these data should prove to be reliable even when implemented in varied VA centers throughout the United States. However, cautious implementation and monitoring is necessary to ensure that each developed AI algorithm is beneficial at VA centers.

Although the VA database spans millions of veterans, there are still many veterans who are not included in the system. For example, homelessness is a large challenge for the veteran population, and if these veterans are not included within the VA system, they cannot receive the available benefits and treatment [16]. Our internal calculations from the American Community Survey conducted by the Census Bureau suggest that there are roughly 18 million veterans in the United States, whereas the VA only covers roughly 9 million of them [17]. To ensure that AI applications produce reliable recommendations for all veterans, it is important to ensure that the data being fed into predictive models is representative.

In addition to the importance of maintaining a representative sample, researchers and clinicians must use appropriate AI techniques. One particularly large challenge with clinical decision support tools and the use of electronic health records is the presence of missing data and small sample sizes. While sample size is less of a challenge within the VA because of the size of its EHR database, missing data can be a source of bias if they are not missing at random [18]. Some ML techniques, such as gradient boosting and decision trees, can deal well with missing data; however, researchers need to be careful about applying ML and automation in these environments. There is also a well-known bias that can emerge against specific groups, whether by race or even socioeconomic status, which can be propagated at scale if ML algorithmics are not trained and "de-biased" properly [19]. However, it is becoming clear that researchers developing predictive models for clinical use need to transcend traditional conversations about algorithmic bias and think harder about the broader and structural forces that are at play in the observed phenomena [20].

3. Understandable

> *…ensure the operations and outcomes of their AI applications are sufficiently interpretable and understandable by subject matter experts, users, and others as appropriate* [9]

A concern for AI development is the necessity for algorithms to be explainable. Explainability is the concept that users should be able to understand how algorithms function, and it is conceptualized along a continuum where relatively simple algorithms based upon branching decision trees and linear regression are feasible to understand [21]. However, the use of deep neural networks (DNNs), where decision-making is spread across multiple layers of interconnected decision-making nodes, currently produces results that are difficult to accurately interpret. Although DNNs provide great utility in analyzing complex data sets, there is concern over the "black box" nature of DNNs, although new methods are being developed to provide explainability to DNNs [22]. Explainable algorithms will foster trust in AI by both clinicians and patients at the VA.

These new principles for the promotion of trustworthy AI build upon an existing framework developed in the VA, Ethical Principles for Access to and Use of Veteran Data [23], that safeguards veterans and their data and ensures that veterans benefit from research. In other words, research is not an end in and of itself—it is a means toward delivering value to veterans. Moreover, these principles are rooted in the legacy of the Belmont Report from 1979 [5], which emphasized privacy, beneficence, and justice in applications of technology. At its root, technology exists to help improve well-being, whether through heightened productivity or quality of the services provided. Together, these principles provide a signpost for clinicians and researchers to work collaboratively so that AI is

developed and deployed for social good, especially for vulnerable groups and veterans.

Moreover, these ethical principles developed and operationalized within the VA can be extended across the broader health care sector. For example, large university hospital systems that exist within the research ecosystem can adopt these ethical principles to guide their strategic investments and the development and deployment of AI tools. In fact, these university ecosystems have many similarities to the VA because they bring together a combination of researchers and clinicians under a common umbrella and institutional resources. Researchers and clinicians can work hand-in-hand to ensure that research and development investments are fundamentally driven by areas of great need and potential impact.

These processes for the development and application of ethical AI extend beyond veterans. In particular, members of any vulnerable group are beneficiaries of adherence to these processes because, by definition, they may find it harder to benefit from AI. For example, while AI is also leading to the invention of new jobs and tasks in the labor market, AI also reduces the demand for other skills that are more routine and manual, which may affect veterans more if they are concentrated in those types of jobs and occupations. In this sense, applications of AI aimed at improving the transition of service members into the civilian sector could not only help veterans directly by, for example, providing them with tools to more efficiently match into jobs that suit their preferences and abilities, but could also improve trust and confidence in the benefits of AI. Moreover, other vulnerable groups likely face similar challenges; therefore, processes for the development and application of AI would help them too.

Our paper explains some of the most important ingredients for ensuring that AI advances are applied in ways that promote improved veteran outcomes. Furthermore, the VA could serve as a model organization, protecting VA patient data and leveraging it for their good and ultimately cutting health care costs, increasing efficiency, and enhancing health care for veterans. If the United States can successfully scale AI under a technology governance structure using these principles, the possibilities are limitless.

Conflicts of Interest

None declared.

References

1. Yu K, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018 Oct;2(10):719-731. [doi: 10.1038/s41551-018-0305-z] [Medline: 31015651]
2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan 7;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
3. Borkowski AA, Wilson CP, Borkowski SA, Thomas LB, Deland LA, Grewe SJ, et al. Comparing artificial intelligence platforms for histopathologic cancer diagnosis. Fed Pract 2019 Oct;36(10):456-463 [FREE Full text] [Medline: 31768096]
4. Overview of VA research on Health Equity. US Department of Veterans Affairs, Office of Research & Development. URL: https://www.research.va.gov/topics/health_equity.cfm [accessed 2021-05-25]
5. The Belmont Report. US Department of Health and Human Services. 1979. URL: https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html [accessed 2021-04-01]
6. Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature 2019 Jul 31;572(7767):116-119. [doi: 10.1038/s41586-019-1390-1]
7. Badal VD, Graham SA, Depp CA, Shinkawa K, Yamada Y, Palinkas LA, et al. Prediction of loneliness in older adults using natural language processing: exploring sex differences in speech. Am J Geriatr Psychiatry 2020 Sep 12:1-14 [FREE Full text] [doi: 10.1016/j.jagp.2020.09.009] [Medline: 33039266]
8. Fonseka TM, Bhat V, Kennedy SH. The utility of artificial intelligence in suicide risk prediction and the management of suicidal behaviors. Aust N Z J Psychiatry 2019 Oct 26;53(10):954-964. [doi: 10.1177/0004867419864428] [Medline: 31347389]
9. Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government. Executive Office of the President of the United States. URL: https://www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government [accessed 2021-05-25]
10. Makridis C, Hirsch B. Labor Market Earnings of Veterans: Is Time in the Military More Valuable or Less than is Civilian Experience? Social Sciences Research Network Working Paper. 2019 Oct 18. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3466518 [accessed 2021-05-25]
11. Makridis C, Zhao D, Bejan C, Alterovitz G. Leveraging machine learning to characterize the role of socio-economic determinants on physical health and well-being among veterans. SSRN Preprint posted online on October 19, 2020. [doi: 10.2139/ssrn.3686845]
12. Riley GF, Lubitz JD. Long-term trends in Medicare payments in the last year of life. Health Serv Res 2010 Apr;45(2):565-576 [FREE Full text] [doi: 10.1111/j.1475-6773.2010.01082.x] [Medline: 20148984]
13. Ahadi S, Zhou W, Schüssler-Fiorenza Rose SM, Sailani MR, Contrepois K, Avina M, et al. Personal aging markers and ageotypes revealed by deep longitudinal profiling. Nat Med 2020 Jan 13;26(1):83-90. [doi: 10.1038/s41591-019-0719-5]

14. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. Lancet Digit Health 2020 Sep;2(9):e489-e492. [doi: 10.1016/s2589-7500(20)30186-2]
15. Kaushal A, Altman R, Langlotz C. Geographic distribution of US cohorts used to train deep learning algorithms. JAMA 2020 Sep 22;324(12):1212-1213 [FREE Full text] [doi: 10.1001/jama.2020.12067] [Medline: 32960230]
16. Peterson R, Gundlapalli AV, Metraux S, Carter ME, Palmer M, Redd A, et al. Identifying Homelessness among Veterans Using VA Administrative Data: Opportunities to Expand Detection Criteria. PLoS ONE 2015 Jul 14;10(7):e0132664 [FREE Full text] [doi: 10.1371/journal.pone.0132664]
17. Department of Veterans Affairs. 2021. Veterans Health Administration (VHA). URL: https://www.va.gov/health/aboutvha.asp [accessed 2021-01-04]
18. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. JAMA Internal Medicine 2018 Nov 01;178(11):1544-1547. [doi: 10.1001/jamainternmed.2018.3763]
19. Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. JAMA 2019 Dec 24;322(24):2377-2378. [doi: 10.1001/jama.2019.18058] [Medline: 31755905]
20. Pfohl SR, Foryciarz A, Shah NH. An empirical characterization of fair machine learning for clinical risk prediction. Journal of Biomedical Informatics 2021 Jan;113:103621. [doi: 10.1016/j.jbi.2020.103621]
21. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inform Fusion 2020 Jun;58:82-115. [doi: 10.1016/j.inffus.2019.12.012]
22. Xie N, Ras G, van Gerven M, Doran D. Explainable deep learning: a field guide for the uninitiated. ArXiv Preprint posted online on April 30, 2020 [FREE Full text]
23. The DOVA. Ethical principles for access to and use of veteran data. Department of Veterans Affairs. 2020. URL: https://vaww.oit.va.gov/oit/office-technical-integration/ethical-data-use/ [accessed 2021-01-04]

Abbreviations

AI: artificial intelligence
DNN: deep neural network
EHR: electronic health record
ML: machine learning
VA: Department of Veterans Affairs

Edited by C Lovis; submitted 18.03.21; peer-reviewed by J Liew, A Amritphale; comments to author 03.04.21; revised version received 23.04.21; accepted 27.04.21; published 02.06.21.

Please cite as:
Makridis C, Hurley S, Klote M, Alterovitz G
Ethical Applications of Artificial Intelligence: Evidence From Health Research on Veterans
JMIR Med Inform 2021;9(6):e28921
URL: https://medinform.jmir.org/2021/6/e28921
doi:10.2196/28921
PMID:34076584

Original Paper

Intention to Use Wiki-Based Knowledge Tools: Survey of Quebec Emergency Health Professionals

Patrick Archambault[1,2,3], MD, MSc; Stéphane Turcotte[4], MSc; Pascal Y Smith[4], PhD; Kassim Said Abasse[5*], PhD; Catherine Paquet[6], PhD; André Côté[5], PhD; Dario Gomez[7], MBA; Hager Khechine[7], MBA, PhD; Marie-Pierre Gagnon[3,8], PhD; Melissa Tremblay[9], MD; Nicolas Elazhary[9], MD; France Légaré[2,3], MD, PhD; Wiki-Based Knowledge Tool Investigators[10*]

[1]Département de médecine d'urgence, Centre intégré de santé et de services sociaux de Chaudière-Appalaches, Lévis, QC, Canada
[2]Department of Family Medicine and Emergency Medicine, Faculty of Medicine, Université Laval, Québec, QC, Canada
[3]VITAM - Centre de recherche en santé durable, Université Laval, Québec, QC, Canada
[4]Centre intégré de santé et de services sociaux de Chaudière-Appalaches, Lévis, QC, Canada
[5]Département de management, Faculté des sciences de l'administration, Université Laval, Québec, QC, Canada
[6]Département de marketing, Faculté des sciences de l'administration, Université Laval, Québec, QC, Canada
[7]Département de systèmes d'information organisationnels, Faculté des sciences de l'administration, Université Laval, Québec, QC, Canada
[8]Faculté des sciences infirmières, Université Laval, Québec, QC, Canada
[9]Department of Family Medicine and Emergency Medicine, Faculty of Medicine, Université de Sherbrooke, Sherbrooke, QC, Canada
[10]see Acknowledgments
[*]these authors contributed equally

Corresponding Author:
Patrick Archambault, MD, MSc
Département de médecine d'urgence
Centre intégré de santé et de services sociaux de Chaudière-Appalaches
143 rue Wolfe
Lévis, QC, G6V3Z1
Canada
Phone: 1 418 835 7121 ext 13907
Email: patrick.archambault@fmed.ulaval.ca

Abstract

Background: Clinical decision support systems are information technologies that assist clinicians in making better decisions. Their adoption has been limited because their content is difficult to adapt to local contexts and slow to adapt to emerging evidence. Collaborative writing applications such as wikis have the potential to increase access to existing and emerging evidence-based knowledge at the point of care, standardize emergency clinical decision making, and quickly adapt this knowledge to local contexts. However, little is known about the factors influencing health professionals' use of wiki-based knowledge tools.

Objective: This study aims to measure emergency physicians' (EPs) and other acute care health professionals' (ACHPs) intentions to use wiki-based knowledge tools in trauma care and identify determinants of this intention that can be used in future theory-based interventions for promoting the use of wiki-based knowledge tools in trauma care.

Methods: In total, 266 EPs and 907 ACHPs (nurses, respiratory therapists, and pharmacists) from 12 Quebec trauma centers were asked to answer a survey based on the theory of planned behavior (TPB). The TPB constructs were measured using a 7-point Likert scale. Descriptive statistics and Pearson correlations between the TPB constructs and intention were calculated. Multiple linear regression analysis was conducted to identify the salient beliefs.

Results: Among the eligible participants, 57.1% (152/266) of EPs and 31.9% (290/907) of ACHPs completed the questionnaire. For EPs, we found that attitude, perceived behavioral control (PBC), and subjective norm (SN) were significant determinants of the intention to use wiki-based knowledge tools and explained 62% of its variance. None of the sociodemographic variables were related to EPs' intentions to use wiki-based knowledge tools. The regression model identified two normative beliefs ("approval by physicians" and "approval by patients") and two behavioral beliefs ("refreshes my memory" and "reduces errors"). For ACHPs, attitude, PBC, SN, and two sociodemographic variables (profession and the previous personal use of a wiki) were significantly

related to the intention to use wiki-based knowledge tools and explained 60% of the variance in behavioral intention. The final regression model for ACHPs included two normative beliefs ("approval by the hospital trauma team" and "people less comfortable with information technology"), one control belief ("time constraints"), and one behavioral belief ("access to evidence").

Conclusions: The intentions of EPs and ACHPs to use wiki-based knowledge tools to promote best practices in trauma care can be predicted in part by attitude, SN, and PBC. We also identified salient beliefs that future theory-based interventions should promote for the use of wiki-based knowledge tools in trauma care. These interventions will address the barriers to using wiki-based knowledge tools, find ways to ensure the quality of their content, foster contributions, and support the exploration of wiki-based knowledge tools as potential effective knowledge translation tools in trauma care.

(JMIR Med Inform 2021;9(6):e24649) doi:10.2196/24649

KEYWORDS

knowledge management; knowledge translation; implementation science; collaborative writing applications; wikis; trauma care

Introduction

Background

Emergency physicians (EPs) and other acute care health professionals (ACHPs), such as nurses, respiratory therapists, and pharmacists, working in fast-paced emergency departments (EDs) rely on heuristic clinical reasoning that can falter and lead to unconscious acts of omission and contribute to medical errors [1-4]. Overuse of diagnostic modalities has also become a major challenge, which exposes patients to unwarranted tests and procedures [5]. Clinical decision support systems (CDSSs) are health information technologies that have been proposed as solutions to assist clinicians in making better decisions [6]. These technologies are of great importance for knowledge management, organizational learning, and knowledge-building purposes in ways that allow decision making to be more productive, agile, innovative, and reputable [7]. Systematic reviews have found that CDSS can help professionals in implementing best practices [8,9] and be effective in promoting changes in a variety of clinical areas and environments [10-14]. CDSS may also reduce health care professionals' cognitive load in stressful high-intensity situations, increase access to evidence-based information at the point of care, and standardize emergency clinical decision making [9,13,15]. However, CDSSs have not been universally adopted because of the perceptions of clinicians and administrators that they are expensive, lack usability, and that their content is difficult to adapt to local context [6,16-23].

Wikis can be an innovative component of a CDSS, which may support their implementation by addressing local adaptability issues and costs [24]. Wikis are collaborative writing technologies [25] that allow the creation of interactive, rapidly expanding, and low-cost knowledge databases [22,26]. Wikis allow people not only to consume content but also to produce and edit knowledge [27,28]. In the health care context, wikis (eg, WikEM [29] and Canadian Computerized Provider Order Entry Toolkit [30]) allow knowledge users (eg, physicians and administrators) to create and maintain a knowledge base that can quickly adapt to the local context at a low cost [26,31]. Wikis offer several advantages, including an immediate access to new or updated knowledge and interinstitutional integration [10-14,26]. As such, a wiki can act as the organizational memory of learning organizations where multiple interprofessional stakeholders can create, update, and share knowledge that promotes best practices [1,26,31-33]. This knowledge can take the form of explicit knowledge tools (eg, protocols, order sets, reminders, care pathways, and decision aids) created to support decision making by clinicians and patients based on the best evidence available from rigorous clinical practice guidelines and systematic reviews [34-37]. Relying on wiki capacities to manage knowledge, some health organizations have begun using wiki-based knowledge tools to support the implementation of best practices [19,25,38-44]. Given the potential of wiki-based knowledge tools to improve clinical practice, it is important to understand the factors that would contribute to their uptake by health care professionals.

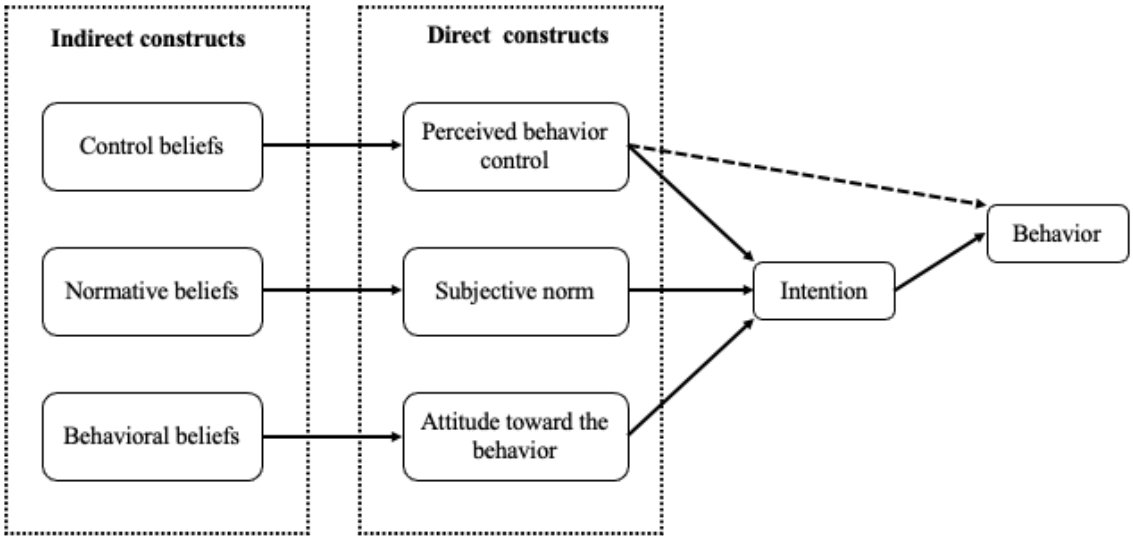
Conceptual Framework

The theory of planned behavior (TPB; Figure 1) has been successfully applied [9,45-47] to study a wide range of health care professionals' behaviors. A recent systematic review has shown that internet-based interventions based on the TPB tend to exert substantial effects on behavior [9]. According to Ajzen [48], the adoption of a new behavior is predicted by the person's intention to engage in that behavior. Intention depends on three main behavioral determinants (direct constructs)—attitude, subjective norm (SN), and perceived behavioral control (PBC). Ajzen [48] also identifies three types of beliefs (indirect constructs) that may influence behavioral determinants—behavioral, normative, and control beliefs (Figure 1). For example, a clinician's intention to use a wiki-based knowledge tool could be strongly influenced by the barriers to access the wiki in the workplace (control belief), a departmental chief not supporting the use of the wiki (normative belief), or a belief that the wiki will help access up-to-date clinical evidence (behavioral belief) [31]. von Haeften et al [49] affirm that to change an intention (and its corresponding behavior), it is necessary to identify and change the determinants of that intention.

According to the TPB as described above, we hypothesize that we can identify the salient beliefs that determine the EPs' and ACHPs' intention to use wiki-based knowledge tools. Moreover, based on our previous qualitative exploration of EPs and ACHPs beliefs demonstrating different beliefs for each professional group [1], we hypothesized that the salient beliefs influencing the intention to use wiki-based knowledge tools would be different for EPs and ACHPs. Identifying the beliefs that have the strongest influence on EP and ACHP intentions will allow us to build a theory-based intervention specific to each

professional group for promoting the use of wiki-based knowledge tools in trauma centers. The ultimate goal of such an intervention is to improve the quality of care within learning health organizations [1,40].

Figure 1. Theory of planned behavior model.



Methods

Study Design, Setting, Population, and Protocol

We conducted our survey using 2 previously developed and tested TPB questionnaires to evaluate EPs' and ACHPs' intention to use wiki-based knowledge tools [31,50] and report its results using the Checklist for Reporting Results of Internet E-Surveys [51-53] (Multimedia Appendix 1 [50]). These questionnaires were previously developed and tested in French with Quebec EPs and ACHPs and revealed adequate internal consistency and stability over time [31,50]. The TPB questionnaires aimed to identify the behavioral determinants that had the greatest influence on the intention to use wiki-based knowledge tools.

The study was conducted in 12 designated trauma centers [54], including 1 level I, 5 level II, and 6 level III trauma centers in the province of Quebec, Canada. Quebec is Canada's second most populous province [55]. The trauma system in Quebec was launched in 1993 and involves an integrated continuum of care from rural community hospitals to urban trauma centers. This system relies on certified ACHPs and EPs who use standardized care protocols across the province. The trauma center designation levels are revised periodically with on-site visits according to the American College of Surgeons criteria [56]. Trauma care services in Quebec are based on transfer agreements between hospitals and a no-refusal transfer policy [57]. Level I, II, and III centers are designated trauma centers with varying levels of services being provided. Level I trauma centers are large, urban hospitals with 24×7 orthopedic, vascular, neurosurgical, and trauma surgery coverage, along with emergency and specialized intensive care services. Level II trauma centers offer full-time, year-round coverage of orthopedic and general surgeries and run an intensive care unit staffed by full-time certified intensivists and an ED staffed by certified EPs. Level III trauma centers offer full-time, year-round coverage of general surgery and partial coverage of orthopedic surgery; they run an ED staffed by general practitioners. They also have an intensive care unit, but they are not run by full-time certified intensivists [56,57]. For the purposes of this study, participants were EPs (excluding residents and medical students) and certified ACHPs (nurses, respiratory therapists, and pharmacists) involved in caring for patients with trauma. Professionals not involved in emergency trauma care were excluded from the study. We purposefully established our list of 12 participating centers based on their geographic location and trauma level of care to recruit a proportion of trauma centers across the province that would reflect the same province-wide proportion of level I, II, and III centers.

To recruit participants, we sent an email to the head physician, nurse, respiratory therapist, and pharmacist of each ED. We asked them to send our invitations to all their respective department members with a web-based link to an electronic survey (SurveyMonkey [58]). Questionnaires were available only in French. A 2-week reminder to complete the web-based survey was sent in the same way. A final reminder was sent after 4 weeks to all potential participants using a ready-to-print PDF version. In total, 266 EPs and 907 ACHPs from 12 Quebec trauma centers were invited to participate. Participants were offered an incentive to participate by offering the chance to win 1 of the 3 electronic tablets. Data were collected between February 2014 and June 2015.

Before responding to the survey, participants were asked to view a 6-minute video (described elsewhere [50]) about wiki-based knowledge tools in trauma care to help them better understand the behavior being investigated. Briefly, participants were shown 1 of the 4 videos that were created specifically for their profession, demonstrating the use of a wiki-based knowledge tool in a simulated trauma case. After watching the appropriate video, the participants filled out 1 of 2 questionnaires according to their profession: EPs filled out the questionnaire for EPs, whereas nurses, respiratory therapists, and pharmacists filled out the questionnaire for ACHPs.

This study was approved by the Research Ethics Committee at the Centre de Santé et Services Sociaux Alphonse-Desjardins as a multicenter research study and by the local ethics review board of each participating center, under the study protocol number MP-23-2014-222. All ED directors approved our project before sending our survey to their members. Participation in the study was voluntary, and the completion of the electronic and paper survey implied consent for participation. To ensure participant privacy and anonymity, no personal information, including internet protocol addresses, was collected.

Measurements

The EP questionnaire comprised 45 items and the ACHP questionnaire comprised 43 items. Briefly, the questionnaires measured sociodemographic, and direct and indirect TPB constructs, as explained elsewhere [1,45,50]. For both questionnaires, the items were measured on a 7-point Likert scale ranging from 1 to 7 (eg, “strongly disagree” [score of 1] to “strongly agree” [score of 7] with “neither agree nor disagree” at the center [score of 4]). Both questionnaires contained 12 sociodemographic questions (eg, age, gender, profession, years of work experience, and previous experience of wiki use in either professional or personal life) and took approximately 10 minutes to complete. SurveyMonkey automatically collected the data for the web version in an Excel spreadsheet, and the responses were manually entered into a spreadsheet for the paper-based questionnaires.

Data Analysis

Before commencing any statistical analyses, data were visually inspected for outliers and checked for normality. Descriptive statistics (means, SDs, and frequencies) summarized and compared demographic information and TPB variables for EP and ACHP participants. For each TPB construct with more than 2 questionnaire items, missing data on items were imputed by using the mean of the other items. The internal consistency of each TPB construct was verified using Cronbach α coefficients for constructs measured using three questionnaire items.

Bivariate analyses were performed between the outcome variable (intention) and the independent variables (demographic information and TPB constructs) using Pearson correlations and Student two-tailed *t* tests. For each type of participant (EP vs ACHP), we then performed a first linear regression model including only TPB direct constructs. We then used a backward approach to test the model adjustment with demographic variables ($P<.10$) [49]. Then, we calculated the proportion of variance (R^2) explained by the model. Then, to identify significant underlying beliefs, we replaced significant direct constructs (PBC, SN, and attitude) that predicted professionals’ intention to use wiki-based knowledge tools with their associated indirect constructs ("control", "normative", and "behavioral beliefs"). Following a backward approach, we only retained significant beliefs (salient beliefs; $P<.05$). Linear regression assumptions were verified for all models. All analyses were performed using the statistical analysis SAS software (SAS Institute Inc) version 9.4 for Windows.

Results

Flow of Participants and Participants’ Characteristics

The demographic characteristics of the participants are presented in Table 1, and their flowchart is presented in Figure 2. Overall, 57.1% (152/266) of EPs and 31.9% (290/907) of ACHPs responded to our survey from 12 trauma centers (level I, II, and III). Among the 442 participants, 337 (76.2%) were women. Their ages ranged from 21 to 69 years, with a mean of 37 (SD 9) years for EPs and 37 (SD 10) years for ACHPs. Among EPs, 49% (74/151) had a special competence in emergency medicine from the College of Family Physicians of Canada, 7.9% (12/151) were certified in emergency medicine as fellows of the Royal College of Physicians and Surgeons of Canada, and 43% (65/151) had no specific certification in emergency medicine. The 290 ACHPs comprised 3 groups of professionals: 196 (67.6%) were nurses, 61 (21%) were respiratory therapists, and 33 (11.4%) were pharmacists (Table 1).

Table 1. Baseline characteristics of participating emergency physicians and ACHPs[a].

Variables	Emergency physicians	ACHPs
Trauma center level, n (%)		
Level III	39 (25.7)	90 (31)
Level II	87 (57.2)	138 (47.6)
Level I	26 (17.1)	62 (21.4)
Age (years)[b]		
Value, mean (SD)	37 (9)	37 (10)
Value, min-max[c]	25-59	21-69
Clinical experience (years)[b]		
Value, mean (SD)	10 (8)	14 (10)
Gender[b], n (%)		
Women	94 (62.3)	243 (84.1)
Men	57 (37.7)	46 (15.9)
Emergency medicine certification[b], n (%)		
CCFP-EM[d]	74 (49)	N/A[e]
FRCPC[f]	12 (7.9)	N/A
No certification	65 (43)	N/A
ACHPs		
Nurses	N/A	196 (67.6)
Respiratory therapist	N/A	61 (21)
Pharmacist	N/A	33 (11.4)

[a]ACHP: acute care health professional.

[b]Missing data: emergency physicians=1; acute care health professionals=1.

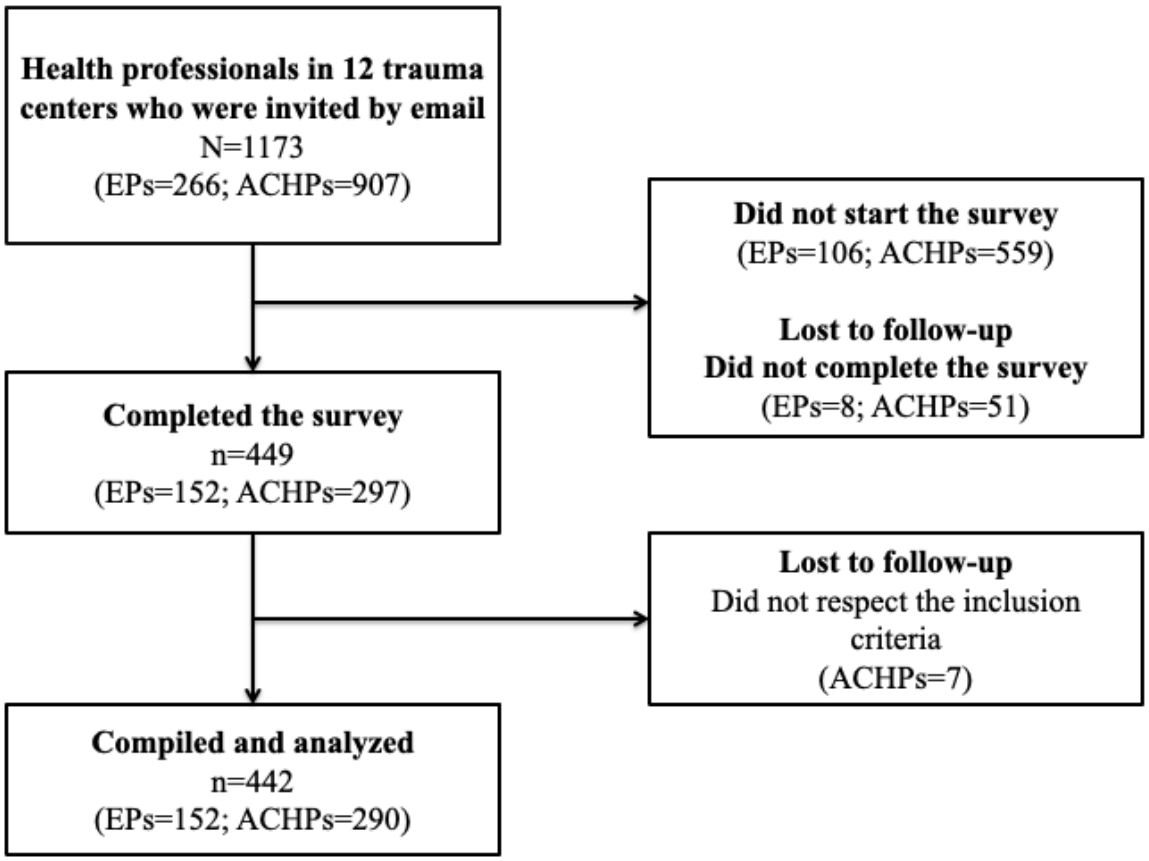
[c]Range.

[d]CCFP-EM: College of Family Physicians of Canada.

[e]N/A: not applicable.

[f]FRCPC: Fellows of the Royal College of Physicians and Surgeons of Canada.

Figure 2. Flowchart of participants of the 12 designated trauma centers. ACHP: acute care health professional; EP: emergency physician.



Descriptive Analysis of the Theoretical Constructs

For EPs, the internal consistency was adequate for all direct TPB constructs (Cronbach α=.76-.90). For ACHPs, the intention and attitude constructs had an appropriate internal consistency (Cronbach α=.85 and Cronbach α=.80, respectively). For PBC and SN constructs, one question was removed from each construct to obtain appropriate internal consistency. The results in Table 2 indicate that participants expressed a high intention (EPs: mean 5.68, SD 1.04; ACHPs: mean 5.49, SD 1.11; on a 7-point Likert scale) to use wiki-based knowledge tools. The PBC (EPs: mean 6.2, SD 0.93; ACHPs: mean 5.85, SD 1.39) was the highest rated direct construct in both groups. In addition, the SN was higher for ACHPs (mean 5.35, SD 1.08) than for EPs (mean 3.65, SD 1.3; *P*<.001).

Table 2. Descriptive analysis of the theoretical variables[a].

Direct construct	Emergency physicians		ACHPs[b]		*P* value
	Value, mean (SD)	Cronbach α	Value, mean (SD)	Cronbach α	
Intention	5.68 (1.04)	.90	5.49 (1.11)	.85	.08
PBC[c]	6.20 (0.93)	.79	5.85 (1.39)	.74	.002
Subjective norm	3.65 (1.3)	.76	5.35 (1.08)	.58	<.001
Attitude	6 (0.89)	.89	5.59 (0.89)	.8	<.001

[a]All scores vary between 1 and 7.

[b]ACHP: acute care health professional.

[c]PBC: perceived behavioral control.

Bivariate and Multivariable Analysis

Results for EPs

The matrix of correlations between direct model variables is presented in Table 3. All independent variables correlated significantly with intention (r=0.33-0.74). On the basis of the strong correlation between age and experience (Pearson correlation; r=0.88), only age was considered in the analysis. Bivariate analyses are presented in Multimedia Appendix 2. Among all demographic variables measured in the questionnaire, three associations with EPs' intention were found to be significant (P<.10). Older EPs had a lower intention to use wiki-based knowledge tools in trauma centers (Pearson correlation; r=−0.14; P=.06) than younger EPs. Similarly, EPs certified as a Fellow of the Royal College of Physicians and Surgeons of Canada had a lower intention (mean 4.86, SD 1.42) to use a wiki-based knowledge tool (P=.02) than EPs without certification (mean 5.71, SD 0.94) or with a College of Family Physicians of Canada certification (mean 5.78, SD 1.01). Previous professional use of wikis was associated with an increased intention (mean 6.03, SD 0.803) in using wiki-based knowledge tools (P=.09).

Table 3. Correlation analysis for emergency physicians and ACHPs[a].

Correlation analysis	Intention	PBC[b]	SN[c]	Attitude
Pearson emergency physicians				
Intention				
r	1	0.43	0.33	0.74
P value	—[d]	<.001	<.001	<.001
PBC				
r	0.43	1	0.02	0.43
P value	<.001	—	.84	<.001
SN				
r	0.33	0.02	1	0.15
P value	<.001	.84	—	.06
Attitude				
r	0.74	0.43	0.15	1
P value	<.001	<.001	.06	—
Pearson ACHPs				
Intention				
r	1	0.46	0.61	0.68
P value	—	<.001	<.001	<.001
PBC				
r	0.46	1	0.31	0.36
P value	<.001	—	<.001	<.001
SN				
r	0.61	0.31	1	0.55
P value	<.001	<.001	—	<.001
Attitude				
r	0.68	0.36	0.55	1
P value	<.001	<.001	<.001	—

[a]ACHP: acute care health professional.

[b]PBC: perceived behavioral control.

[c]SN: subjective norm.

[d]Not applicable.

The linear regression model with the TPB direct constructs and demographic variables indicated that all three direct TPB constructs were associated with the intention to use wiki-based knowledge tools (Table 4). This model, based on TPB direct constructs, explained 62% of the variance in the intention to use wiki-based knowledge tools. Attitude (β=.75) was the most important predictor of EP use of wiki-based knowledge tools

to promote best practices in trauma care. None of the EPs' sociodemographic variables remained significant in this model.

Table 4. Multiple linear regression analysis for emergency physicians and ACHPs[a].

Variable	Estimated value of parameters (SE)	*P* value
Emergency physicians' final TPB[b] model for direct constructs		
Intercept	−0.52 (0.43)	.24
PBC[c]	0.16 (0.06)	.01
SN[d]	0.19 (0.04)	<.001
Attitude	0.75 (0.07)	<.001
ACHPs' final TPB model for direct constructs		
Intercept	−0.30 (0.29)	.30
PBC	0.17 (0.03)	<.001
SN	0.32 (0.05)	<.001
Attitude	0.56 (0.06)	<.001
Profession (respiratory therapist)	−0.42 (0.11)	.001
Profession (pharmacist)	−0.11 (0.14)	.45
Wiki for personal use	0.19 (0.09)	.03

[a]ACHP: acute care health professional.

[b]TPB: theory of planned behavior.

[c]PBC: perceived behavioral control.

[d]SN: subjective norm.

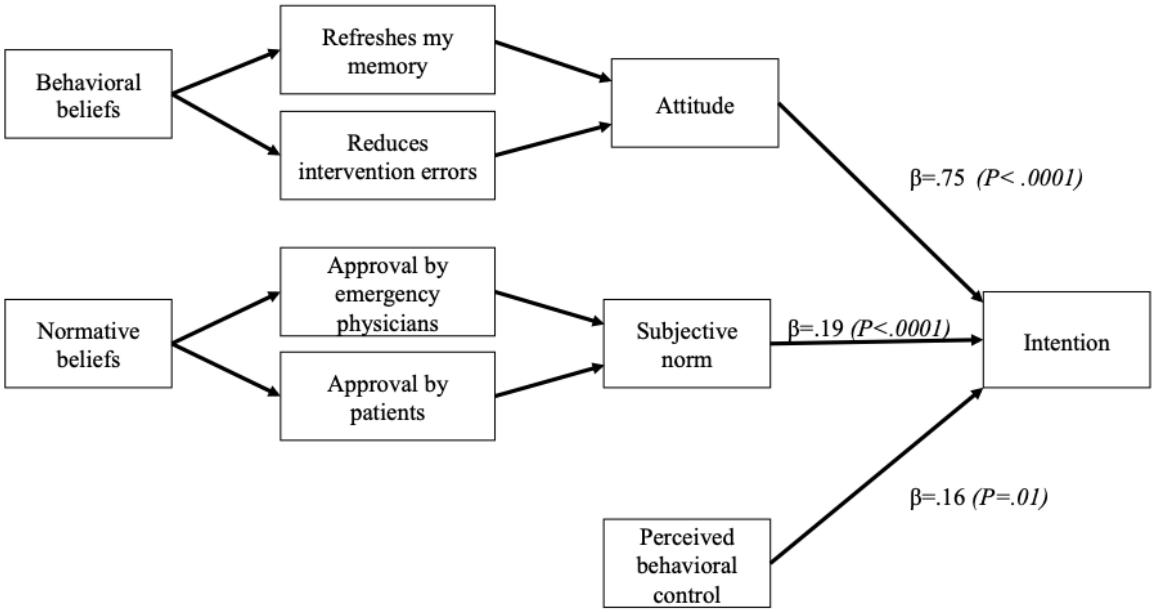
To determine the salient beliefs for predicting EPs' intention to use wiki-based knowledge tools, all significant TPB direct constructs in the first linear regression model were replaced with their associated beliefs in a second regression model. The final model (Table 5) identified significant normative beliefs ("approval from EPs" and "patients") and two behavioral beliefs (wiki-based knowledge tools "refresh my memory" and "reduce intervention errors"; Multimedia Appendices 3 and 4). Figure 3 presents a summary of all the constructs that influence EPs' intention to use wiki-based knowledge tools.

Table 5. Salient belief analysis for both emergency physicians and ACHPs[a].

Variable	Estimated value of parameters (SE)	*P* value
Emergency physicians		
Intercept	−0.92 (0.61)	.13
Support by emergency physicians	0.27 (0.07)	.001
Support by patients	0.19 (0.04)	<.001
Refreshes my memory	0.43 (0.09)	<.001
Reduces intervention errors	0.21 (0.08)	.009
ACHPs		
Intercept	0.80 (0.37)	.03
Time constraints	0.14 (0.03)	.001
Supported by people less comfortable with information technology	0.10 (0.04)	.01
Supported by my hospital trauma team	0.32 (0.05)	<.001
Access to evidence	0.29 (0.05)	<.001

[a]ACHP: acute care health professional.

Figure 3. Emergency physicians' final theory of planned behavior model with direct and indirect constructs (β weights and *P* values in parentheses).



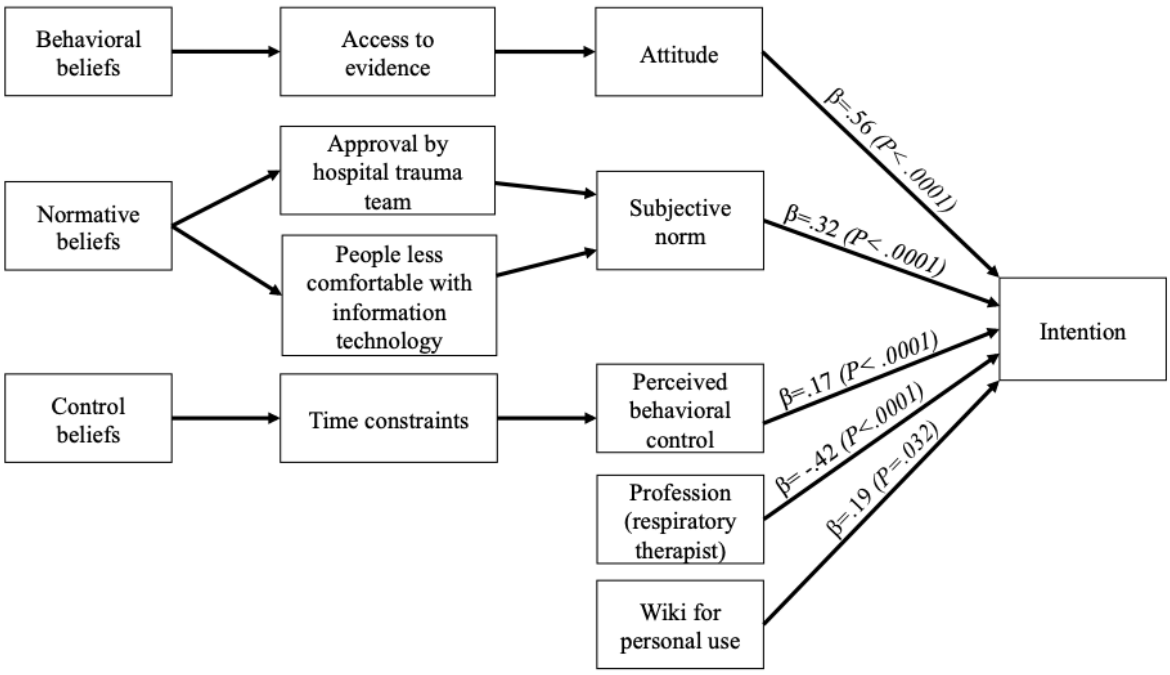
Results for ACHPs

For ACHPs, the matrix of correlations between the direct constructs is shown in Table 3. All independent variables correlated significantly with intention (*r*=0.46-0.68). Correlations between the independent variables were also important (*r*=0.31-0.55). Among all demographic variables measured in the ACHP questionnaire, three significant associations with ACHP intentions were found. ACHPs who did not have access to a computer had a lower intention to use wiki-based knowledge tools than ACHPs with computer access (*P*=.04). Moreover, ACHPs who previously used a wiki in their workplace had a higher intention to use wiki-based knowledge tools (*P*=.009). ACHPs in level I hospitals had a higher intention to use wiki-based knowledge tools than ACHPs in level II and III hospitals (*P*<.001). Otherwise, no significant bivariate associations were found with the type of profession (*P*=.36) or a previous personal use of a wiki (*P*=.13). Bivariate analyses are shown in Multimedia Appendix 5.

The results of the multiple linear regression model using the direct TPB constructs and demographic variables indicated that all three direct constructs were significantly associated with the intention to use wiki-based knowledge tools (PBC: *P*<.001; SN: *P*<.001; attitude: *P*<.001). Two sociodemographic variables remained significant in this model: profession and previous use of a wiki for personal use. The final model is presented in Table 4. This model explains 60% of the variance in the intention to use wiki-based knowledge tools. Attitude (β=.56) was the most important predictor of ACHPs' use of wiki-based knowledge tools to promote best practices in trauma care centers.

To identify the salient beliefs that predict ACHPs' intention to use wiki-based knowledge tools, all significant TPB direct constructs were replaced with their associated indirect constructs (beliefs) in a second linear regression model. The final model obtained using the backward selection approach is presented in Table 5. We found that two normative beliefs ("people less comfortable with information technology" and "my hospital trauma team"), one control belief ("I would use wikis even if I had time constraints"), and one behavioral belief ("If I used a wiki, it would give me access to evidence") were significant, as shown in Multimedia Appendices 3 and 6. Figure 4 presents a summary of all the constructs that influence ACHPs' intention to use wiki-based knowledge tools.

Figure 4. Acute care health professionals' final theory of planned behavior model with direct and indirect constructs (β weights and *P* values in parentheses).



Discussion

Principal Findings

This study identified the salient beliefs in emergency health care professionals (EPs and ACHPs) that can predict the intention to use wiki-based knowledge tools for promoting best practices in trauma care centers. With these results, we can better understand how wiki-based knowledge tools can be used to increase evidence-based practices in trauma care and how to maximize the use and benefits of wiki-based knowledge tools. This will inform the construction of novel educational interventions to address specific beliefs to increase EPs and ACHPs use a wiki-based knowledge tool.

The research reported here provides data from a system-wide survey conducted in a wide range of trauma centers that increases its applicability to promote the implementation of best practices in trauma care across the full range of the trauma continuum. Other theory-based investigations [1,59-62] have been conducted to explore behaviors with respect to contributing to a wiki or using wiki content in contexts other than quality improvement in health care, but this study identifies the specific behavioral determinants needed to address in the context of a health care wiki-based quality improvement intervention. Overall, the intention to use wiki-based reminders to support best practice implementation was high for both EPs and ACHPs. These findings are similar to those reported by Gupta et al [41] and Wright et al [19] regarding the use of wikis in the context of the collaborative design of an asthma action plan and the sharing of clinical decision support content with Web 2.0, respectively. Other researchers have found lower expressed intentions to use wiki-based information in various contexts [63,64]. A randomized controlled trial comparing an in-person nominal group approach with an internet-based wiki-inspired alternative for engaging stakeholders in chronic kidney disease research prioritization identified a low correlation in rankings as compared with the wiki groups, with less satisfaction and perceptions of active engagement [64]. We believe that our positive results regarding EPs' and ACHPs' intentions to use our wiki model probably reflect our participants' trust in the expert-created content model we proposed in our videos as opposed to a model of layperson-created content such as Wikipedia.

In our study, age, gender, years of experience, access to a computer with internet, the frequency of using another professional wiki, previous wiki edition experience, or trauma committee membership did not have any influence on either EPs' or ACHPs' intention to use wiki-based knowledge tools for promoting trauma care best practices. We found that the ACHP profession type was related to the intention to use wiki-based knowledge tools with pharmacists and respiratory therapists, both having a lower intention to use wiki-based knowledge tools compared with nurses. Conversely, the previous use of a wiki for personal reasons increased the ACHPs' intention to use a wiki-based knowledge tool. Our analysis showed that the level of the trauma center did not influence the intention to use wiki-based knowledge tools.

We also found that ACHPs were a heterogeneous group and had different behavioral determinants toward using wiki-based knowledge tools. The ACHPs were nurses, respiratory therapists, and pharmacists, all of whom had different clinical tasks. We suggest that future studies should consider the particularities of each profession. We have demonstrated that both EPs and ACHPs have a good perception of their ability (PBC) to use a wiki-based knowledge tool. In other words, in general, they feel confident that they will be able to use this type of technology. However, our salient belief analysis showed that some ACHP

subgroups feel less comfortable with information technology. Other studies have also shown that certain health professionals such as nurses express the need for educational programs to enhance their level of comfort with information technology [65-67] and with wiki technology [25] in particular.

ACHPs also perceived time constraints as a potential barrier to the use of wiki-based knowledge tools. Although time constraint was not a salient belief for EPs in our study, this contrasts with earlier studies that have identified time constraints as an important control belief in technology adoption [9,45,68] and in other contexts as well [9,69] for EPs and ACHPs alike. Given the tight time constraints associated with trauma care, ACHPs appear to appreciate brevity and efficiency [13,15]. Although our study did not show time constraints as a significant salient belief for EPs, we do not believe EPs will differ from ACHPs in this aspect based on previous studies [25]. Consequently, interventions targeting these control beliefs will most likely need to be oriented toward showing the efficiency of using wiki-based knowledge tools to improve trauma care decision making for EPs and ACHPs alike.

EPs and ACHPs are also more likely to engage in using wiki-based knowledge tools if they know that using such tools will refresh their memory, give them access to evidence-based knowledge tools, and reduce intervention errors. Consequently, educational interventions targeting these behavioral beliefs will have to show that using a wiki-based knowledge tool can help EPs and ACHPs reduce medical errors and remind them about the best evidence to use [9,27,69]. Although our results indicate that EPs feel less social pressure to use wiki-based knowledge tools than ACHPs, both EPs and ACHPs are both more likely to engage in using wiki-based knowledge tools if they feel supported by their colleagues and their patients. Therefore, we could develop common behavioral change techniques that support the collaborative use of wiki-based knowledge tools, interprofessional communication, and local champions to lead the implementation of wiki-based reminders promoting practice change. Considering the value EPs place in support from patients, involving patient partners could also support using a wiki-based reminder system. The existing recommendations for patient-oriented research could help in engaging patients and clinicians in a collaborative quality improvement platform [70].

Our results also indicate that both ACHPs and EPs share the need for support from their peers (other EPs and trauma teams). This means that a common intervention targeting both EPs and ACHPs in trauma teams could improve the use of wiki-based knowledge tools. Interprofessional collaboration has been proposed as an important facilitator in the implementation of best trauma care practices [71,72].

This study adds to the understanding of using wiki-based knowledge tools to support the implementation of best practices in trauma care by using the TPB. In terms of the significance of the variables, our results are similar to those presented in previous studies that identified barriers and facilitators. For example, others have shown that the scientific quality of information resources [16,45] influences their use. We also found that wiki-based knowledge tool use will also be influenced by access to high-level evidence (ACHPs) and potentially reduce intervention errors (EPs). The analytical strategy used in this study provides scientific evidence to identify the most important determinants of EPs' and ACHPs' intentions to design an intervention aimed at promoting the use of wiki-based knowledge tools. We found that EPs' and ACHPs' intention to use wiki-based knowledge tools can be predicted by the three direct TPB constructs—attitude toward the behavior, SN, and PBC. We have also identified the salient beliefs that will help us develop a theory-based training program to promote the use of wiki-based knowledge tools in trauma care centers for EPs and ACHPs [40,73]. These salient beliefs will also inform the development of interventions that support the implementation of future wiki-based knowledge tools for other acute care contexts, such as optimal ED elder care [74] and pandemic knowledge management [75].

Limitations

This study has several limitations. First, the principal limitation of our study is not being able to measure the actual behavior. This is a preliminary study that will help us construct a wiki system containing knowledge tools to promote best practices in trauma care that will consider all the identified behavioral determinants [12,45]. According to the TPB, intention is assumed to be an immediate antecedent of behavior, and measures of behavioral intention are frequently used as a proxy for actual behavior [45].

Second, this study was conducted in 12 publicly funded health organizations in the province of Quebec, a French-speaking region of Canada. Thus, the results may not be generalizable to other types of organizations and other settings. However, given the strong predictive power of the theoretical model, we believe that our approach can inform similar studies in other locations.

Third, we did not separate specific beliefs for each ACHP category. However, our results suggest that ACHP characteristics need to be considered while evaluating the intention to use wiki-based knowledge tools. We suggest that future studies should consider the particularities of each type of health professional. Finally, there are other limitations related to our survey methodology. Our study involved voluntary participation, which may have introduced a selection bias. Study participants may have had more experience or a stronger intention to use wiki-based knowledge tool than nonparticipants. For this reason, it is possible that a social desirability bias positively influenced our results. Moreover, this survey was conducted in 2014 and 2015. Although this does not affect the internal validity of our results, it might potentially affect the applicability of the paper in today's context as technology and its acceptance may have evolved. Furthermore, our linear regression model for ACHPs seems to be affected by two variables (profession and the previous use of a wiki for personal use) with a small confounding effect. Unbalanced data between categorical modalities of these two variables may have attenuated the true relation with intention in bivariate analyses.

Conclusions

This study allows us to better understand how a wiki-based knowledge tool can be used to increase evidence-based practices

and maximize their benefits. This will be useful in constructing an implementation intervention that supports the best practices in trauma care. This study contributes to knowledge translation and organizational learning by proposing a strong theoretical basis to assess the determinants of using wiki-based knowledge tools in trauma care centers. Future studies are needed to assess the impact of using wiki-based knowledge tools on health care professionals' knowledge, attitudes, skills, and behaviors in practice as well as to address the barriers to their use, to find ways to ensure the quality of their content, to foster contributions, and to make these tools effective knowledge translation tools for different stakeholders.

Acknowledgments

Funding for this project was provided by the Canadian Institutes for Health Research (Knowledge Synthesis Grant, FRN116632); Knowledge Translation Canada; Fondation de l'Hôtel-Dieu de Lévis; and Canadian Foundation for Healthcare Improvement, and a research grant was provided from the Département de médecine familiale et médecine d'urgence de l'Université Laval. PA is the recipient of a clinical scholar award from the Fonds de Recherche du Québec-Santé. PMA was also the recipient of a Canadian Institutes of Health Research Embedded Clinician Researcher Award. FL is the Canada Research Chair in Shared Decision Making and Knowledge Translation. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

The authors would like to thank all the Wiki-Based Knowledge Tool investigators who coordinated the local data collection and who made this project possible; Centre intégré de santé et de services sociaux de Chaudière-Appalaches: Audrey Dupuis, Carrie Anna McGinn, Émilie Papillon-Dion, Annie Prévost, Hugo Grenier, Sabrina Chevanel, Sandra McKetchum, Micheline Vigneault, Pierre Faucher; Centre intégré universitaire de santé et de services sociaux du Nord-De-L'île-De-Montréal: Jean-Marc Chauny, Chantal Lanthier; Centre intégré de santé et de services sociaux du Bas-Saint-Laurent: Agnès Pascot, Pierre-Luc Sylvain, Esther Otis, Doris Arbour, Julie Lagacé, Jocelyn Deschênes; Centre intégré de santé et de services sociaux de la Gaspésie: Claudia Plourde, Vincent Tremblay, Diane Henry, Nancy Richard, Sylvain Levac, Marie-Claude Boudreau; Centre intégré de santé et de services sociaux de la Mauricie-et-du-Centre-du-Québec: François Parent, Marcel Rheault; Centre intégré de santé et de services sociaux des Laurentides: Martin Recher, Lucie Dugré, Sylvain Marcil, Karine Sanogo, Daniel Bellemare, Sylvie Côté; Centre intégré universitaire de santé et de services sociaux de l'Estrie - Centre hospitalier universitaire de Sherbrooke: Claudie Gagnon, Mélanie Fauteux.

The authors also gratefully thank Marie-Hélène Savard for her collaboration in reviewing and commenting on our manuscript. The authors also gratefully thank Susie Gagnon for her coordination during this research project. The authors would also like to thank Eddy Lang and Jean Lapointe, who helped obtain funding to conduct this project from KT Canada, as well as all participants in the 12 trauma centers and local research assistants and coordinators.

Authors' Contributions

PA wrote the original protocol and obtained funding. PA, ST, PYS, DG, and KSA led the design, data acquisition, data analysis, and drafting of the first manuscript. CP, AC, DG, HK, MPG, MT, NE, and FL were responsible for revising the manuscript multiple times for methodological, conceptual, and intellectual content. Members of the Wiki-Based Knowledge Tool investigators assisted with data acquisition. All authors read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Checklist for Reporting Results of Internet E-Surveys guideline report.
[DOCX File , 194 KB - medinform_v9i6e24649_app1.docx]

Multimedia Appendix 2
Bivariate analysis for emergency physicians.
[DOCX File , 29 KB - medinform_v9i6e24649_app2.docx]

Multimedia Appendix 3
Emergency physicians' and acute care health professionals' indirect constructs.
[DOCX File , 15 KB - medinform_v9i6e24649_app3.docx]

Multimedia Appendix 4
Emergency physicians' salient belief analysis.

[PPT File (Microsoft PowerPoint Presentation), 60 KB - medinform_v9i6e24649_app4.ppt]

Multimedia Appendix 5
Bivariate analysis for acute care health professionals.
[DOCX File , 33 KB - medinform_v9i6e24649_app5.docx]

Multimedia Appendix 6
Acute care health professionals' salient beliefs' analysis.
[PPT File (Microsoft PowerPoint Presentation), 60 KB - medinform_v9i6e24649_app6.ppt]

References

1. Archambault PM, Bilodeau A, Gagnon M, Aubin K, Lavoie A, Lapointe J, et al. Health care professionals' beliefs about using wiki-based reminders to promote best practices in trauma care. J Med Internet Res 2012 Apr 19;14(2):e49 [FREE Full text] [doi: 10.2196/jmir.1983] [Medline: 22515985]
2. Brehaut JC, Hamm R, Majumdar S, Papa F, Lott A, Lang E. Cognitive and social issues in emergency medicine knowledge translation: a research agenda. Acad Emerg Med 2007 Nov 1;14(11):984-990 [FREE Full text] [doi: 10.1197/j.aem.2007.06.025] [Medline: 17893396]
3. Croskerry P, Nimmo G. Better clinical decision making and reducing diagnostic error. J R Coll Physicians Edinb 2011 Jun;41(2):155-162. [doi: 10.4997/JRCPE.2011.208] [Medline: 21677922]
4. McDonald CJ. Protocol-based computer reminders, the quality of care and the non-perfectability of man. N Engl J Med 1976 Dec 9;295(24):1351-1355. [doi: 10.1056/NEJM197612092952405] [Medline: 988482]
5. Bérubé M, Moore L, Leduc S, Farhat I, Lesieur M, Lamontagne J, et al. Low-value injury care in the adult orthopaedic trauma population: a protocol for a rapid review. BMJ Open 2020 Mar 18;10(3):e033453 [FREE Full text] [doi: 10.1136/bmjopen-2019-033453] [Medline: 32193261]
6. Pope C, Halford S, Turnbull J, Prichard J, Calestani M, May C. Using computer decision support systems in NHS emergency and urgent care: ethnographic study using normalisation process theory. BMC Health Serv Res 2013 Mar 23;13(1):111 [FREE Full text] [doi: 10.1186/1472-6963-13-111] [Medline: 23522021]
7. Holroyd BR, Bullard MJ, Graham TA, Rowe BH. Decision support technology in knowledge translation. Acad Emerg Med 2007 Nov 1;14(11):942-948 [FREE Full text] [doi: 10.1197/j.aem.2007.06.023] [Medline: 17766733]
8. Jaspers MW, Smeulers M, Vermeulen H, Peute LW. Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. J Am Med Inform Assoc 2011 May 1;18(3):327-334 [FREE Full text] [doi: 10.1136/amiajnl-2011-000094] [Medline: 21422100]
9. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. Br Med J 2005 Apr 2;330(7494):765 [FREE Full text] [doi: 10.1136/bmj.38398.500764.8F] [Medline: 15767266]
10. Balas E, Weingarten S, Garb C, Blumenthal D, Boren S, Brown G. Improving preventive care by prompting physicians. Arch Intern Med 2000 Feb 14;160(3):301-308. [doi: 10.1001/archinte.160.3.301] [Medline: 10668831]
11. Buntinx F, Winkens R, Grol R, Knottnerus JA. Influencing diagnostic and preventive performance in ambulatory care by feedback and reminders. A review. Fam Pract 1993 Jun;10(2):219-228. [doi: 10.1093/fampra/10.2.219] [Medline: 8359615]
12. Mandelblatt J, Kanetsky PA. Effectiveness of interventions to enhance physician screening for breast cancer. J Fam Pract 1995 Feb;40(2):162-171. [Medline: 7654272]
13. Sequist TD, Gandhi TK, Karson AS, Fiskio JM, Bugbee D, Sperling M, et al. A randomized trial of electronic clinical reminders to improve quality of care for diabetes and coronary artery disease. J Am Med Inform Assoc 2005;12(4):431-437 [FREE Full text] [doi: 10.1197/jamia.M1788] [Medline: 15802479]
14. Wensing M, Vedsted P, Kersnik J, Peersman W, Klingenberg A, Hearnshaw H, et al. Patient satisfaction with availability of general practice: an international comparison. Int J Qual Health Care 2002 Apr;14(2):111-118. [doi: 10.1093/oxfordjournals.intqhc.a002597] [Medline: 11954680]
15. Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG. Epidemiology and reporting characteristics of systematic reviews. PLoS Med 2007 Mar 27;4(3):e78 [FREE Full text] [doi: 10.1371/journal.pmed.0040078] [Medline: 17388659]
16. Haug P, Gardner R, Evans R, Rocha B, Rocha R. Clinical Decision Support at Intermountain Healthcare. New York, USA: Springer International Publishing; 2007.
17. Bullard MJ, Emond SD, Graham TA, Ho K, Holroyd BR. Informatics and knowledge translation. Acad Emerg Med 2007 Nov 1;14(11):996-1002 [FREE Full text] [doi: 10.1197/j.aem.2007.06.032] [Medline: 17967961]
18. Sahota N, Lloyd R, Ramakrishna A, Mackay JA, Prorok JC, Weise-Kelly L, CCDSS Systematic Review Team. Computerized clinical decision support systems for acute care management: a decision-maker-researcher partnership systematic review of effects on process of care and patient outcomes. Implement Sci 2011 Aug 3;6(1):91 [FREE Full text] [doi: 10.1186/1748-5908-6-91] [Medline: 21824385]

19. Wright A, Bates DW, Middleton B, Hongsermeier T, Kashyap V, Thomas SM, et al. Creating and sharing clinical decision support content with web 2.0: issues and examples. J Biomed Inform 2009 Apr;42(2):334-346 [FREE Full text] [doi: 10.1016/j.jbi.2008.09.003] [Medline: 18935982]
20. Anooj P. Clinical decision support system: risk level prediction of heart disease using weighted fuzzy rules. Int J Res Rev Comput Sci 2011;1(4):482-498. [doi: 10.2478/s13537-011-0032-y]
21. Bright T, Wong A, Dhurjati R, Bristow E, Bastian L, Coeytaux R. Effect of Clinical Decision-Support Systems: A Systematic Review. London, UK: Centre for Reviews and Dissemination; 2012.
22. Parthiban L, Subramanian R. An Intelligent Agent for Detection of Erythemato- Squamous Diseases Using Co-active Neuro-Fuzzy Inference System and Genetic Algorithm. In: International Conference on Intelligent Agent & Multi-Agent Systems. 2009 Presented at: IAMS'09; July 22-24, 2009; Chennai, India. [doi: 10.1109/iama.2009.5228016]
23. Ray J, Ratwani R, Sinsky C, Frankel R, Friedberg M, Powsner S, et al. Six habits of highly successful health information technology: powerful strategies for design and implementation. J Am Med Inform Assoc 2019 Oct 1;26(10):1109-1114 [FREE Full text] [doi: 10.1093/jamia/ocz098] [Medline: 31265064]
24. Lara B, Cañas F, Vidal A, Nadal N, Rius F, Paredes E, et al. Knowledge management through two virtual communities of practice (endobloc and pneumobloc). Health Informatics J 2017 Sep 21;23(3):170-180 [FREE Full text] [doi: 10.1177/1460458216639739] [Medline: 27102887]
25. Archambault PM, van de Belt TH, Grajales FJ, Faber MJ, Kuziemsky CE, Gagnon S, et al. Wikis and collaborative writing applications in health care: a scoping review. J Med Internet Res 2013 Oct 8;15(10):e210 [FREE Full text] [doi: 10.2196/jmir.2787] [Medline: 24103318]
26. Majchrzak A, Wagner C, Yates D. The impact of shaping on knowledge reuse for organizational improvement with wikis. MISQ 2013 Feb 2;37(2):455-469. [doi: 10.25300/misq/2013/37.2.07]
27. Tapscott D, Williams AD. Wikinomics: how mass collaboration changes everything. Choice Rev 2007 Aug 1;44(12):44-6933. [doi: 10.5860/choice.44-6933]
28. Yates D, Wagner C, Majchrzak A. Factors affecting shapers of organizational wikis. J Am Soc Inf Sci 2009. [doi: 10.1002/asi.21266]
29. Donaldson R, Ostermayer D, Banuelos R, Singh M. Development and usage of wiki-based software for point-of-care emergency medical information. J Am Med Inform Assoc 2016 Nov;23(6):1174-1179. [doi: 10.1093/jamia/ocw033] [Medline: 27121610]
30. Theal J, Protti D. CPOE with evidence-based clinical decision support improves patient outcomes: the journey to date for a Canadian hospital. Healthc Q 2014 May 5;17(1):24-29. [doi: 10.12927/hcq.2014.23780] [Medline: 24844717]
31. Archambault PM, Légaré F, Lavoie A, Gagnon M, Lapointe J, St-Jacques S, et al. Healthcare professionals' intentions to use wiki-based reminders to promote best practices in trauma care: a survey protocol. Implement Sci 2010 Jun 11;5(1):45 [FREE Full text] [doi: 10.1186/1748-5908-5-45] [Medline: 20540775]
32. Archambault P, Blouin D, Poitras J, Fountain R, Fleet R, Bilodeau A, et al. Emergency medicine residents' beliefs about contributing to a Google Docs presentation: a survey protocol. Inform Prim Care 2011 Jul 1;19(4):207-216 [FREE Full text] [doi: 10.14236/jhi.v19i4.815] [Medline: 22828575]
33. Graham I, Tetroe J, KT Theories Research Group. Some theoretical underpinnings of knowledge translation. Acad Emerg Med 2007 Nov;14(11):936-941 [FREE Full text] [doi: 10.1197/j.aem.2007.07.004] [Medline: 17967955]
34. Gaddis G, Greenwald P, Huckson S. Toward improved implementation of evidence-based clinical algorithms: clinical practice guidelines, clinical decision rules, and clinical pathways. Acad Emerg Med 2007 Nov;14(11):1015-1022 [FREE Full text] [doi: 10.1197/j.aem.2007.07.010] [Medline: 17967964]
35. Straus S, Tetroe J, Graham I. Knowledge Translation in Health Care || Audit and feedback interventions. New York: John Wiley & Sons; 2009:2013-2018.
36. Shekelle P, Woolf S, Grimshaw JM, Schünemann HJ, Eccles MP. Developing clinical practice guidelines: reviewing, reporting, and publishing guidelines; updating guidelines; and the emerging issues of enhancing guideline implementability and accounting for comorbid conditions in guideline development. Implement Sci 2012 Jul 4;7(1):62 [FREE Full text] [doi: 10.1186/1748-5908-7-62] [Medline: 22762242]
37. Weisz G, Cambrosio A, Keating P, Knaapen L, Schlich T, Tournay V. The emergence of clinical practice guidelines. Milbank Q 2007 Dec;85(4):691-727 [FREE Full text] [doi: 10.1111/j.1468-0009.2007.00505.x] [Medline: 18070334]
38. Archambault PM, Beaupré P, Bégin L, Dupuis A, Côté M, Légaré F. Impact of implementing a wiki to develop structured electronic order sets on physicians' intention to use wiki-based order sets. JMIR Med Inform 2016 May 17;4(2):e18 [FREE Full text] [doi: 10.2196/medinform.4852] [Medline: 27189046]
39. Barondeau G. Understanding wiki collaboration in Quebec healthcare organizations. In: Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration. 2012 Presented at: WikiSym '12; August 2012; Linz, Austria URL: https://dl.acm.org/doi/proceedings/10.1145/2462932 [doi: 10.1145/2462932.2462968]
40. Archambault PM, Turgeon AF, Witteman HO, Lauzier F, Moore L, Lamontagne F, Canadian Critical Care Trials Group. Implementation and evaluation of a wiki involving multiple stakeholders including patients in the promotion of best practices in trauma care: the WikiTrauma interrupted time series protocol. JMIR Res Protoc 2015 Feb 19;4(1):e21 [FREE Full text] [doi: 10.2196/resprot.4024] [Medline: 25699546]

41. Gupta S, Wan FT, Newton D, Bhattacharyya OK, Chignell MH, Straus SE. WikiBuild: a new online collaboration process for multistakeholder tool development and consensus building. J Med Internet Res 2011 Dec 8;13(4):e108 [FREE Full text] [doi: 10.2196/jmir.1833] [Medline: 22155694]
42. van de Belt TH, Faber MJ, Knijnenburg JM, van Duijnhoven NT, Nelen WL, Kremer JA. Wikis to facilitate patient participation in developing information leaflets: first experiences. Inform Health Soc Care 2014 Mar 11;39(2):124-139. [doi: 10.3109/17538157.2013.872107] [Medline: 24517459]
43. Grigori M. Advances in learning software organizations : (Banff, 20-21 June 2004). In: Lecture Notes in Computer Science. Canada: Springer; 2004.
44. Theal J, Protti D. Cpoe with evidence-based clinical decision support improves patient outcomes: part 2--proof from a Canadian hospital. Healthc Q 2014;17(4):68-74. [doi: 10.12927/hcq.2015.24121] [Medline: 25906469]
45. Ajzen I. The theory of planned behaviour: reactions and reflections. Psychol Health 2011 Sep;26(9):1113-1127. [doi: 10.1080/08870446.2011.613995] [Medline: 21929476]
46. Agoritsas T, Heen AF, Brandt L, Alonso-Coello P, Kristiansen A, Akl EA, et al. Decision aids that really promote shared decision making: the pace quickens. Br Med J 2015 Feb 10;350:g7624 [FREE Full text] [doi: 10.1136/bmj.g7624] [Medline: 25670178]
47. Godin G, Bélanger-Gravel A, Eccles M, Grimshaw J. Healthcare professionals' intentions and behaviours: a systematic review of studies based on social cognitive theories. Implement Sci 2008 Jul 16;3(1):36 [FREE Full text] [doi: 10.1186/1748-5908-3-36] [Medline: 18631386]
48. Ajzen I. From Intentions to Actions: A Theory of Planned Behavior. Berlin, Heidelberg: Springer; Jan 22, 1985:11-39.
49. von Haeften I, Fishbein M, Kasprzyk D, Montano D. Analyzing data to obtain information to design targeted interventions. Psychol Health Med 2001 May;6(2):151-164. [doi: 10.1080/13548500125076]
50. Archambault PM, Gagnon S, Gagnon M, Turcotte S, Lapointe J, Fleet R, et al. Development and validation of questionnaires exploring health care professionals' intention to use wiki-based reminders to promote best practices in trauma. JMIR Res Protoc 2014 Oct 3;3(3):e50 [FREE Full text] [doi: 10.2196/resprot.3762] [Medline: 25281856]
51. Eysenbach G. Improving the quality of web surveys: the checklist for reporting results of internet e-surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
52. Burns KE, Duffett M, Kho ME, Meade MO, Adhikari NK, Sinuff T, ACCADEMY Group. A guide for the design and conduct of self-administered surveys of clinicians. Can Med Assoc J 2008 Jul 29;179(3):245-252 [FREE Full text] [doi: 10.1503/cmaj.080372] [Medline: 18663204]
53. Légaré F, Ratté S. Improving the reporting of surveys of clinicians. Can Med Assoc J 2008 Oct 7;179(8):801-802 [FREE Full text] [doi: 10.1503/cmaj.1080103] [Medline: 18838456]
54. Trauma Center Levels Explained. American Trauma Society. URL: https://www.amtrauma.org/page/traumalevels [accessed 2016-09-28]
55. Population and Demography Statistics. Canada Statistics. URL: https://www.statcan.gc.ca/eng/start [accessed 2021-05-20]
56. Nathens AB, Cryer HG, Fildes J. The American College of Surgeons Trauma Quality Improvement Program. Surg Clin North Am 2012 Apr;92(2):441-454. [doi: 10.1016/j.suc.2012.01.003] [Medline: 22414421]
57. Sampalis JS, Denis R, Lavoie A, Fréchette P, Boukas S, Nikolis A, et al. Trauma care regionalization: a process-outcome evaluation. J Trauma 1999 Apr;46(4):565-79; discussion 579. [doi: 10.1097/00005373-199904000-00004] [Medline: 10217218]
58. Free Online Survey Software: Questionnaire Tool. SurveyMonkey. URL: https://www.surveymonkey.com [accessed 2016-09-28]
59. Morley DA. Enhancing networking and proactive learning skills in the first year university experience through the use of wikis. Nurse Educ Today 2012 Apr;32(3):261-266. [doi: 10.1016/j.nedt.2011.03.007] [Medline: 21481500]
60. Culley JM, Polyakova-Norwood V. Synchronous online role play for enhancing community, collaboration, and oral presentation proficiency. Nurs Educ Perspect 2012 Jan;33(1):51-54. [doi: 10.5480/1536-5026-33.1.51] [Medline: 22416543]
61. McGowan BS, Wasko M, Vartabedian BS, Miller RS, Freiherr DD, Abdolrasulnia M. Understanding the factors that influence the adoption and meaningful use of social media by physicians to share medical information. J Med Internet Res 2012 Sep 24;14(5):e117 [FREE Full text] [doi: 10.2196/jmir.2138] [Medline: 23006336]
62. Stutsky B. Empowerment and Leadership Development in an Online Story-based Learning Community. In: Association for the Advancement of Computing in Education. 2009 Presented at: AACE'09; April 11-14, 2009; New York, USA URL: http://learntechlib.org/p/32795/
63. Giordano R. An Investigation of the Use of a Wiki to Support Knowledge Exchange in Public Health. In: Proceedings of the 2007 International ACM Conference on Supporting Group Work. 2007 Presented at: ACM'07; March 1-6, 2007; Sanibel Island, Florida, USA. [doi: 10.1145/1316624.1316664]
64. Elliott MJ, Straus SE, Pannu N, Ahmed SB, Laupacis A, Chong GC, et al. A randomized controlled trial comparing in-person and wiki-inspired nominal group techniques for engaging stakeholders in chronic kidney disease research prioritization. BMC Med Inform Decis Mak 2016 Aug 24;16(1):113 [FREE Full text] [doi: 10.1186/s12911-016-0351-y] [Medline: 27553026]

65. Malo C, Neveu X, Archambault PM, Emond M, Gagnon M. Exploring nurses' intention to use a computerized platform in the resuscitation unit: development and validation of a questionnaire based on the theory of planned behavior. Interact J Med Res 2012 Sep 13;1(2):e5 [FREE Full text] [doi: 10.2196/ijmr.2150] [Medline: 23611903]
66. Asan O, Flynn KE, Azam L, Scanlon MC. Nurses' perceptions of a novel health information technology: a qualitative study in the pediatric intensive care unit. Int J Hum Comput Interact 2017 Jan 11;33(4):258-264 [FREE Full text] [doi: 10.1080/10447318.2017.1279828] [Medline: 31595138]
67. Kuo K, Liu C, Ma C. An investigation of the effect of nurses' technology readiness on the acceptance of mobile electronic medical record systems. BMC Med Inform Decis Mak 2013 Aug 12;13(1):88 [FREE Full text] [doi: 10.1186/1472-6947-13-88] [Medline: 23938040]
68. Côté F, Gagnon J, Houme PK, Abdeljelil AB, Gagnon MP. Using the theory of planned behaviour to predict nurses' intention to integrate research evidence into clinical decision-making. J Adv Nurs 2012 Oct;68(10):2289-2298. [doi: 10.1111/j.1365-2648.2011.05922.x] [Medline: 22229522]
69. Guerrier M, Légaré F, Turcotte S, Labrecque M, Rivest L. Shared decision making does not influence physicians against clinical practice guidelines. PLoS One 2013 Apr 24;8(4):e62537 [FREE Full text] [doi: 10.1371/journal.pone.0062537] [Medline: 23638111]
70. Archambault PM, McGavin C, Dainty KN, McLeod SL, Vaillancourt C, Lee JS, et al. Recommendations for patient engagement in patient-oriented emergency medicine research. Can J Emerg Med 2018 May 25;20(3):435-442. [doi: 10.1017/cem.2018.370] [Medline: 29690943]
71. Nilsson U, Gruen R, Myles P. Postoperative recovery: the importance of the team. Anaesthesia 2020 Jan;75(Suppl 1):e158-e164 [FREE Full text] [doi: 10.1111/anae.14869] [Medline: 31903575]
72. Courtenay M, Nancarrow S, Dawson D. Interprofessional teamwork in the trauma setting: a scoping review. Hum Resour Health 2013 Nov 5;11(1):57 [FREE Full text] [doi: 10.1186/1478-4491-11-57] [Medline: 24188523]
73. Plante P, Angulo Mendoza GA, Archambault P. Analyse, développement et évaluation d'une formation médicale en ligne. Med Med 2019 Nov 15(2):6-28. [doi: 10.52358/mm.vi2.95]
74. Archambault PM, Rivard J, Smith PY, Sinha S, Morin M, LeBlanc A, Network Of Canadian Emergency Researchers. Learning integrated health system to mobilize context-adapted knowledge with a wiki platform to improve the transitions of frail seniors from hospitals and emergency departments to the community (learning wisdom): protocol for a mixed-methods implementation study. JMIR Res Protoc 2020 Aug 5;9(8):e17363 [FREE Full text] [doi: 10.2196/17363] [Medline: 32755891]
75. Collaborative Writing Applications in Support of Knowledge Translation and Management during Global Pandemics: A Scoping Review Protocol. OSFHome. URL: https://osf.io/dprwa/ [accessed 2021-06-04]

Abbreviations

ACHP: acute care health professional
CDSS: clinical decision support system
ED: emergency department
EP: emergency physician
PBC: perceived behavioral control
SN: subjective norm
TPB: theory of planned behavior

Edited by G Eysenbach; submitted 29.09.20; peer-reviewed by C Jacob, A Brettle, T Chan; comments to author 13.01.21; revised version received 16.02.21; accepted 07.05.21; published 18.06.21.

Please cite as:
Archambault P, Turcotte S, Smith PY, Said Abasse K, Paquet C, Côté A, Gomez D, Khechine H, Gagnon MP, Tremblay M, Elazhary N, Légaré F, Wiki-Based Knowledge Tool Investigators
Intention to Use Wiki-Based Knowledge Tools: Survey of Quebec Emergency Health Professionals
JMIR Med Inform 2021;9(6):e24649
URL: https://medinform.jmir.org/2021/6/e24649
doi:10.2196/24649
PMID:34142977

Original Paper

The Clinical Decision Support System AMPEL for Laboratory Diagnostics: Implementation and Technical Evaluation

Maria Beatriz Walter Costa[1,2*], PhD; Mark Wernsdorfer[1,2*], PhD; Alexander Kehrer[3], MSc; Markus Voigt[4], MSc; Carina Cundius[4], PhD; Martin Federbusch[1]; Felix Eckelt[1]; Johannes Remmler[1], MD; Maria Schmidt[1,2], MSc; Sarah Pehnke[1]; Christiane Gärtner[1,2], MD; Markus Wehner[5], MD; Berend Isermann[1], MD; Heike Richter[5], PhD; Jörg Telle[3], MBA; Thorsten Kaiser[1], MD

[1]Institute of Laboratory Medicine, Clinical Chemistry und Molecular Diagnostics, University of Leipzig Medical Center, Leipzig, Germany
[2]Faculty of Medicine, University of Leipzig, Leipzig, Germany
[3]Xantas AG, Leipzig, Germany
[4]Information Management, University of Leipzig Medical Center, Leipzig, Germany
[5]Muldental Clinics GmbH Non-Profit Company, Hospital Grimma and Wurzen, Grimma, Germany
[*]these authors contributed equally

Corresponding Author:
Maria Beatriz Walter Costa, PhD
Institute of Laboratory Medicine, Clinical Chemistry und Molecular Diagnostics
University of Leipzig Medical Center
Paul-List-Str 13/15
Leipzig, 04103
Germany
Phone: 49 15153157803
Email: bia.walter@gmail.com

Abstract

Background: Laboratory results are of central importance for clinical decision making. The time span between availability and review of results by clinicians is crucial to patient care. Clinical decision support systems (CDSS) are computational tools that can identify critical values automatically and help decrease treatment delay.

Objective: With this work, we aimed to implement and evaluate a CDSS that supports health care professionals and improves patient safety. In addition to our experiences, we also describe its main components in a general manner to make it applicable to a wide range of medical institutions and to empower colleagues to implement a similar system in their facilities.

Methods: Technical requirements must be taken into account before implementing a CDSS that performs laboratory diagnostics (labCDSS). These can be planned within the functional components of a reactive software agent, a computational framework for such a CDSS.

Results: We present AMPEL (Analysis and Reporting System for the Improvement of Patient Safety through Real-Time Integration of Laboratory Findings), a labCDSS that notifies health care professionals if a life-threatening medical condition is detected. We developed and implemented AMPEL at a university hospital and regional hospitals in Germany (University of Leipzig Medical Center and the Muldental Clinics in Grimma and Wurzen). It currently runs 5 different algorithms in parallel: hypokalemia, hypercalcemia, hyponatremia, hyperlactatemia, and acute kidney injury.

Conclusions: AMPEL enables continuous surveillance of patients. The system is constantly being evaluated and extended and has the capacity for many more algorithms. We hope to encourage colleagues from other institutions to design and implement similar CDSS using the theory, specifications, and experiences described in this work.

(JMIR Med Inform 2021;9(6):e20407) doi:10.2196/20407

KEYWORDS

clinical decision support system (CDSS); laboratory medicine; digital health; reactive software agent; computational architecture

Introduction

Background

Clinical decision support systems (CDSS) are computational or technological systems designed to attend specific demands in health care [1-3]. CDSS aim to assist physicians and nurses in making better informed clinical decisions and ultimately improve patient safety [4]. Importantly, the implementation of a CDSS in a health care environment requires the integration of the system into the pre-existing dataflow, computational infrastructure, and clinical procedures [3,5].

Commonly, hospitals use enterprise software in their infrastructure due to the amount, complexity, and sensitivity of data. SAP (SAP Software Solutions, Walldorf, Germany) [6] is an enterprise software widely used in health care organizations worldwide for data administration and processing. SAP and similar platforms require robust and complex infrastructures, which includes database management systems (DBMS). The SAP platform includes a Business/Data Warehouse (BW) module with either in-disk SQL DBMS or the more modern, highly performant in-memory HANA (SAP Software Solutions, Walldorf, Germany) database. The BW is a central platform that synchronizes and standardizes data structures from different systems. It provides input data that can be used by CDSS and algorithms to perform complex analyses using well-established techniques from statistics and computer science [7-9].

Remarkably, laboratory diagnostics are highly standardized, quality assured, and one of the most important sources for clinical decision making [10,11]. Many results are of numeric nature (eg, K^+ [potassium] = 3.9 mmol/L), allowing for rule sets of control values to be defined by medical specialists. This, in turn, creates an ideal context for a CDSS based on programming logic. With this approach, thousands of results can be evaluated, and notifications can be sent whenever critical conditions are detected, providing valuable assistance to health care professionals. Systems that evaluate laboratory results for specific purposes of their facilities have been reviewed in [12]. To our knowledge, however, until now, there are no systems available that are capable of working with different laboratory biomarkers in parallel, are scalable, and can identify complex laboratory assemblies, apart from AMPEL (Analysis and Reporting System for the Improvement of Patient Safety through Real-Time Integration of Laboratory Findings). Alert fatigue is a recorded problem of CDSS [13,14] and should be carefully addressed. However, laboratory measurements are patient-specific and, when used in a CDSS context, show a good balance between overalerting and underalerting [15].

Objectives

In this contribution, we first present guidelines and a theory for a reactive software agent (RSA) for implementing a CDSS that performs laboratory diagnostics (labCDSS). To offer practical support, we present a general system with its minimal components, so that other colleagues can implement a labCDSS in their own medical facilities. We also present our specific implementation and its evaluation at the University of Leipzig Medical Center (ULMC), Germany: AMPEL, which means traffic light in German.

AMPEL is essentially a notification system [16]. When the value of a laboratory parameter falls outside of predefined reference limits, a quick medical reaction is of paramount importance (eg., $K^+ \leq 2.0$ mmol/L, defining severe hypokalemia). The longer this medical response is delayed, the higher the risk to patient safety due to an associated risk between severe hypokalemia and sudden cardiac death [17]. The AMPEL system notifies the medical personnel when the patient's parameters increase the chance of life-threatening conditions. Our motivation for this research project was to design, implement, and evaluate a CDSS that improves patient safety and aids health care professionals in detecting life-threatening conditions.

Methods

Overview

In this section, we describe the technical requirements and components for a labCDSS. With these guidelines, health information technology (IT) specialists can plan the implementation of a labCDSS at a particular medical facility. As an example, we describe an overview of the computational infrastructure of the ULMC, where we implemented AMPEL.

Technical Requirements

Automated systems in medical facilities must satisfy certain technical requirements. Prior planning is crucial for successful implementation of a labCDSS. Afterwards, the system's components should be constantly monitored and potentially optimized.

The following 2 sections contain (1) a collection of general CDSS requirements, which are tailored to (2) a labCDSS. For broader and legally detailed requirements, see the work by Harer and Baumgartner [18], especially chapters 2 and 5.

System Functionality

The main functionality of a CDSS is to deploy different types of notifications. These are sent to clinicians or to the laboratory staff who will communicate directly with them whenever a patient's parameter is not within a safe range or a medical treatment is delayed.

System Performance

The performance is measured as the time a notification takes to be delivered. Delays can result from a high number of notifications or notification configurations that should be optimized. These issues influence the choices of the appropriate system components, such as database configurations, underlying hardware, and notification means (ie, socket-based, REST-interface, among others).

Component Failure

Component failure can also cause delays, which could ultimately affect patients. To avoid it, a working state of the CDSS should be maintained, even if some of its components fail. This is particularly challenging for a system that is partially integrated into a public infrastructure (ie, internet and power), which is

subjected to planned (eg, maintenance) and unplanned (eg, wear) outages. Users should be properly informed about such eventualities.

Some components can be uncoupled from public providers (eg, using a facility-wide intranet or a separate power grid). In general, however, it is more practical to distribute them over multiple redundant components.

Human Error

Two types of human error are noteworthy when planning a CDSS. The first is incorrect algorithm design. To avoid this, algorithms must be precisely tailored to specific laboratory parameters or medical diagnosis as well as based on (1) medical literature, (2) retrospective statistical analysis of available data, (3) expertise from medical laboratory specialists, and (4) expertise from clinical specialists.

The second type of error can occur during the display of notifications. Notifications in the graphical user interface (GUI) must be unambiguous and descriptive. This should be carefully considered during graphical design, along with the intended level of intrusiveness in the busy environment of an inpatient unit [19].

In addition to technical requirements, medical products are also subject to clinical evaluation before they are deployed at a particular facility. The new European Union regulation 2017/745 about general safety and performance requirements in medical devices [20] requires clinical investigations to show that the medical device (1) reaches the intended purpose planned in its design and works as expected and/or (2) provides the expected clinical benefit and/or (3) shows an acceptable risk/benefit ratio, weighing the side effects against the benefits to be achieved by the device [21].

The presented labCDSS, AMPEL, complies with these safety requirements by featuring a quantifiable benefit, transparent actions, the impossibility of adverse effects, and ease of reversibility (more details in the Discussion).

The Functional Components of a Reactive Software Agent

Agent-based architectures have been implemented for different purposes and in the medical domain [22,23]. They share 4 structural components [24]: (1) perceptive elements that receive input, (2) digitally represented domain knowledge, (3) an inference module that generates action by applying knowledge to the input, and (4) an output component that presents these actions to the patient or the attending physician. See [25] for a systematic review.

In a software agent, information continually flows into the system as well as out of it. The stream of input data is processed by the inference module using expert knowledge, which enables input interpretation and output generation.

The first medical agent architectures were designed for home care (for examples, see [26,27]; for an overview, see [28]). Currently, various agent-based software solutions have been published with the aim of improving patient care. For example, see the work by Schaaf and colleagues [29] with a system for diagnostic support in patients with rare diseases or Nguyen and colleagues [30] with a system that supports optimal antibiotic therapy.

A Reactive Software Agent for a CDSS That Performs Laboratory Diagnostics

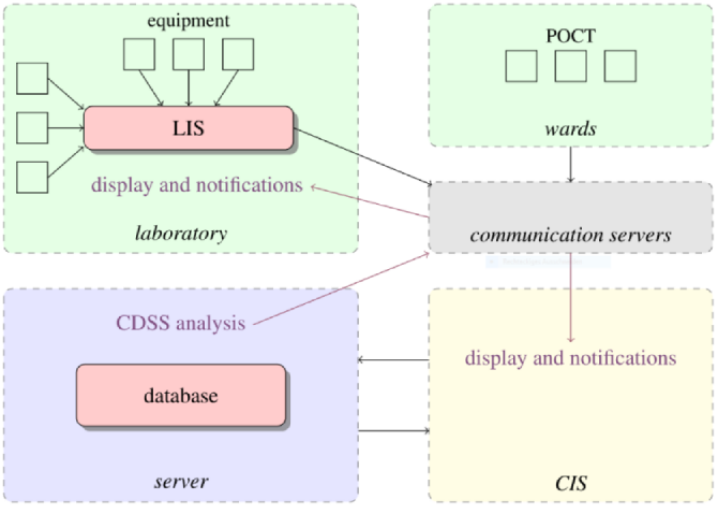
An RSA is comprised of 3 components, which are described in the following sections in the context of a labCDSS.

Input Component

The input component is depicted by green rectangles in Figure 1. The labCDSS receives input data from the laboratory information system (LIS) at a central server, which stores, manages, and processes data from various sources [31]. The input data are, however, not restricted to the LIS. The central server can also receive laboratory results over the internet, for instance.

Network packages should be encoded in standard eHealth formats, such as Health Level Seven International (HL7) or American Society for Testing and Materials (ASTM).

Figure 1. Conceptual computational infrastructure and data flow of a clinical decision support system (CDSS) that performs laboratory diagnostics following a reactive software agent framework. Input components are colored green, knowledge representation and inference components are blue, and output components are yellow. Laboratory parameters are measured by Point-of-Care Testing (POCT) devices and/or at a dedicated laboratory, stored in a Laboratory/Clinic Information System (LIS/CIS), and sent over to one or more information management systems and a server. Nodes indicate components that either generate or process data, while the edges indicate directional data transfer between the components over the internet.



Knowledge Representation and Inference

The knowledge representation and inference is shown as the blue rectangle in Figure 1. This is the key component of the labCDSS. Although parameter evaluation could be performed in other components, an appropriate one is the central server, which often holds a data warehouse (DW) and performs extract, transform, and load (ETL) processes. The DW is the central collection point for data that are generated by various sources.

ETL processes are performed by the DW at regular intervals to keep the data up to date. ETL processes receive data from various sources, clean and unify them, and are then processed or loaded within the DW. How often data are received and formatted depends on the institutional specificities [32-34].

Output Component

The output components are shown in the yellow rectangle in Figure 1. The output of the labCDSS is supplied to the CIS, which is usually connected to the various systems of modern clinical facilities, such as pathology and radiology. The CIS stores information in an electronic medical record, which clinicians have access to. The format of the electronic medical record can be modified to include new data elements and entry methods. Display and special features can also be included.

Local Computational Infrastructure

The ULMC has 6000 employees and 1451 beds with 57,000 inpatients, 374,000 outpatients, and 32,000 emergency patients annually [35]. Around 15,000 laboratory parameters are measured every day at the ULMC. Most of them are measured at the Institute for Laboratory Medicine, Clinical Chemistry and Molecular Diagnostics (ILM). The ILM is a medical laboratory that provides measurements of patients' samples in daily routine diagnostics. It offers these services to the inpatient and outpatient units, the departments of the Faculty of Medicine, the Medical Care Center of the ULMC [36], and external institutes.

Some of the laboratory diagnostics are performed directly at the inpatient units, via point-of-care testing (POCT) devices. These include ABL90 Flex and ABL800 Flex blood-gas, metabolite, and electrolyte analyzers (Radiometer, Krefeld, Germany) [37] as well as blood glucose meters (Roche, Basel, Switzerland) [38]. After measurements, either at the ILM or directly at the inpatient units, diagnostic data is sent to the CIS and LIS via communication servers: Aqure (Radiometer, Krefeld, Germany), Kodix (in-house development), and IT-1000 (Roche, Basel, Switzerland).

SAP [6] is the CIS of the ULMC and performs several important tasks. It is not as efficient in handling laboratory results and has only limited reporting options. For these reasons, the ULMC uses LabCentre (i-Solutions health, Mannheim, Germany) [39] for this task. LabCentre is the LIS of the ULMC and is of utmost importance to the AMPEL project. It is accessed by the AMPEL team for control as well as for coordination of phone calls via an implemented phone call documentation system. This system is implemented in the LIS, so that for each algorithm or even for the whole report, telephone calls can be documented. The documentation includes the call time, phone number, and name of the called party. In parallel to LabCentre, SAP [6] contains several modules, each dedicated to a specific purpose (eg, patient care, data storage, controlling, and logistics). The SAP R/3 module has a visual interface that displays patient data and is used for patient care purposes by medical personnel. The core of AMPEL analyses occurs in the SAP BW module. The output of AMPEL is sent from the BW to the SAP R/3 module and is integrated into the inpatient unit overview in the form of a dedicated AMPEL column.

The computational infrastructure of the ULMC can be represented as a graph, in which the nodes indicate components that either generate or process data and the edges indicate directional data transfer between the components via a shared resource or socket. The BW of the ULMC is model-driven and based on the SAP NetWeaver ABAP platform [40], which runs

the release version SAP NetWeaver Server Version 7 with EHP 4 for both SAP BASE and SAP BW modules.

The BW receives data from heterogeneous sources (SAP and non-SAP alike) and cleans and stores them for downstream analysis. In addition, there are 2 MSSQL databases for data storage, one connected to the BW and the other connected to SAP R/3. Communication servers transfer data between the machines that measure the laboratory parameters and the BW, CIS, and LIS. The standard format HL7 [41] is used for all data transfer, with the exception of the AMPEL system, which uses a custom format, and the transfer between Aqure and LabCentre, which uses the ASTM [42] format.

Results

Overview

In this section, we detail the AMPEL labCDSS as well as its implementation in the ULMC. AMPEL was planned according to the technical requirements defined in the "Technical Requirements" subsection in the Methods and implemented within the RSA framework described in "The Functional Components of a Reactive Software Agent" subsection in the Methods.

It can also be implemented in different computational infrastructures that provide the functions defined in the "A Reactive Software Agent for a CDSS that Performs Laboratory Diagnostic" subsection in the Methods.

AMPEL at the ULMC

Input Component

The input to AMPEL comes from the SAP R/3 module, which, in turn, receives data from the POCT machines at the inpatient and outpatient units as well as the machines from the ILM. AMPEL is implemented in the SAP BW. Consequentially, the communication was adapted to SAP's native data transport interfaces.

The exact content of these communication packages varies, but some core information is required: (1) a timestamp that indicates the measurement time, (2) the patient ID to associate the data with the particular medical record, (3) the case ID to relate the values to the current treatment, (4) an order ID to identify the batch of the measurements throughout the current treatment, (5) the laboratory parameter, and (6) its result.

Knowledge Representation and Inference

Incoming data packages are stored and processed in the SAP BW. The SAP BW is the central server of the ULMC (Figure 1, bottom left) and uses ETL processes to make the evaluations. The algorithms that compose AMPEL encode medical domain knowledge to transform the input into medically relevant output. The output of each algorithm is a classification of the patient's parameter results regarding a particular parameter or diagnosis. Notifications are triggered whenever a value falls outside the defined range.

After processing, 1 of 3 possible outputs is generated: (1) all values are within safe ranges; (2) at least one value needs to be monitored, but no immediate action is required; or (3) at least one value is critical, and immediate action is required.

Output Component

Once the output has been generated, the host system converts it into an outbound message. AMPEL makes use of the secure internet communication framework (SICF) integrated into SAP and sends the resulting data as a HTTP post package to its target destination. Other communication channels (eg, REST interfaces) would also be possible.

Implementation Details

When the laboratory results of a patient become available at the SAP BW, they are subjected to AMPEL's algorithms. The mean time between a critical condition being detected by the algorithms and a notification being sent to the AMPEL team is 36 minutes. This time can be shorter or longer depending on how fast the data are processed throughout the computational system (for more details, refer to the Limitations section). Each algorithm is specific to a laboratory parameter or a diagnosis (eg, hypokalemia and potassium). Five algorithms have already been implemented (Table 1), and two more complex algorithms are in the prerelease phase (Table 1). In addition, up to 23 other algorithms are being developed. Importantly, each algorithm is tailored to the specificities of the parameter, diagnosis, and other relevant medical aspects. Some of them (eg, hypokalemia and hypercalcemia) have straightforward control rule sets, while others, such as creatinine, for the detection of renal failure, or procalcitonin (PCT), to diagnose and monitor infections, require more complex rule sets. Each algorithm is carefully developed by a team of physicians, scientists, and IT personnel (computer scientists, computer engineers, and bioinformaticians) under close consideration of literature and extensive practical experience. Theoretically, algorithms can be defined in any formal language that enables the description of rational functions over input values and time.

Table 1. Laboratory parameters, rule sets, time to control (TTC), and diagnosis from AMPEL. Each line comprises a specific algorithm. If any parameter value falls outside the defined range (rule set), the diagnosis is documented, and notifications are sent to warn clinicians in the inpatient units.

Laboratory parameter	Rule set	TTC	Diagnosis
K^+ (potassium)	<2.5 mmol/L	≥6 hours	Hypokalemia
Ca^{++} (calcium)	>3.5 mmol/L or >2.0 mmol/L ionized	≥12 hours	Hypercalcemia
Na^+ (sodium)	<120 mmol/L	≥12 hours	Hyponatremia
Lactate	>4 mmol/L	≥ 6 hours	Hyperlactatemia
Creatinine	Complex rule set	Immediately	Acute kidney injury
Procalcitonin (PCT)[a]	Complex rule set	Prerelease	Sepsis
Troponin T[a]	Complex rule set	Prerelease	Myocardial infarction

[a]Requires more complex rule sets and are currently under development.

Most of the algorithms were based on the following: First, the result of the laboratory parameter should indicate a critical medical finding with a high specificity. Second, the system should report if there is evidence of delayed medical interventions. Both aspects indicate the seriousness of the medical situation. For instance, a blood potassium deficiency of K^+ ≤1.9 mmol/L is acutely life threatening and requires immediate response. Therefore, the laboratory instantly informs the inpatient and outpatient units of these cases, and adequate treatment must be initiated. If the potassium of these patients has not been checked within the last 6 hours, this critical finding may have been overlooked, and a life-threatening condition is likely to occur, due to the risk of heart rhythm disturbances.

AMPEL detects such cases computationally, and clinicians in the inpatient and outpatient units are notified to the situation via a phone call. These notifications originate at the BW, in which text messages are generated according to the algorithms. The notifications appear as items on a telephone list of the LIS, which are manually processed, either by the AMPEL team (business hours) or by a medical technologist of the ULMC (24 hours, 7 days a week).

Feedback on the system is obtained at the end of each notification call. Two questions are asked, and the answers are manually entered into the LIS: “Has the therapy started (based on laboratory findings)?” and “Do you think the notification helped in treating the patient?” This direct communication is part of the evaluation of the AMPEL system and incorporates user feedback.

The AMPEL displays at the SAP R/3 are assigned 1 of 3 colors: green, yellow, or red. The colors indicate the status of the evaluated parameter with regard to the health risk to the patient. As a summary of all analyzed parameters, 1 unique color will be assigned to each patient in a dedicated column at the SAP R/3 interface (see Figure 2). This assignment follows a hierarchy: If there is at least 1 red parameter, the overview color is set to red; if there are no red parameters but at least 1 yellow, the overview color is set to yellow. Green is set otherwise.

If the connection between the server and the terminals at the inpatient units fails, the AMPEL column of SAP R/3 is simply not filled with any color and remains empty. The terminal interface queries the AMPEL server every 600 seconds. As soon as a connection has been reestablished, the column appears again on the terminal.

Moreover, a user can click on a specific AMPEL symbol and get a full AMPEL report for a patient (Figure 3). This represents an overview of all the patient’s notifications and their individual assessment. Importantly, when a clinician checks all yellow and red AMPEL notifications and attends to the patient, they manually click on the checkmark in the SAP R/3 AMPEL column. The checkmark is afterwards displayed in SAP (Figure 2), informing other clinicians that the patient has already been attended to.

Figure 2. Screenshot of the SAP R/3 visual interface with highlight on the dedicated AMPEL column, with all three possible colors (summary report) displayed: green, yellow, or red.

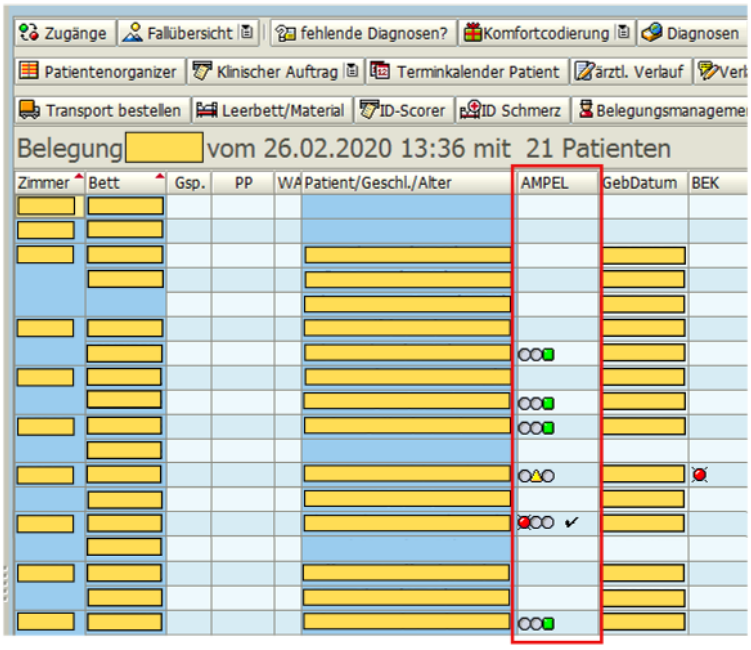


Figure 3. Screenshot of the interactive AMPEL report for a patient at the SAP R/3 visual interface. All algorithms that have been analyzed by the AMPEL system are displayed, each in a highlighted line. The user can access the specific reports by clicking on the highlighted line. In this example, the creatinine (AKIN 2) report for acute kidney injury has been chosen. The full report displays additional information to clinicians, such as internal standards or specifics from medical guidelines. The blue box at the top states that the analysis has been carried out automatically by AMPEL, an ongoing research project, and should therefore be used with caution.

Discussion

Principal Findings

We designed and implemented the AMPEL system in the ULMC as well as the Muldental Clinics in Grimma and Wurzen in Germany. AMPEL is a labCDSS with 5 active algorithms and 2 under development. Each algorithm is specific for a laboratory parameter or a diagnosis and is carefully developed by an interdisciplinary team. If a critical condition is detected, notifications are sent to warn the patient's attending clinicians. AMPEL is constantly being evaluated regarding functionality and usefulness. In addition, we are currently collecting data for prospective analysis and scientific evaluation of each algorithm.

Safety Requirements

This section is divided into 4 main categories. In extent and detail, these features of the AMPEL system address and go beyond annex I of EU regulation 2017/745 mentioned in the Technical Requirements subsection in the Methods.

Quantifiable Benefit

During the development phase, ULMC patients are assigned to a control or intervention group so as to quantify AMPEL's effects. Odd patient IDs make up the control group and even IDs the intervention group. The assignment of patient IDs is

incremental and nonsystematic and not expected to create a bias for any of the test groups. Both groups undergo the AMPEL analysis but only the results for the intervention group are presented in the AMPEL column of the SAP R/3 terminal (for an example, see Figure 3).

Preliminary results show that the time to control (TTC) is considerably lower for patients of the test group ([16] and unpublished data). Based on AMPEL's transparency of actions, its deterministic algorithms, and the randomized controlled experiment setting, there is solid reason to assume this is due to AMPEL's notifications. To further investigate these preliminary indications, trials are currently ongoing, and analysis are carried out for specific patient cohorts (eg, male versus female, patients with liver and kidney disease, and others).

Transparent Actions

The notifications of the system are determined by manually designed rule sets. As long as the involved components perform according to their specifications, the system successfully notifies the clinicians of all cases that are singled out by the rule set. It is important that any system action can be traced back to one unambiguous cause. This is necessary for the direction of subsequent medical attention as well as investigation in case of possible erroneous notifications. This issue is especially important to keep in mind in case the system is extended at some point — for example with a machine learning module — to guarantee the impossibility of adverse effects.

Impossibility of Adverse Effects

Component failure is not addressed by AMPEL as a research project because of extremely rare outages. This will become relevant, however, when AMPEL matures into a medical product. If the system is to be deployed at places with less reliable infrastructure, a coordinated distribution of its components is reasonable and should be incorporated.

The AMPEL system was designed to avoid unnecessary notifications. To achieve this, we simulate the system with retrospective patient data and evaluate the amount of less helpful notifications in coordination with clinical specialists. Furthermore, the system is continually adapted and evaluated prospectively. Up to the writing of this article, the feedback of clinicians has been very positive, although individual, less helpful notifications cannot be fully avoided. For patients in palliative medical conditions, the system can be individually silenced by the system administrator.

Ease of Reversibility

In the unlikely case that the system consistently generates false notifications, this could result in an unfavorable redistribution of a clinician's attention. If this occurs, the system can be easily removed from the computational infrastructure.

AMPEL does not replace any other notification system. Therefore, problems caused by the removal of the AMPEL column (see Figure 2) might arise only if staff has started to depend on AMPEL as their sole means of a reminder.

Limitations

Considering the complexity of the computational infrastructure and the amount of input data at the ULMC, unforeseen issues may arise in the AMPEL system, such as delays or connectivity outages. In order to localize and solve these problems, thorough documentation is required, along with continuous monitoring and communication between users and system management.

After several months of evaluation, delays were identified within the current implementation of AMPEL. The most frequent one was found within the on-disk SAP BW database itself, contrary to a faster in-memory, and is only accessed by the BW in fixed time intervals (every 60 minutes). Due to this delay, follow-up examinations might already have taken place but are still displayed as pending in the SAP R/3 AMPEL column. We are in the process of moving our system to an in-memory infrastructure to minimize this delay.

Another reason for delay has been identified in the Kodix communication server. It sends parameter values to SAP R/3 only every 15 minutes. In case of a connection break, parameters must be sent manually by the Kodix IT team. Kodix belongs to the critical IT infrastructure. Outside working hours, a computer scientist is available on call at the hospital.

In order to avoid delays of the AMPEL evaluation, we developed systematic documentation in the form of a checkbox protocol, based on the dataflow of a parameter measurement. As an example, if we notice a large time lag between the measurement of the POCT machine and the timestamp of the Aqure connecting server, we can pinpoint the delay to the connecting servers Aqure or Kodix. Alternatively, if the lag occurred between the TTC and the timestamp of the email (both being assigned at the SAP BW), we can pinpoint the delay to the internal system of AMPEL in the SAP BW.

Problems regarding the components of the computational infrastructure are inherited by the AMPEL system. If one of the communication servers is under maintenance, messages to the LIS can be sent twice. In our settings, unscheduled suspension of the Kodix server, for example, leads to a resubmission of messages from the previous day, causing duplications.

Another issue was detected with overlooked or misjudged notifications at the LIS interface (on which phone calls are based). Part of this can be avoided simply by reducing delays in message delivery, such that there is no ambiguity as to the notification's current relevance and whether it should still be considered.

The data of the ULMC are mirrored in the SAP BW server. This ensures availability of data and provides a redundant means of storage. Conservative default settings set the maximum query frequency for the secondary database to 1 hour. In some cases, this can lead to notifications arriving after the necessary treatment has been initiated. Given the proper functionality of all system components, the maximum delay between measurement and notification is 85 minutes (up to 15 minutes for the Kodix server to send data, up to 60 minutes for the SAP BW to access the on-disk database, and up to 10 minutes for the inpatient terminal display to update.)

Conclusions

A labCDSS can be used to notify clinicians of important follow-up procedures. It interprets and processes laboratory data at a central node and delivers the results to a team who notifies the attending clinicians.

An overwhelming amount of digitally recorded medical data is available in the CIS of a medical facility, which can yield valuable information, but is impossible to analyze manually. The rule sets of the AMPEL system are developed and verified by laboratory medicine specialists under critical consideration of current medical literature. These rules are implemented in the algorithms of AMPEL, which automatically process the large amount of available laboratory parameters and discover associations between the parameter results and diseases.

AMPEL and similar labCDSS could be a valuable asset in clinics in remote areas as well as smaller facilities that do not have in-house laboratory medicine experts, dedicated to screen laboratory results for improved patient safety. We presented in this contribution the technical requirements and functional components of a CDSS for laboratory diagnostics and exemplified them by detailing our implementation of the AMPEL system at the ULMC. We hope to encourage colleagues to also design and implement a labCDSS within their institutions.

Acknowledgments

The authors would like to thank the IT personnel from ULMC, especially Franziska Jeromin and Thomas Thalheim, and the anonymous reviewers who contributed greatly for improving this paper.

This project is co-financed through public funds according to the budget decided by the Saxon State Parliament under the RL eHealthSax 2017/18 grant (eHealthSax-Richt-linie Nr.: 100331796 - Diese Maßnahme wird mitfinanziert mit Steuermitteln auf Grundlage des von den Abgeordneten des Sächsischen Landtags beschlossenen Haushaltes).

Authors' Contributions

TK, JT, HR, MW, and BI designed the study. MWC and MW formalized the AMPEL system in the context of the ULMC computational infrastructure and wrote the manuscript. AK, MV, and CC implemented and administrate the AMPEL computational system. MF, FE, JR, MS, SP, and CG developed and monitor the theoretical algorithms and communicate with users. TK, MS, MF, FE, JR, and CG contributed to writing the manuscript. All authors approved the final manuscript.

Conflicts of Interest

AMPEL is currently a public-funded research project and runs at ULMC and Muldental Clinics in Grimma and Wurzen. After completion of the project, it will be transferred to the controlling software Vismedica of Xantas AG to be commercialized. AK and JT from Xantas AG as well as all other co-authors declare that the future commercialization of AMPEL had no influence on the research or writing of the manuscript.

References

1. Silveira DV, Marcolino MS, Machado EL, Ferreira CG, Alkmim MBM, Resende ES, et al. Development and Evaluation of a Mobile Decision Support System for Hypertension Management in the Primary Care Setting in Brazil: Mixed-Methods Field Study on Usability, Feasibility, and Utility. JMIR Mhealth Uhealth 2019 Mar 25;7(3):e9869 [FREE Full text] [doi: 10.2196/mhealth.9869] [Medline: 30907740]
2. Wang J, Bao B, Shen P, Kong G, Yang Y, Sun X, et al. Using electronic health record data to establish a chronic kidney disease surveillance system in China: protocol for the China Kidney Disease Network (CK-NET)-Yinzhou Study. BMJ Open 2019 Aug 28;9(8):e030102 [FREE Full text] [doi: 10.1136/bmjopen-2019-030102] [Medline: 31467053]
3. Adnan M, Peterkin D, McLaughlin A, Hill N. HL7 Middleware Framework for Laboratory Notifications for Notifiable Diseases. Stud Health Technol Inform 2015;214:1-7. [Medline: 26210410]
4. Courbis A, Murray RB, Arnavielhe S, Caimmi D, Bedbrook A, Van Eerd M, et al. Electronic Clinical Decision Support System for allergic rhinitis management: MASK e-CDSS. Clin Exp Allergy 2018 Dec 20;48(12):1640-1653. [doi: 10.1111/cea.13230] [Medline: 29999223]
5. Schuh C, de Bruin JS, Seeling W. Clinical decision support systems at the Vienna General Hospital using Arden Syntax: Design, implementation, and integration. Artif Intell Med 2018 Nov;92:24-33. [doi: 10.1016/j.artmed.2015.11.002] [Medline: 26706047]
6. SAP Software Solutions. URL: https://www.sap.com/index.html [accessed 2020-03-24]
7. Schrodt J, Dudchenko A, Knaup-Gregori P, Ganzinger M. Graph-Representation of Patient Data: a Systematic Literature Review. J Med Syst 2020 Mar 12;44(4):86 [FREE Full text] [doi: 10.1007/s10916-020-1538-4] [Medline: 32166501]
8. Roth J, Goebel N, Sakoparnig T, Neubauer S, Kuenzel-Pawlik E, Gerber M, PATREC Study Group. Secondary use of routine data in hospitals: description of a scalable analytical platform based on a business intelligence system. JAMIA Open 2018 Oct;1(2):172-177 [FREE Full text] [doi: 10.1093/jamiaopen/ooy039] [Medline: 31984330]

9. Kreuzthaler M, Martínez-Costa C, Kaiser P, Schulz S. Semantic Technologies for Re-Use of Clinical Routine Data. Stud Health Technol Inform 2017;236:24-31. [Medline: 28508775]
10. Regan M, Forsman R. The impact of the laboratory on disease management. Dis Manag 2006 Apr;9(2):122-130. [doi: 10.1089/dis.2006.9.122] [Medline: 16620198]
11. Forsman R. The value of the laboratory professional in the continuum of care. Clin Leadersh Manag Rev 2002;16(6):370-373. [Medline: 12506827]
12. Slovis B, Nahass TA, Salmasian H, Kuperman G, Vawdrey DK. Asynchronous automated electronic laboratory result notifications: a systematic review. J Am Med Inform Assoc 2017 Nov 01;24(6):1173-1183 [FREE Full text] [doi: 10.1093/jamia/ocx047] [Medline: 28520977]
13. Desmedt S, Spinewine A, Jadoul M, Henrard S, Wouters D, Dalleur O. Impact of a clinical decision support system for drug dosage in patients with renal failure. Int J Clin Pharm 2018 Oct 21;40(5):1225-1233. [doi: 10.1007/s11096-018-0612-1] [Medline: 29785684]
14. Pfistermeister B, Sedlmayr B, Patapovas A, Suttner G, Tektas O, Tarkhov A, et al. Development of a Standardized Rating Tool for Drug Alerts to Reduce Information Overload. Methods Inf Med 2018 Jan 08;55(06):507-515. [doi: 10.3414/me16-01-0003]
15. Muylle K, Gentens K, Dupont AG, Cornu P. Evaluation of context-specific alerts for potassium-increasing drug-drug interactions: A pre-post study. Int J Med Inform 2020 Jan;133:104013. [doi: 10.1016/j.ijmedinf.2019.104013] [Medline: 31698230]
16. Eckelt F, Remmler J, Kister T, Wernsdorfer M, Richter H, Federbusch M, et al. [Improved patient safety through a clinical decision support system in laboratory medicine]. Internist (Berl) 2020 May 27;61(5):452-459. [doi: 10.1007/s00108-020-00775-3] [Medline: 32221627]
17. Kjeldsen K. Hypokalemia and sudden cardiac death. Exp Clin Cardiol 2010;15(4):e96-e99 [FREE Full text] [Medline: 21264075]
18. Harer J, Baumgartner C. Anforderungen an Medizinprodukte: Praxisleitfaden für Hersteller und Zulieferer. Munich, Germany: Hanser Fachbuchverlag; 2018.
19. Pelayo S, Marcilly R, Bernonville S, Leroy N, Beuscart-Zephir MC. Human factors based recommendations for the design of medication related clinical decision support systems (CDSS). Stud Health Technol Inform 2011;169:412-416. [Medline: 21893783]
20. Regulation (EU) 2017/745 of the European Parliament and of the Council. Official Journal of the European Union. 2017 Apr 05. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32017R0745&from=IT#d1e32-94-1 [accessed 2020-04-20]
21. Bianco S, Nunziata A, Pozzoli G. Clinical investigations on medical devices, after the new European regulation (2017/745). Clinical Trials and Practice 2017;1(1):10-14 [FREE Full text] [doi: 10.17140/CTPOJ-1-102]
22. Lanzola G, Gatti L, Falasconi S, Stefanelli M. A framework for building cooperative software agents in medical applications. Artificial Intelligence in Medicine 1999 Jul;16(3):223-249. [doi: 10.1016/s0933-3657(99)00008-1]
23. Nguyen MT, Fuhrer P, Pasquier-Rocha J. Enhancing e-health information systems with agent technology. Int J Telemed Appl 2009;2009:279091-279013 [FREE Full text] [doi: 10.1155/2009/279091] [Medline: 19096509]
24. Kruger GH, Chen C, Blum JM, Shih AJ, Tremper KK. Reactive software agent anesthesia decision support system. Systemics, Cybernetics and Informatics 2011;9(6):30-37 [FREE Full text]
25. Isern D, Moreno A. A Systematic Literature Review of Agents Applied in Healthcare. J Med Syst 2016 Feb 21;40(2):43. [doi: 10.1007/s10916-015-0376-2] [Medline: 26590981]
26. Isern D, Moreno A, Sánchez D, Hajnal A, Pedone G, Varga LZ. Agent-based execution of personalised home care treatments. Appl Intell 2009 Jun 24;34(2):155-180. [doi: 10.1007/s10489-009-0187-6]
27. Kaluža B, Mirchevska V, Dovgan E, Luštrek M, Gams M. An Agent-Based Approach to Care in Independent Living. In: de Ruyter B, editor. Ambient Intelligence. AmI 2010. Lecture Notes in Computer Science, vol 6439. Berlin, Heidelberg: Springer; 2010:177-186.
28. Hudson DL, Cohen ME. Intelligent agents in home healthcare. Ann. Telecommun 2010 May 5;65(9-10):593-600. [doi: 10.1007/s12243-010-0170-6]
29. Schaaf J, Boeker M, Ganslandt T, Haverkamp C, Hermann T, Kadioglu D, et al. Finding the Needle in the Hay Stack: An Open Architecture to Support Diagnosis of Undiagnosed Patients. Stud Health Technol Inform 2019 Aug 21;264:1580-1581. [doi: 10.3233/SHTI190544] [Medline: 31438241]
30. Nguyen A, Hassanzadeh H, Zhang Y, O'Dwyer J, Conlan D, Lawley M, et al. A Decision Support System for Pathology Test Result Reviews in an Emergency Department to Support Patient Safety and Increase Efficiency. Stud Health Technol Inform 2019 Aug 21;264:729-733. [doi: 10.3233/SHTI190319] [Medline: 31438020]
31. Yusof MM, Arifin A. Towards an evaluation framework for Laboratory Information Systems. J Infect Public Health 2016 Nov;9(6):766-773 [FREE Full text] [doi: 10.1016/j.jiph.2016.08.014] [Medline: 27665060]
32. Mukherjee R, Kar P. A Comparative Review of Data Warehousing ETL Tools with New Trends and Industry Insight. 2017 Presented at: IEEE 7th International Advance Computing Conference (IACC); Jan 5-7, 2017; Hyderabad, India. [doi: 10.1109/iacc.2017.0192]

33. Sachin S, Goyal SK, Avinash S, Kamal K. Nuts and Bolts of ETL in Data Warehouse. In: Rathore V, Worring M, Mishra D, Joshi A, Maheshwari S, editors. Emerging Trends in Expert Applications and Security. Advances in Intelligent Systems and Computing, vol 841. Singapore: Springer Publishing Company; 2019:1-9.
34. Vyas S, Vaishnav P. A comparative study of various ETL process and their testing techniques in data warehouse. Journal of Statistics and Management Systems 2017 Nov 16;20(4):753-763. [doi: 10.1080/09720510.2017.1395194]
35. Universitätsmedizin Leipzig in numbers. Universitätsmedizin Leipzig. URL: https://www.uniklinikum-leipzig.de/Seiten/uml-in-zahlen.aspx [accessed 2020-04-04]
36. Institut für Laboratoriumsmedizin. Universitätsmedizin Leipzig. URL: http://ilm.uniklinikum-leipzig.de/ [accessed 2020-04-03]
37. ABL90 FLEX BGA device. Radiometer. URL: https://www.radiometer.de/de-de/produkte-und-l%C3%B6sungen/blutgasanalyseger%C3%A4te/abl90-flex-bga-gerat [accessed 2020-03-31]
38. Accu-Chek Inform II system. Roch. URL: https://diagnostics.roche.com/global/en/products/instruments/accu-chek-inform-ii.html [accessed 2020-03-31]
39. Laboratory. i-Solutions Health. URL: https://i-solutions.de/loesungen/labor.php [accessed 2020-03-24]
40. Data Flow in SAP Business Warehouse. SAP Help Portal. URL: https://help.sap.com/viewer/7511bced67d5418da87c4aa5fdec77d7/7.40.17/en-US/4a1e8b8a46c51977e10000000a42189c.html [accessed 2020-03-03]
41. Introduction to HL7 Standards. HL7 International. URL: https://www.hl7.org/implement/standards/ [accessed 2020-03-16]
42. ASTM International. URL: https://www.astm.org/ [accessed 2020-03-27]

Abbreviations

AMPEL: "Analyse- und Meldesystem zur Verbesserung der Patientensicherheit durch Echtzeitintegration von Laborbefunden," which translates to "Analysis and Reporting System for the Improvement of Patient Safety through Real-Time Integration of Laboratory Findings"
ASTM: American Society for Testing and Materials
BW: Business Warehouse
CDSS: clinical decision support system
CIS: clinical information system
DBMS: database management system
DW: data warehouse
ETL: extract, transform, and load
GUI: graphical user interface
HL7: Health Level Seven International
ILM: Institute for Laboratory Medicine, Clinical Chemistry and Molecular Diagnostics
IT: information technology
K: potassium
labCDSS: CDSS that performs laboratory diagnostics
LIS: laboratory information system
PCT: procalcitonin
POCT: point-of-care testing
RSA: reactive software agent
SICF: secure internet communication framework
TTC: time to control
ULMC: University of Leipzig Medical Center

Edited by G Eysenbach; submitted 15.06.20; peer-reviewed by A Bietenbeck, M Görges; comments to author 05.07.20; revised version received 28.08.20; accepted 20.11.20; published 03.06.21.

Please cite as:
Walter Costa MB, Wernsdorfer M, Kehrer A, Voigt M, Cundius C, Federbusch M, Eckelt F, Remmler J, Schmidt M, Pehnke S, Gärtner C, Wehner M, Isermann B, Richter H, Telle J, Kaiser T
The Clinical Decision Support System AMPEL for Laboratory Diagnostics: Implementation and Technical Evaluation
JMIR Med Inform 2021;9(6):e20407
URL: https://medinform.jmir.org/2021/6/e20407
doi:10.2196/20407
PMID:34081013

Original Paper

Using Electronic Health Records to Mitigate Workplace Burnout Among Clinicians During the COVID-19 Pandemic: Field Study in Iran

Pouyan Esmaeilzadeh[1*], PhD; Tala Mirzaei[1*], PhD

Department of Information Systems and Business Analytics, College of Business, Florida International University, Miami, FL, United States

[*]all authors contributed equally

Corresponding Author:
Pouyan Esmaeilzadeh, PhD
Department of Information Systems and Business Analytics
College of Business
Florida International University
Modesto A Maidique Campus 11200 SW 8th St
RB 261 B
Miami, FL, 33199
United States
Phone: 1 3053483302
Email: pesmaeil@fiu.edu

Abstract

Background: The COVID-19 pandemic spread worldwide in 2020. Notably, in the countries dealing with massive casualties, clinicians have worked in new conditions characterized by a heavy workload and a high risk of being infected. The issue of clinician burnout during the pandemic has attracted considerable attention in health care research. Electronic health records (EHRs) provide health care workers with several features to meet a health system's clinical needs.

Objective: We aim to examine how the use of EHR features affects the burnout of clinicians working in hospitals that have special wards for confirmed COVID-19 cases.

Methods: Using an online survey, we collected data from 368 physicians, physician assistants, and nurses working in six hospitals that have implemented EHRs in the city of Tehran in Iran. We used logistic regression to assess the association between burnout and awareness of EHR features, EHR system usability, concerns about COVID-19, technology solutions, hospital technology interventions, hospital preparedness, and professional efficacy adjusted for demographic and practice characteristics.

Results: The primary outcome of our study was self-reported burnout during the COVID-19 pandemic. Of the 368 respondents, 36% (n=134) reported having at least one symptom of burnout. Participants indicated that the leading cause of EHR-related stress is inadequate training for using technology (n=159, 43%), followed by having less face-to-face time with patients (n=140, 38%). Positive perceptions about the EHR's ease of use were associated with lower odds of burnout symptoms. More interventions, such as clear communication of regulations; transparency in policies, expectations, and goals regarding the use of technology in the clinical workflow; and hospital preparedness to cope with the challenges of the pandemic, were associated with lower odds of burnout.

Conclusions: The use of EHR applications, hospital pandemic preparation programs, and transparent technology-related policies and procedures throughout the epidemic can be substantial mitigators of technology-based stress and clinician burnout. Hospitals will then be better positioned to devise or modify technology-related policies and procedures to support physicians' and nurses' well-being during the COVID-19 pandemic. Training programs, transparency in communications of regulations, and developing a clear channel for informing clinicians of changes in policies may help reduce burnout symptoms among physicians and nurses during a pandemic. Providing easily accessible mentorship through teleconsultation and 24-hour available information technology support may also help to mitigate the odds of burnout.

(JMIR Med Inform 2021;9(6):e28497) doi:10.2196/28497

KEYWORDS

COVID-19; pandemic; clinician burnout; electronic health record; health information technologies; hospital intervention

Introduction

Burnout has attracted more attention in health care since it is considered a trigger for health care professionals' physical and mental problems [1]. Clinicians' burnout may negatively affect the quality of care, cost of health care delivery, productivity, and patient satisfaction [2]. Prior research indicates different organizational, relational, and work-related factors contributing to clinicians' burnout [3,4]. COVID-19 was first identified in Wuhan, China and has since become a global pandemic that has spread to more than 150 countries and affected over 9.6 million people worldwide [5]. Recently with the spread of the COVID-19 pandemic, burnout syndrome and physical exhaustion among clinicians have become even more pronounced. Declaring the COVID-19 outbreak as a pandemic by the World Health Organization has raised concerns about its possible detrimental effects on clinicians' workload and well-being [6].

During the outbreak, clinicians have been exposed to patients with severe or mild symptoms. Since respiratory droplets and close contact are the main transmitters of COVID-19, clinicians are at higher risk of being affected [7]. Besides the health-related stress, health care centers have had a sudden increase in phone calls, patient portal messages, appointment requests, ambulatory care visits, and walk-in patients [8]. Accommodating patients with lots of concerns, questions, and clinical demands has also increased health care professionals' workload. These burnout symptoms are exhibited by not only health care professionals who are specialized in infectious diseases but also physicians with different specialties, and nurses may experience more significant work pressure dealing with suspected and confirmed cases [7]. Previous studies have provided evidence to indicate that clinicians have experienced stress, anxiety, and depression during the COVID-19 pandemic [9,10].

Research to date has identified several contributing factors associated with burnout, including health information technology (HIT). Electronic health record (EHR) systems, for example, are often seen as cumbersome to use, failing to fulfill the promise of improved health care delivery and little more than a means of meeting regulatory and billing requirements [11,12]. However, during the COVID-19 outbreak, health care organizations are establishing different strategies and leveraging several health information technologies to improve pandemic management [13]. One of the technology-based tools that can support the standard management of patients during the current pandemic is the EHR [14,15].

The EHR has some features to enable the use of standardized processes such as scripted triaging, immediate health information exchange, real-time data analytics, electronic check-in, self-screening pages, and telemedicine [16]. Providing remote care through telemedicine and teleconsulting for patients became more prevalent during the outbreak. Using patient portals to provide instructions and medication prescriptions increased as well. There remains considerable debate in the informatics community about the actual role health information technologies play in shaping clinician burnout during the pandemic. Existing research suggests that technologies may be confounded with other important causes, including changes in regulatory mandates, increased clinical volumes, and flexibility to mitigate ad hoc challenges. Regardless of the role information technology (IT) plays in clinician burnout, innovative solutions to alleviate burnout during the pandemic are urgently needed. A recent article explains that "EHR innovations cannot help to mitigate clinician burnout without careful consideration of the socioecological context in which these innovations occur, including organizational culture, the healthcare marketplace, technology ecosystem, and national policy" [17]. There is a lack of research investigating the impact of this pandemic on burnout among clinicians in resource-limited countries and the possible impact of HIT on clinicians' stress and burnout [18].

We investigate how clinicians' awareness of EHR features and perception of EHR system usability as well as their perceptions about the hospital policies and preparedness can help reduce the burnout caused during the pandemic for clinicians in Iran, which was one of the resource-limited countries that had been reported to be among the first 15 countries to have COVID-19 patients and deaths [19]. As of April 8, 2021, Iran has 1,984,348 total confirmed cases and 63,699 deaths [19]. EHRs are still developing in Iran, and various hospitals are at a different level of adoption and use of EHRs. Hospitals in Iran use several EHR applications and informatics infrastructure for outbreak management [14,20]. However, little is known about how the use of EHR capabilities and clinicians' perceptions of this technology in the Iranian health system during the COVID-19 outbreak successfully reverse burnout among clinicians to manage this novel infection. This study's findings can contribute insights to develop better strategies for using EHRs, refining policies, and enhancing preparedness to reduce the increasing rate of burnout symptoms during the pandemics in resource-limited countries.

Methods

Setting

There are 118 hospitals in Tehran, the capital city of Iran, and among them, only 15 hospitals have wards specialized for quarantining and treating patients with COVID-19 (at the time of this research). In this study, we considered six of these hospitals with large administrations. These six hospitals were mainly selected because they were among the first hospitals in Tehran that have been quarantining and treating patients with COVID-19. They encompass outpatient primary and specialty medical and surgical care as well as emergency patient care. These hospitals were pioneer health systems in the country to care for COVID-19–positive patients. They created a special ward for COVID-19–infected patients in February 2020. They served as a quarantine site for confirmed cases and established a center to monitor and adapt to rapidly changing conditions and recommendations from local, state, and federal officials. The hospital characteristics are summarized in Table 1. These hospitals have implemented one of the main commercially available EHR systems (brands) in Iran with similar applications, functionalities, and features. Some of these vendors provide services to public hospitals, and some offer services to private hospitals. All EHR systems offer some of the basic features

such as viewing laboratory results, collecting patient demographics, listing patient problems, and highlighting out-of-range laboratory results. However, other advanced features such as sending and receiving reminders and messages and remote access to health information have been scarcely or not implemented yet. Table 2 shows the prevalence of EHR features implemented in the hospitals of interest in this study.

Table 1. Hospital details.

Type of hospital	Physicians, n	Nurses, n	Outpatients in a year, n	Total patients (annually), n	Beds, n
Private hospital	150	185	16,000	23,000	50
Private hospital	90	170	15,000	20,000	45
Private hospital	70	180	15,000	22,000	50
Public hospital	250	400	11,000	15,000	330
Private hospital	180	320	20,000	31,000	445
Private hospital	120	400	24,000	30,000	200

Table 2. Implementation of EHR features among the hospitals in this study.

EHR[a] feature	Hospitals (N=6), n
View laboratory results electronically	6
Collect patient demographics electronically	6
List patient problems electronically	6
Highlight out-of-range laboratory results	6
List patient medications electronically	5
Create medical history and follow-up notes	5
Order laboratory tests electronically	5
Order radiology tests electronically	5
View electronic clinical notes electronically	4
View imaging results electronically	4
View allergy lists	4
Warn of drug interactions and contraindications	3
Personal human resources (payroll, benefits, training)	3
Preregister systems	3
Send prescriptions electronically to a pharmacy	2
View immunization records	1
Send medical information securely to other health care professionals	1
Receive reminders for guideline-based interventions	1
Receive reminders for preventive screening tests	1
Sending electronic referrals	1
Receive reminders for screens of chronic disease management	0
Sending electronic clinical messaging to patients	0
Remote access to EHR	0

[a]EHR: electronic health record.

Research Design

This study was reviewed and approved by the involved hospitals. Data collection was conducted in the six hospitals using an online survey during May and August 2020. For data collection, we could not reach out to all physicians and nurses working in the six hospitals for two reasons. First, due to the COVID-19 pandemic, a large number of clinicians did not work in the same shifts as they worked before the outbreak. Second, many physicians and nurses of the hospitals did not work at the COVID-19 specialized wards. Therefore, our clinician population size was shortened to clinicians working in different shifts at the COVID-19–dedicated wards. Based on the information provided by the six hospitals' administrations, the

total population size was 586 clinicians. We asked the administrators of the COVID-19 specialized wards to distribute the survey online among all clinicians working in different shifts at the COVID-19 wards during May and August 2020 (4 months).

Since WhatsApp is a popular communication app in Iran, the survey was mainly administered via WhatsApp to physicians (with different specialties) and nurses who were directly exposed to patients who were suspected of or infected with COVID-19. The survey was designed to measure facets of HIT use, including EHR functionality, electronic prescribing, health information exchange, other technology-based tools, and informatics applications.

Survey questions also examined the effects of using HIT on workflow, patient care, and clinician burnout during the COVID-19 pandemic. The survey was first prepared in English; then, it was translated to Farsi by one of the authors. The survey was reviewed by a physician working in Iran, and based on the feedback, some modifications were made for clarity purposes. Next, the survey was translated back to English by a university professor in Iran to ensure the correct terms were used and the questions were clear and understandable to participants. Before data collection, survey questions were reviewed, evaluated, and approved by the IT departments and hospital administrations. Finally, the survey was conducted anonymously at the individual clinician level. The survey questions are provided in Multimedia Appendix 1.

In this study, the dependent variable was clinician burnout. We used a single item to measure clinician burnout. This single question was adapted from the physician's work–life study [21,22]. The measure has six answers: (1) no symptoms of burnout, (2) under stress but do not feel burned out, (3) having one or more symptoms of burnout such as emotional exhaustion, (4) thinking about work frustrations as symptoms of burnout will not go away, (5) feeling completely burned out and seeking help, and (6) completely burned out and getting help now [23]. Following previous studies [18], we dichotomized this measure into "no symptoms of burnout" (≤2 on the 5-point scale) and "one or more symptoms of burnout" (≥3 on the 5-point scale). This single-item measure was validated for clinicians and exhibited a better sensitivity and specificity compared with the Maslach Burnout Inventory survey [24].

Next, we were interested in identifying the factors that influence burnout during the COVID-19 pandemic. Indicators of burnout include awareness of EHR features, EHR system usability, concern about COVID-19, use of technology solutions, use of hospital technology interventions, hospital preparedness, and professional efficacy. Table 3 shows the description of each indicator.

Table 3. Description and source of variables used in this study.

Measures	Description	Sources
Awareness of EHR[a] features	The extent to which participants are aware of available EHR features	[25]
EHR system usability	Adapted from the System Usability Scale: giving a global view of subjective assessments of EHR system usability	[26]
Concerned about COVID-19	The level of concern of the participants about the effect of the COVID-19 pandemic in their life	[7]
Use of technology solutions	The extent to which individuals use technology solutions for coping with work-related stress	[27,28]
Use of hospital technology interventions	The extent to which individuals agree with the use of different organizational interventions at their workplace (hospital)	[29,30]
Perceived hospital preparedness	The extent to which clinician perceives the hospital is prepared to deal with the COVID-19 crisis	[31]
Professional efficacy	Adapted from Maslach Burnout Inventory-General Survey: the extent to which an individual's perceived effectiveness and accomplishment at work	[32]

[a]EHR: electronic health record.

We included several control variables such as patients' demographics and practice setting. Respondents provided information about their age, gender, marital status, years of practice, role, specialty, area of work, and hospital setting.

There is considerable debate in the health informatics community about the actual role health information technologies play in clinician burnout [1]. We are interested in identifying the specific factors related to EHRs that may cause stress among clinicians. Further, once we understand the causes of burnout, we are left to ponder innovative solutions to prevent or mitigate burnout. Existing research suggests that some technological advances may help to reduce burnout. However, more studies need to investigate the confounding effect of regulatory mandates and interventions that aim at improving communication, changing workflow, or addressing clinician concerns via quality improvement projects [33]. Therefore, we added additional questions in the survey to understand EHR-related causes of stress, other technology-based solutions, and policy interventions that may help clinicians reduce burnout. Respondents were asked to report stressors associated with the use of EHRs on a 5-point scale. The factors causing stress as a result of using EHRs were increasing computerization of practice, too much time spent on EHRs at work, spending an enormous amount of time on the EHRs at home, insufficient time for documentation, too much data entry, having less face-to-face time for conversation and examination with patients,

and inadequate technology-related training for using EHRs at work. These measures were adopted from previous studies [23].

Using a 5-point scale, participants were asked to indicate to what extent using technology-based solutions can help them cope with work-related stress. The solutions were availability of training programs; availability of responsive IT support; possibility for telecommuting (working from home); using technology and tools to follow up with patients remotely (telemedicine); use of mobile apps for meditation, breathing, and relaxation; availability of help desk services; mentorship programs through teleconsultation; and communication groups via messaging apps (eg, WhatsApp, Telegram, and Viber).

Using a 5-point scale, we asked respondents to evaluate the effectiveness of the hospital's policies, strategies, and interventions regarding the implementation and use of technology-based tools and informatics infrastructure at the workplace. The policies and interventions were developing standards for order entry and reporting among hospitals; using data analysts (specialists) to analyze patient data and elaborate on the patterns; designing an online survey to regularly measure satisfaction with technology and possible technology-related risks and stress; establishing a regular assessment to evaluate the effectiveness of technology in hospitals; using a systematic way to measure workplace burnout and analyze the results; formulating transparent policies, expectations, and goals regarding the use of technology in the clinical workflow; applying clear regulations about the responsibility of clinicians when medical errors occur using technology; and providing incentives for using technology meaningfully for health care delivery. These items were followed by an open box for the participants to add effective interventions or strategies.

Statistical Analysis

We used SPSS Statistics V21.0 (IBM Corp) to conduct all statistical analyses. We used univariable statistics to describe the respondents' demographic characteristics, the prevalence of burnout, general causes of stress at the workplace, EHR-related stress, and technology solutions to cope with stress. We used multivariable logistic regression to measure the association between burnout and measures of respondents' awareness of available EHR features, perceived EHR system usability, level of concerns about COVID-19, use of technology solutions for coping with stress, use of technology at the workplace, hospital preparedness for the pandemic, and perceived professional efficacy. We controlled the model for respondents' age, gender, marital status, role, area of work, specialty, and the number of years they have been in practice to ensure that these factors did not create bias in the results. We performed a sensitivity analysis using an ordered logit model to examine whether the dependent variable (burnout) was represented by its ordinal response categories instead of the dichotomized response categories included in the primary analysis. We also tested the burnout association with the number of patients with COVID-19 that have been cared for by the respondents to identify whether more patients are independently associated with burnout. The results of the sensitivity analysis are reported in Multimedia Appendix 2. Further, to ensure the robustness of the coefficients, we repeated the analysis by considering within-hospital correlations and controlling for the hospital effect. Results are reported in Multimedia Appendix 3 and show that all coefficients remain robust.

Results

Due to a large amount of abrupt tension and irregularity caused by the pandemic, only 373 clinicians initially attempted the survey. After removing incomplete answers, we finally collected 368 fully completed surveys that included 147 nurses, 161 physicians, and 60 physician assistants providing care for patients with COVID-19 (Table 4). The majority of the respondents were female (n=266, 72%), and 68% (n=252) were married. Among the respondents, 22% (n=80) were younger than 35 years, 42% (n=154) were aged between 35 and 44 years, 21% (n=78) were aged between 45 and 54 years, and 16% (n=56) were 55 years or older. There was about an equal number of respondents working in emergency departments, intensive care, and operating rooms (all n=67, 18%), while 34% (n=124) reported working in other inpatient services, and 12% (n=43) reported working in outpatient services. About a quarter of the respondents reported having less than 5 years of experience in practice, while 20% (n=73) reported between 6 to 10 years of experience, 19% (n=71) reported between 11 to 15 years of experience, 22% (n=82) between 16 and 20 years of experience, and 13% (n=46) reported having more than 20 years of experience in practice. The majority of the respondents were from private hospitals (n=310, 84%). About 43% (n=159) of the respondents reported having at least one extra night shift per week since the beginning of the pandemic.

Table 4. Characteristics of respondents.

Characteristics	Sample (N=368), n (%)
Age (years)	
<35	80 (22)
35-44	154 (42)
45-54	78 (21)
55-64	35 (10)
≥65	21 (6)
Gender	
Male	102 (28)
Female	266 (72)
Marital status	
Married	252 (68)
Single	116 (32)
Role	
Nurse	147 (40)
Physician	161 (44)
Physician assistant	60 (16)
Area of work	
Emergency department	67 (18)
Intensive care unit	67 (18)
Other inpatient services	124 (34)
Outpatient services	43 (12)
Operating rooms	67 (18)
Specialty	
Emergency medicine	27 (7)
Family medicine	38 (10)
Surgery	56 (16)
Anesthesiology	67 (18)
Gynecology	50 (14)
Nursing	56 (14)
Years in practice	
<1	3 (0.01)
1-5	96 (26)
6-10	73 (20)
11-15	71 (19)
16-20	82 (22)
>20	46 (13)
Hospital setting	
Public hospital	58 (16)
Private hospital	310 (84)
Burnout prevalence	
I *enjoy* my work. I have no symptoms of burnout.	115 (31)
I am under *stress* and don't always have as much energy as I did, but I don't feel burned out.	119 (32)

Characteristics	Sample (N=368), n (%)
I am *definitely burning out* and have one or more symptoms of burnout (eg, emotional exhaustion).	66 (18)
The symptoms of burnout I am experiencing won't go away. I think about work *frustrations* a lot.	24 (7)
I feel completely burned out. I am at the point where I may need to *seek help*.	28 (8)
I am completely burned out, and *I am getting help*.	16 (4)

A total of 36% (134/368) of the respondents reported having at least one symptom of burnout. The highest level of burnout was reported among the physician assistants (32/60, 53%), followed by nurses (Figure 1).

Regarding the area of work, most burnout symptoms were reported in intensive care units, followed by emergency department and other outpatient services (Figure 2).

Figure 1. Percent of respondents reporting at least one symptom of burnout in their role.

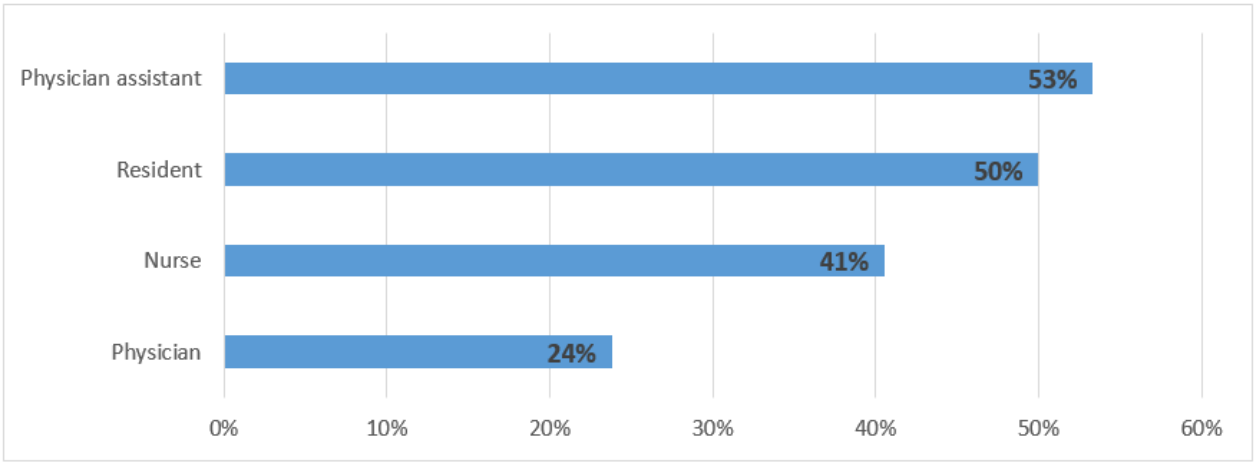
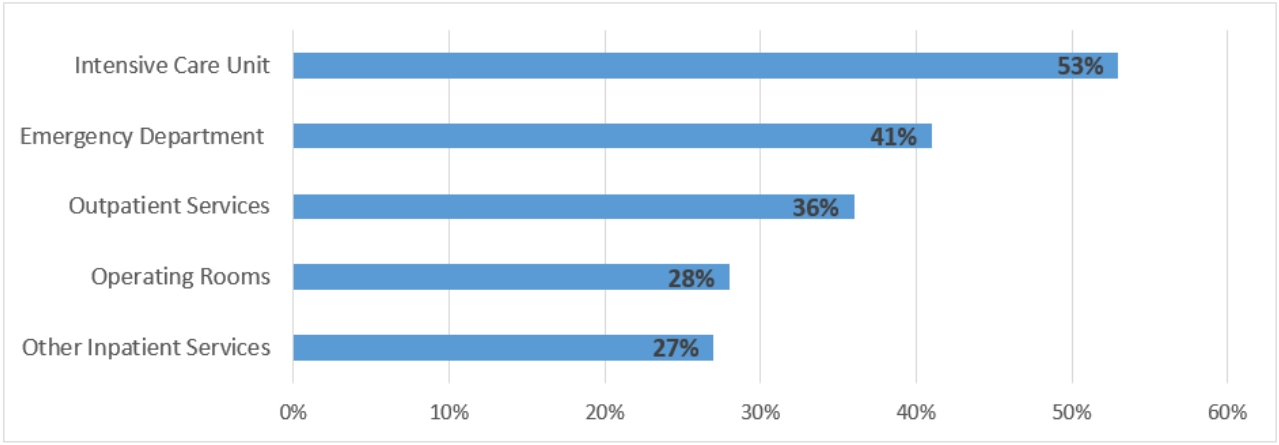
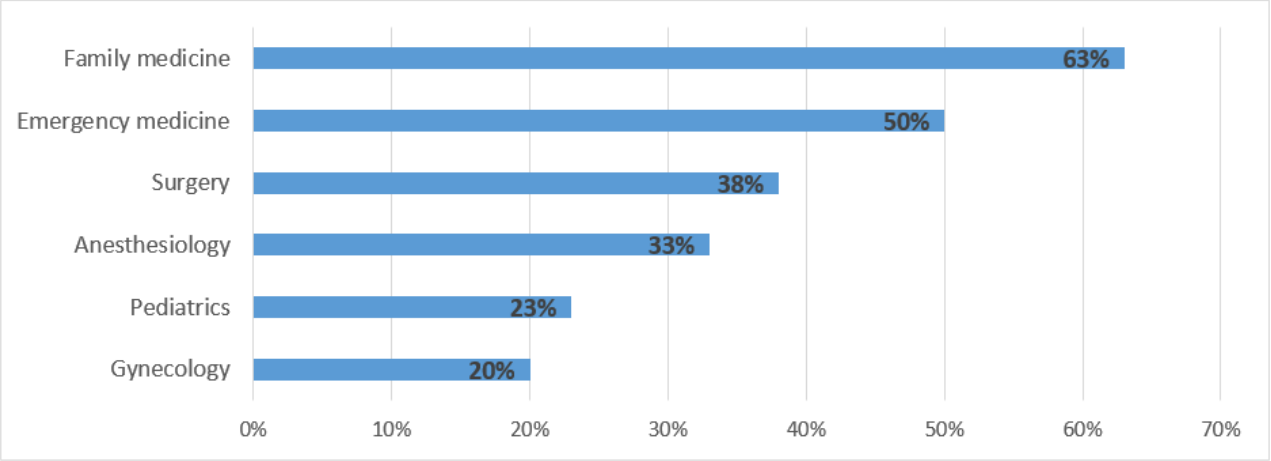


Figure 2. Percent of respondents reporting at least one symptom of burnout by area of work.



The specialties who experienced the most from burnout practiced family medicine and emergency medicine, followed by surgery and anesthesiology (Figure 3). Results show that family medicine physicians experienced the most from burnout. This finding is in line with several previous studies [34]. Family medicine physicians usually feel the most stress because they need to provide comprehensive health care for people of all ages. During the pandemic, family medicine visited more patients with similar symptoms (eg, common cold or flu). At the time of our study, people who felt that they might be at risk of being COVID-19 positive usually visited their trusted family medicine first for a check-up or advice. This also has some cultural reasons because family doctors are generally considered the most trusted physicians for families and the first choice to visit in case of an ailment. Thus, family medicine physicians visited various patients across all ages, genders, and diseases, and they felt the most stress and frustration under the pandemic situation.

Figure 3. Percent of respondents reporting at least one symptom of burnout by specialty.



Before running the logistic regression model, we assessed the reliability of the proposed measures. Table 5 presents some indicators and descriptive statistics for each measure. We assessed the reliability using Cronbach alpha for all measures. The recommended threshold value is above .70 [35], which implies adequate reliability for all measures in this study. For awareness of EHR features, since it was measured as a set of dichotomous items, we separately calculated an aggregate score adjusted for each hospital and used that score in the logistic regression model to assess its impact on burnout. Higher scores show a higher level of awareness about the available EHR features.

Table 5. Summary of the properties of the measurements.

Measures	Measurement type	Mean (SD)	Cronbach α
EHR[a] system usability	10-item scale measured on a 5-point Likert scale	3.20 (0.71)	.85
Concerned about COVID-19	4 items measured on 5-point Likert scale	4.43 (0.67)	.72
Use of technology solutions	8 items measured on a 5-point Likert scale	3.12 (0.80)	.81
Use of hospital technology interventions	8 items measured on a 5-point Likert scale	2.89 (0.85)	.86
Hospital preparedness	5 items measured on a 5-point Likert scale	3.50 (0.86)	.87
Professional efficacy	4 items measured on a 5-point Likert scale	3.75 (0.79)	.88

[a]EHR: electronic health record.

Next, based on the multivariable logistic regression model while controlling for age, gender, marital status, role, area of work, specialty, years of experience, and type of hospital (Table 6), we identified that perception of EHR system usability, level of concern about COVID-19, hospital technology-related interventions, and level of hospital preparedness to cope with the pandemic are the significant factors associated with burnout among the clinicians. Positive perceptions about the ease of use of EHRs and a high level of confidence in understanding its function were associated with lower odds of burnout symptoms (OR –0.16, 95% CI –0.24 to –0.09; *P*<.001). In other words, the nurses and physicians who find EHRs useable and are confident in their knowledge and ability to work with it have lower odds of having any burnout symptoms compared to those who find EHRs difficult and cumbersome to use. Similarly, the more use of technology interventions implemented in the hospitals was associated with lower odds of burnout (OR –0.38, 95% CI –0.51 to –0.26; *P*<.001).

Table 6. The estimate of the association between demographic, practice, and EHR characteristics and 1 or more symptoms of burnout.

Variable	OR[a] (95% CI)	SE	*Z* value	*P* value
Awareness of EHR[b] features	0.03 (–0.07 to 0.13)	0.05	0.54	.59
EHR system usability	–0.16 (–0.24 to –0.09)	0.04	–4.16	<.001
Concerned about COVID-19	0.29 (0.11 to 0.48)	0.09	3.05	.002
Use of technology solutions	–0.06 (–0.14 to 0.03)	0.04	–1.31	.19
Use of hospital technology interventions	–0.38 (–0.51 to –0.26)	0.06	–5.95	<.001
Hospital preparedness	–0.23 (–0.38 to –0.09)	0.07	–3.05	.002
Professional efficacy	–0.06 (–0.23 to 0.11)	0.09	–0.70	.49
Age (years; reference group: ≥65 years)				
<35	–0.11 (–3.48 to 3.09)	1.65	–0.07	.95
35-44	–2.13 (–5.49 to 0.79)	1.58	–1.35	.18
45-54	–0.88 (–4.05 to 1.98)	1.51	–0.59	.56
55-64	1.61 (–1.35 to 4.73)	1.51	1.06	.29
Gender (reference group: female)				
Male	–0.75 (–1.98 to 0.44)	0.61	–1.22	.22
Marital status (reference group: single)				
Married	–0.30 (–1.35 to 0.72)	0.53	–0.57	.57
Role (reference group: physician assistant)				
Nurse	–0.93 (–2.24 to 0.34)	0.66	–1.42	.16
Physician	–2.34 (–3.96 to –0.86)	0.78	–2.99	.003
Area of work (reference group: operating rooms)				
Emergency department	–1.38 (–2.89 to 0.05)	0.75	–1.85	.07
Intensive care unit	0.55 (–1.06 to 2.2)	0.83	0.66	.51
Other inpatient services	–1.96 (–3.27 to –0.73)	0.65	–3.04	.002
Outpatient services	–1.18 (–2.95 to 0.49)	0.87	–1.36	.18
Specialty (reference group: nursing)				
Emergency medicine	–1.47 (–3.61 to 0.49)	1.04	–1.41	.16
Family medicine	–0.76 (–2.8 to 1.27)	1.03	–0.73	.46
Surgery	–2.25 (–4.42 to –0.25)	1.05	–2.14	.03
Anesthesiology	–1.14 (–2.41 to 0.09)	0.63	–1.80	.07
Gynecology	0.83 (–0.62 to 2.34)	0.75	1.11	.27
Years in practice (reference group: 16-20)				
<1	–1.25 (–3.57 to 1.12)	1.18	–1.06	.29
1-5	1.33 (–0.79 to 3.61)	1.11	1.21	.23
6-10	–0.93 (–3.02 to 1.2)	1.06	–0.88	.38
11-15	–0.12 (–2.12 to 1.95)	1.02	–0.12	.91
Type of hospital (reference group: public)				
Private	–0.21 (–1.65 to 1.19)	0.71	–0.29	.77
(Intercept)	11.30 (6.35 to 16.86)	2.65	4.27	<.001

[a]OR: odds ratio.

[b]EHR: electronic health record.

We realize that the nurses and physicians working in hospitals that implemented various interventions regarding the use of technology at the workplace reveal lower odds of having any symptom of burnout compared to those working in hospitals

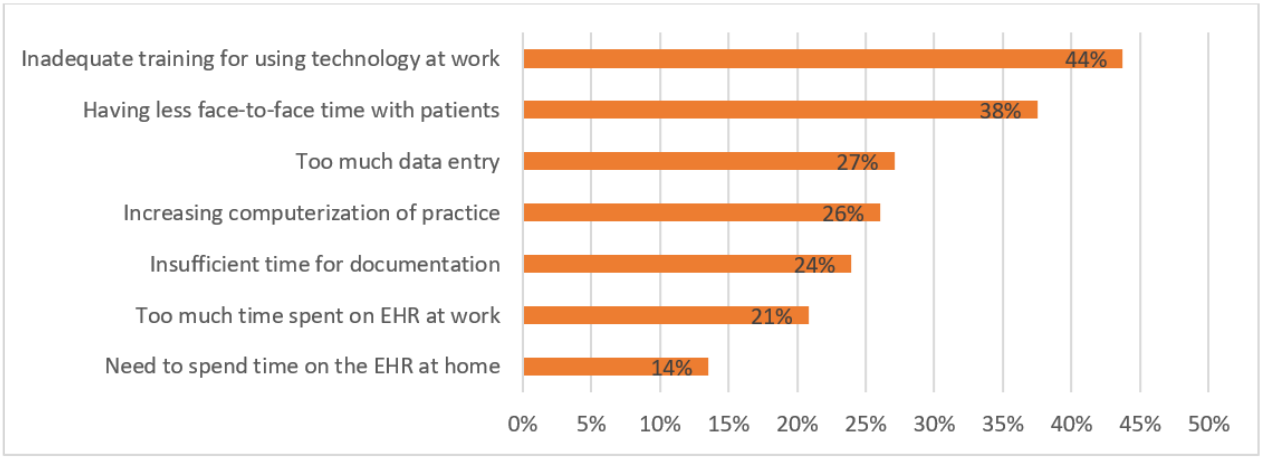
that did not implement such interventions. Hospital preparedness was also associated with a reduction in burnout symptoms (OR –0.23, 95% CI –0.38 to –0.09; *P*=.002). Regarding the level of concern about COVID-19, we identified that the more the clinicians are concerned about the pandemic situation, the higher the odds of burnout (OR 0.29, 95% CI 0.11-0.48; *P*=.002). We did not find a significant association between the odds of having at least one symptom of burnout and the awareness about EHR features, use of technology solutions, and respondents' professional efficacy.

Among the control variables, physicians (as a role), other inpatient services (as an area of work), and surgery (as a specialty) exhibited significantly lower burnout symptoms. Among the work areas, health care workers in other inpatient centers (eg, services delivered in rehabilitation centers, psychiatric departments, or addiction treatment centers) exhibited lower burnout.

We asked respondents about the factors that cause stress in their work in general prior to the pandemic. Almost one-fifth of the participants listed "insufficient compensation" or "family responsibilities" as the leading cause of stress. Almost one-fourth of the participants listed "workload" and almost one-fifth listed "health risks" as the second most important stressors. During the pandemic, the clinicians' workload increased, and the concerns over contracting the virus or transmitting it to family members amplified. Almost one-fifth of the participants reported that they have work on average about 20 hours per week during the pandemic compared to the before the pandemic. Almost one-fourth of the participants reported that they have to work over the weekends more often during the pandemic compared to the time before it. Almost half of the participants reported an increase in the number of nights that they needed to be on call per week during the pandemic.

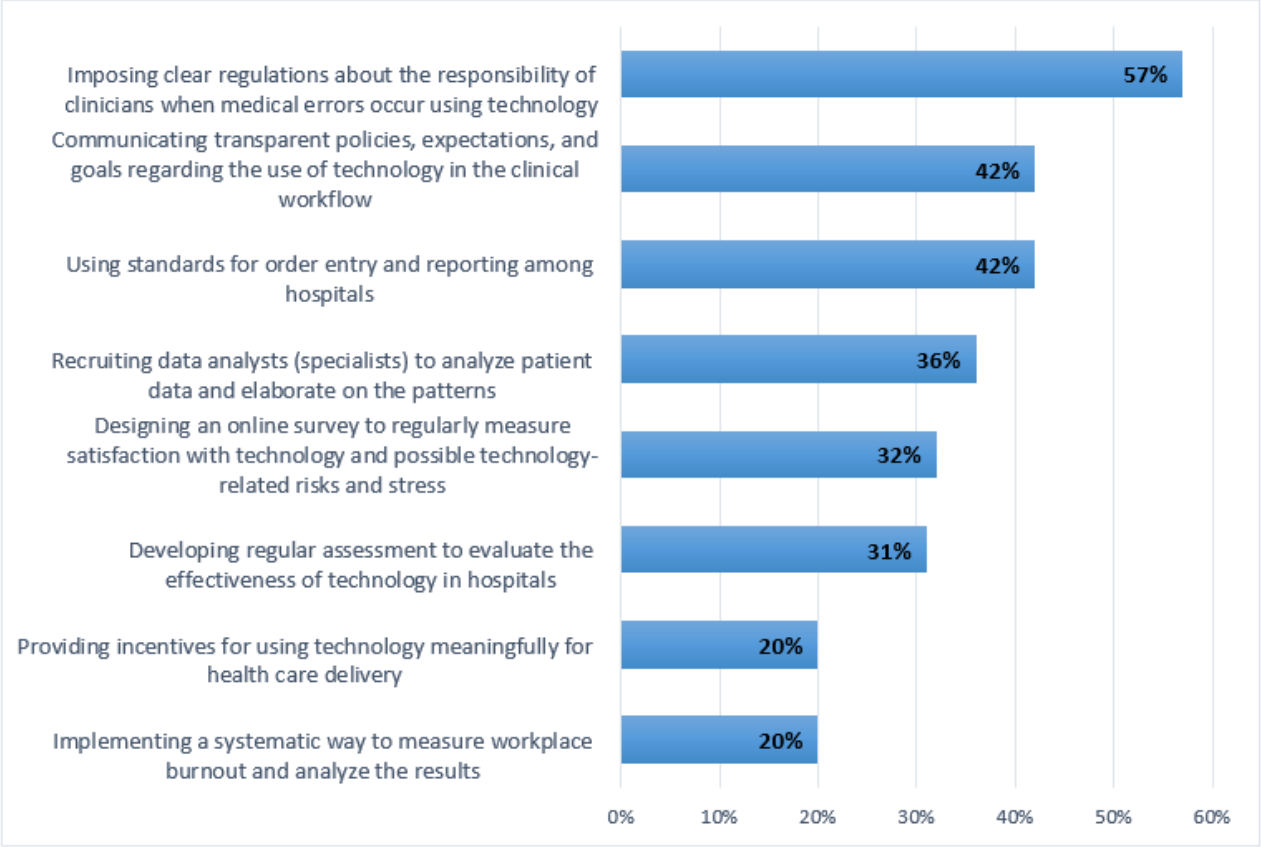
We were also interested in understanding the EHR-related causes of stress during the pandemic. Figure 4 shows several EHR-related stress factors during the pandemic. During the pandemic, an extension was made to EHRs to register and keep track of COVID-19 tests administered and to send the results electronically to a national online repository. For the tests with positive results, the record of hospitalization or death was populated in the system. Respondents strongly agreed that the most important leading cause of EHR-related stress is inadequate training for using technology, followed by having less face-to-face time with patients, too much time spent on data entry, and increasing computerization at work. Since during the pandemic clinicians' shifts changed and they had an additional workload with an unclear use policy, the likelihood of creating errors through using EHRs increased. Since the implementation of EHRs is still at an early stage, respondents mentioned inadequate training for using EHR features. This training issue, coupled with other key issues such as having less face-to-face time with patients (due to social distancing) and involving too much data entry (COVID-19 test results), caused more stress during the pandemic compared to normal times since the workload increased and job pressure was higher. The situational factors (due to the pandemic) exacerbated stress associated with EHR use in hospitals.

Figure 4. Leading causes of EHR-related stress during the COVID-19 pandemic. EHR: electronic health record.



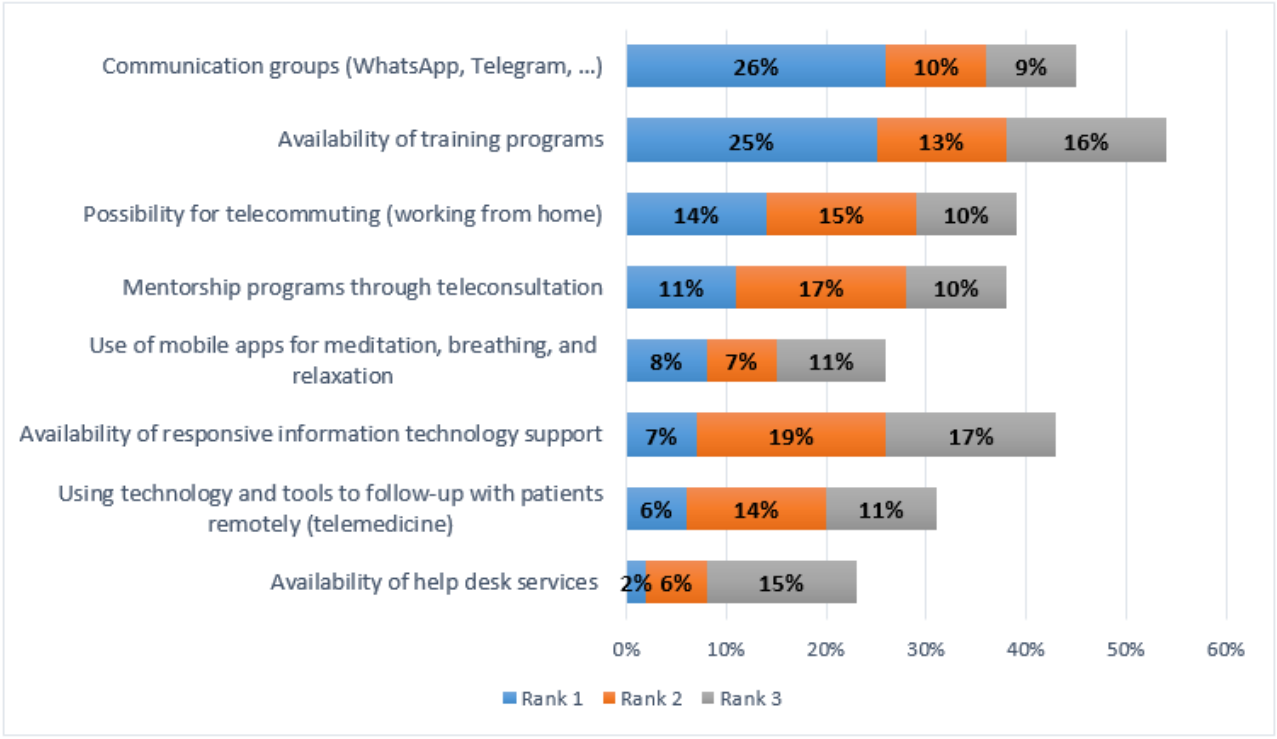
Next, we asked participants about the effectiveness of implementing different interventions at the hospitals during the pandemic. Respondents believed that the hospitals' most effective intervention was implementing clear regulations about the responsibility of clinicians when medical errors occur using technology. Clinicians are more prone to medical errors when they work under pressure and stress imposed by the pandemic. Therefore, the hospitals need to develop and implement proper regulations and guidelines, and communicate them clearly with the clinicians. Other effective interventions include imposing transparent policies, expectations, and goals regarding the use of technology in the clinical workflow and developing unified standards for order entry and reporting among hospitals (Figure 5).

Figure 5. Effective hospital interventions during the COVID-19 pandemic.



The most effective solutions that the respondents selected for addressing their burnout symptoms were using communication tools that facilitate group interactions (eg, WhatsApp and Telegram) and providing training programs for learning EHR systems. Respondents identified the second most effective solution as the availability of telecommuting (working from home), followed by participating in mentorship programs through teleconsultation (Figure 6). These findings show that the most prominent solution for reducing burnout among clinicians during the pandemic is facilitating the communication and information channels for them in a fast-paced environment to get the timely information they need to perform their activities. To practice social distancing, which is one of the main factors that can help reduce the transmission of COVID-19, the clinician needs to provide some of the care services remotely through telemedicine or teleconsultation. Our results show that special training to facilitate these activities for the clinician seems to be among the essential solutions to reduce clinicians' burnout.

Figure 6. Effective technology solutions during the COVID-19 pandemic.



Discussion

Statement of Principal Findings

The COVID-19 pandemic imposed serious challenges to the health care systems of numerous countries. As this virus spreads worldwide, more research is needed to address clinicians' burnout during the pandemic. This study aims to investigate burnout issues among physicians and nurses working in Iranian hospitals that directly provide care for patients suspected of or infected with COVID-19. In this study of 368 clinicians working in hospitals in Tehran, we found that EHR usability, technology-based hospital interventions, hospital preparedness, and level of concern about COVID-19 were significantly associated with workplace burnout. This finding is consistent with previous studies suggesting poor EHR usability will lead to clinicians' dissatisfaction and, in turn, burnout [36]. EHRs have also been found to be a useful tool to support outbreak management [16]. Clinicians who believe that EHRs are not unnecessarily complicated and that they can have the support of technical personnel to use the system are less likely to experience burnout symptoms. Moreover, clinicians who consider that the various functions in EHRs are well integrated and that they can learn the features quickly are less prone to exhibit burnout symptoms. Given the negative association between EHR usability and burnout, hospitals need to improve EHR user-friendliness and convenience to reduce clinician burnout. Based on this study's findings, we recommend that hospitals measure EHR usability and burnout among their clinicians in a regular and systematic way. Those clinicians who reported higher burnout symptoms could be the subject of a focus group to determine what changes are required by hospitals in implementing EHRs (for instance, availability of EHR-related training or technical support).

Interpretation Within the Context of the Wider Literature

This study shows that the more caregivers are familiar with the EHR systems and feel confident in their ability to use them, the less they feel burnout symptoms. Interestingly in our sample, the most effective solution for reducing burnout symptoms was providing a training program for EHR use. Our results are in line with recent studies on EHR-related perceptions of workload that suggest individualized EHR training can improve the knowledge of EHR tools and satisfaction [37,38]. Other studies indicated that a considerable amount of time that the care providers need to spend recording information in the EHR system and data entry was the essential contributor to burnout [18]. However, our study reveals that the respondents are less worried about the data entry and their time using the system either at work or at home. They feel that more useful education and training programs are required to improve their symptoms of burnout. We also call attention to other important solutions to address burnout, such as easily accessible mentorship programs through teleconsultation and the availability of responsive IT support. Advanced EHR education through highly interactive personalized mentorship and hands-on workshops can help care providers alleviate workplace burnout and improve care quality [39,40]. EHR mentorship programs may improve the care providers' self-efficacy, and the availability of 24-hour IT support provides the peace of mind that help is always accessible for them. This solution will help them combat burnout symptoms and psychological stress.

Our findings suggest that technology-related hospital interventions can decrease burnout among clinicians. We highlight several potential interventions to reduce the odds of burnout. Hospitals need to develop shared standards for data entry and exchanging health information among hospitals [41].

This intervention plays a significant role in the pandemic since prompt information sharing between hospitals and other health care organizations is required. Hospitals can also recruit business intelligence [38] specialists to analyze patient data, extract patterns, and illuminate risk factors and trends. Using reporting and analytics tools during the outbreak, data analysts can help hospitals devise better strategies by analyzing patient data from different perspectives (eg, location, symptoms, and contact with COVID-19–positive patients). Another possible intervention by hospitals could be using an online survey to regularly measure satisfaction with the technology used in their clinical setting. This approach can help hospitals identify potential technology-related risks and stress, reducing the likelihood of clinician burnout. Furthermore, hospitals can frequently evaluate the effectiveness of current technology. This approach provides hospitals with a better assessment of technology from the users' perspectives. For instance, during the pandemic, hospitals may recognize the need to implement multiple COVID-19–specific tools (eg, an EHR-integrated patient self-triage and self-scheduling tool to manage COVID-19) [8].

Implications for Policy, Practice, and Research

Based on the findings, a significant hospital intervention could be applying a systematic method for measuring workplace burnout and analyzing the results. Clinician burnout surveys could be conducted frequently. However, the findings and feedback could remain uninvestigated, and it will be unclear how burnout affects health care professionals' productivity [42]. For instance, during the COVID-19 pandemic, clinicians may encounter new stressors that might not have been recognized previously [7]. Discovering stressors and analyzing the source of stress (eg, technology-related) can be useful for a rapid response to clinicians' burnout during the COVID-19 outbreak.

Hospitals should establish transparent policies, expectations, and goals regarding the use of technology-based tools in the clinical workflow. For instance, the primary purpose of EHR implementation, the extent to which EHR features should be used, and clear regulations and guidelines for using EHRs should be communicated to clinicians working in hospitals. Otherwise, the lack of clear use strategies regarding clinician–technology interactions can lead to confusion, stress, and burnout [43]. First, we recommend that hospitals clarify clinicians' responsibility, accountability, and associated regulations when medical errors occur using technology (eg, EHRs) in the clinical setting. Second, hospitals should define an *EHR meaningful use program* to elucidate what features of EHRs are essential and why their use is mandatory and, moreover, what applications are advanced or voluntary. Third, policy makers can offer incentives to hospitals and clinicians for the meaningful use of technology-based tools for health care delivery. These guidelines and instructions not only set a direction for hospitals on how to successfully implement and use EHRs but also mitigate the likelihood of clinician burnout. We need to acknowledge that any new policies, instructions, and procedures could be sources of burnout at the beginning (during a pandemic in particular). The success factor is implementing the new strategies, guidelines, and processes based on a robust execution plan, enough resources, and transparent communications and collaborations among clinicians in hospitals.

Clinicians working in the COVID-19 ward are significantly worried about becoming infected or concerned about their family becoming infected. This high level of concern about the pandemic significantly raises their burnout symptoms. The findings also demonstrate that hospital preparedness can enable the efficiencies of the COVID-19 care management program by reducing clinician burnout symptoms. Burnout is a system problem [44], and hospitals need to pose rapid solutions to address it during the pandemic. Hospitals can mitigate the odds of burnout among physicians and nurses by providing useful pandemic training, protective equipment, and support programs. Pandemic training (eg, classes, brochures, and meetings) can affect clinician stress by disseminating helpful statistics, insights, and information about the epidemic [45]. Through protective equipment (eg, personal protective equipment, masks, gloves, and sanitizer), hospitals are able to decrease workforce anxiety by providing clinicians facing patients with COVID-19 with various tools to avoid infection [46]. Support programs (eg, online safety alerts and wellness guidance, digital access to benefits, and work from home practices) can reduce burnout by improving work flexibility and helping clinicians access EHRs and other technology across distances [47].

Limitations

This study is subject to some limitations. First, due to data collection limitations, our research is conducted based on a small group of physicians and nurses facing COVID-19. Second, this study is conducted in a resource-limited country with different pandemic preparedness plans and technology infrastructure in the health care system. Therefore, caution should be exercised when generalizing the findings to other countries. Third, because this study was conducted in the critical episode of the epidemic in Iran, selection bias may have skewed our results if highly stressed clinicians or those with little stress decided not to participate. Fourth, we conducted data collection electronically; thus, clinicians who were more comfortable with computers or mobile devices were more likely to participate. Finally, due to the time of data collection, our findings may not be generalized to the onset or the time that COVID-19 is under control in Iran. We recommend that future studies extend our research in the same or other contexts by collecting more comprehensive data using different data collection strategies, considering additional informatics-based tools, and examining further epidemic factors at different times.

Conclusions

Studying technology-related factors affecting clinicians' burnout has attracted much attention in the health care setting. This study adds to the literature by examining how EHR usability, technology-based hospital interventions, hospital preparedness, and concern about COVID-19 are significantly associated with burnout among clinicians during the COVID-19 pandemic in a resource-limited country. The results show the critical role of transparent hospital policies, EHR implementation strategies, training needs, and hospital pandemic support programs in reducing burnout. Interestingly, transparency is identified as the most important intervention that hospitals should implement to address burnout symptoms. Transparency in regulations addressing medical errors using technology and transparency

dealing with policies and expectations of clinical workflows were identified as critical factors to alleviate workplace burnout. Due to the critical effects of burnout on physicians, nurses, patients, and the health care system, hospitals need to design robust interventions to address stress generated by HIT. Through this process, hospitals can mitigate the clinician burden and improve quality of care and patient safety.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey questions.
[DOCX File , 27 KB - medinform_v9i6e28497_app1.docx]

Multimedia Appendix 2
Sensitivity analysis.
[DOCX File , 19 KB - medinform_v9i6e28497_app2.docx]

Multimedia Appendix 3
Results after controlling for a hospital effect.
[DOCX File , 23 KB - medinform_v9i6e28497_app3.docx]

References

1. National Academies of Sciences, Engineering, and Medicine. Taking Action Against Clinician Burnout: A Systems Approach to Professional Well-Being. Washington, DC: The National Academies Press; 2019.
2. Salyers MP, Bonfils KA, Luther L, Firmin RL, White DA, Adams EL, et al. The relationship between professional burnout and quality and safety in healthcare: a meta-analysis. J Gen Intern Med 2017 May;32(4):475-482 [FREE Full text] [doi: 10.1007/s11606-016-3886-9] [Medline: 27785668]
3. Friedberg M, Chen P, Van Busum KR, Aunon F, Pham C, Caloyeras J, et al. Factors affecting physician professional satisfaction and their implications for patient care, health systems, and health policy. Rand Health Q 2014;3(4):1 [FREE Full text] [Medline: 28083306]
4. Shanafelt T, Noseworthy JH. Executive leadership and physician well-being: nine organizational strategies to promote engagement and reduce burnout. Mayo Clin Proc 2017 Jan;92(1):129-146. [doi: 10.1016/j.mayocp.2016.10.004] [Medline: 27871627]
5. He F, Deng Y, Li W. Coronavirus disease 2019: what we know? J Med Virol 2020 Jul;92(7):719-725 [FREE Full text] [doi: 10.1002/jmv.25766] [Medline: 32170865]
6. Adams JG, Walls RM. Supporting the health care workforce during the COVID-19 global epidemic. JAMA 2020 May 21;323(15):1439-1440. [doi: 10.1001/jama.2020.3972] [Medline: 32163102]
7. Wu Y, Wang J, Luo C, Hu S, Lin X, Anderson AE, et al. A comparison of burnout frequency among oncology physicians and nurses working on the frontline and usual wards during the COVID-19 epidemic in Wuhan, China. J Pain Symptom Manage 2020 Jul;60(1):e60-e65 [FREE Full text] [doi: 10.1016/j.jpainsymman.2020.04.008] [Medline: 32283221]
8. Judson T, Odisho A, Neinstein A, Chao J, Williams A, Miller C, et al. Rapid design and implementation of an integrated patient self-triage and self-scheduling tool for COVID-19. J Am Med Inform Assoc 2020 Jun 01;27(6):860-866 [FREE Full text] [doi: 10.1093/jamia/ocaa051] [Medline: 32267928]
9. Chen Q, Liang M, Li Y, Guo J, Fei D, Wang L, et al. Mental health care for medical staff in China during the COVID-19 outbreak. Lancet Psychiatry 2020 Apr;7(4):e15-e16 [FREE Full text] [doi: 10.1016/S2215-0366(20)30078-X] [Medline: 32085839]
10. Koh D. Occupational risks for COVID-19 infection. Occup Med (Lond) 2020 Mar 12;70(1):3-5 [FREE Full text] [doi: 10.1093/occmed/kqaa036] [Medline: 32107548]
11. Nguyen L, Bellucci E, Nguyen LT. Electronic health records implementation: an evaluation of information system impact and contingency factors. Int J Med Inform 2014 Dec;83(11):779-796. [doi: 10.1016/j.ijmedinf.2014.06.011] [Medline: 25085286]
12. Triantafillou P. Making electronic health records support quality management: a narrative review. Int J Med Inform 2017 Aug;104:105-119. [doi: 10.1016/j.ijmedinf.2017.03.003] [Medline: 28599812]
13. Guan W, Ni Z, Hu Y, Liang W, Ou C, He J, et al. Clinical characteristics of coronavirus disease 2019 in China. N Engl J Med 2020 Apr 30;382(18):1708-1720. [doi: 10.1056/nejmoa2002032]
14. Amir-Behghadami M, Janati A. Iranian national COVID-19 electronic screening system: experience to share. Emerg Med J 2020 Jul;37(7):412-413 [FREE Full text] [doi: 10.1136/emermed-2020-209806] [Medline: 32434767]

15. Lenert L, McSwain BY. Balancing health privacy, health information exchange, and research in the context of the COVID-19 pandemic. J Am Med Inform Assoc 2020 Jun 01;27(6):963-966 [FREE Full text] [doi: 10.1093/jamia/ocaa039] [Medline: 32232432]
16. Reeves J, Hollandsworth H, Torriani F, Taplitz R, Abeles S, Tai-Seale M, et al. Rapid response to COVID-19: health informatics support for outbreak management in an academic health system. J Am Med Inform Assoc 2020 Jun 01;27(6):853-859 [FREE Full text] [doi: 10.1093/jamia/ocaa037] [Medline: 32208481]
17. Bakken S. Can informatics innovation help mitigate clinician burnout? J Am Med Inform Assoc 2019 Feb 01;26(2):93-94 [FREE Full text] [doi: 10.1093/jamia/ocy186] [Medline: 30668747]
18. Gardner R, Cooper E, Haskell J, Harris D, Poplau S, Kroth P, et al. Physician stress and burnout: the impact of health information technology. J Am Med Inform Assoc 2019 Feb 01;26(2):106-114 [FREE Full text] [doi: 10.1093/jamia/ocy145] [Medline: 30517663]
19. Johns Hopkins Coronavirus Resource Center. URL: https://coronavirus.jhu.edu/ [accessed 2021-04-08]
20. Faiz S, Riahi T, Rahimzadeh P, Nikoubakht N. Commentary: remote electronic consultation for COVID-19 patients in teaching hospitals in Tehran, Iran. Med J Islam Repub Iran 2020;34:31 [FREE Full text] [doi: 10.34171/mjiri.34.31] [Medline: 32617270]
21. Schmoldt RA, Freeborn DK, Klevit HD. Physician burnout: recommendations for HMO managers. HMO Pract 1994 Jul;8(2):58-63. [Medline: 10135263]
22. Williams ES, Konrad TR, Linzer M, McMurray J, Pathman DE, Gerrity M, et al. Refining the measurement of physician job satisfaction: results from the Physician Worklife Survey. SGIM Career Satisfaction Study Group. Society of General Internal Medicine. Med Care 1999 Dec;37(11):1140-1154. [doi: 10.1097/00005650-199911000-00006] [Medline: 10549616]
23. Britt HR, Koranne R, Rockwood T. Statewide improvement approach to clinician burnout: findings from the baseline year. Burnout Res 2017 Dec;7:29-35. [doi: 10.1016/j.burn.2017.11.002]
24. Dolan ED, Mohr D, Lempa M, Joos S, Fihn SD, Nelson KM, et al. Using a single item to measure burnout in primary care staff: a psychometric evaluation. J Gen Intern Med 2015 May;30(5):582-587 [FREE Full text] [doi: 10.1007/s11606-014-3112-6] [Medline: 25451989]
25. Häyrinen K, Saranto K, Nykänen P. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. Int J Med Inform 2008 May;77(5):291-304. [doi: 10.1016/j.ijmedinf.2007.09.001] [Medline: 17951106]
26. Brooke J. SUS: a 'quick and dirty' usability scale. In: Jordan PW, Thomas B, McClelland IL, Weerdmeester B, editors. Usability Evaluation In Industry. Boca Raton, FL: CRC Press; 1996:189.
27. Brennan F. Technostress and leadership: a case study in higher education during the COVID-19 crisis. Theseus. 2021. URL: https://www.theseus.fi/bitstream/handle/10024/380031/Brennan_Fintan%20.pdf?sequence=3&isAllowed=y [accessed 2021-04-12]
28. Vaziri H, Casper WJ, Wayne JH, Matthews RA. Changes to the work-family interface during the COVID-19 pandemic: examining predictors and implications using latent transition analysis. J Appl Psychol 2020 Oct;105(10):1073-1087. [doi: 10.1037/apl0000819] [Medline: 32866024]
29. Christen P, D'Aeth JC, Løchen A, McCabe R, Rizmie D, Schmit N, et al. The J-IDEA Pandemic Planner: a framework for implementing hospital provision interventions during the COVID-19 pandemic. Med Care 2021 May 01;59(5):371-378 [FREE Full text] [doi: 10.1097/MLR.0000000000001502] [Medline: 33480661]
30. Mahmud E, Dauerman HL, Welt FG, Messenger JC, Rao SV, Grines C, et al. Management of acute myocardial infarction during the COVID-19 pandemic: a position statement from the Society for Cardiovascular Angiography and Interventions (SCAI), the American College of Cardiology (ACC), and the American College of Emergency Physicians (ACEP). J Am Coll Cardiol 2020 Sep 15;76(11):1375-1384 [FREE Full text] [doi: 10.1016/j.jacc.2020.04.039] [Medline: 32330544]
31. Responding to community spread of COVID-19: interim guidance, 7 March 2020. World Health Organization. URL: https://apps.who.int/iris/handle/10665/331421 [accessed 2021-02-10]
32. Schaufeli W, Leiter M, Maslach C, Jackson S. Maslach Burnout Inventory-General Survey. In: Maslach C, Jackson SE, Leiter MP, editors. The Maslach Burnout Inventory: Test Manual. Palo Alto, CA: Consulting Psychologists Press; 1996.
33. Physician burnout. Agency for Healthcare Research and Quality. URL: https://www.ahrq.gov/prevention/clinician/ahrq-works/burnout/index.html [accessed 2021-04-02]
34. Shanafelt T, Hasan O, Dyrbye L, Sinsky C, Satele D, Sloan J, et al. Changes in burnout and satisfaction with work-life balance in physicians and the general US working population between 2011 and 2014. Mayo Clin Proc 2015 Dec;90(12):1600-1613. [doi: 10.1016/j.mayocp.2015.08.023] [Medline: 26653297]
35. Nunnally JC. An overview of psychological measurement. In: Wolman BB, editor. Clinical Diagnosis of Mental Disorders: A Handbook. Boston, MA: Springer; 1978.
36. Babbott S, Manwell LB, Brown R, Montague E, Williams E, Schwartz M, et al. Electronic medical records and physician stress in primary care: results from the MEMO Study. J Am Med Inform Assoc 2014 Mar;21(e1):e100-e106 [FREE Full text] [doi: 10.1136/amiajnl-2013-001875] [Medline: 24005796]

37. DiAngi YT, Stevens LA, Halpern-Felsher B, Pageler NM, Lee TC. Electronic health record (EHR) training program identifies a new tool to quantify the EHR time burden and improves providers' perceived control over their workload in the EHR. JAMIA Open 2019 Jul;2(2):222-230 [FREE Full text] [doi: 10.1093/jamiaopen/ooz003] [Medline: 31984357]
38. Robinson K, Kersey JA. Novel electronic health record (EHR) education intervention in large healthcare organization improves quality, efficiency, time, and impact on burnout. Medicine (Baltimore) 2018 Oct;97(38):e12319. [doi: 10.1097/MD.0000000000012319] [Medline: 30235684]
39. Jordan J, Watcha D, Cassella C, Kaji AH, Trivedi S. Impact of a mentorship program on medical student burnout. AEM Educ Train 2019 Jul;3(3):218-225 [FREE Full text] [doi: 10.1002/aet2.10354] [Medline: 31360814]
40. Perumalswami CR, Takenoshita S, Tanabe A, Kanda R, Hiraike H, Okinaga H, et al. Workplace resources, mentorship, and burnout in early career physician-scientists: a cross sectional study in Japan. BMC Med Educ 2020 Jul 03;20(1):178 [FREE Full text] [doi: 10.1186/s12909-020-02072-x] [Medline: 32493497]
41. Kaelber DC, Bates DW. Health information exchange and patient safety. J Biomed Inform 2007 Dec;40(6 Suppl):S40-S45 [FREE Full text] [doi: 10.1016/j.jbi.2007.08.011] [Medline: 17950041]
42. Dewa CS, Loong D, Bonato S, Thanh NX, Jacobs P. How does burnout affect physician productivity? A systematic literature review. BMC Health Serv Res 2014 Jul 28;14:325 [FREE Full text] [doi: 10.1186/1472-6963-14-325] [Medline: 25066375]
43. Rothenberger DA. Physician burnout and well-being: a systematic review and framework for action. Dis Colon Rectum 2017 Jul;60(6):567-576. [doi: 10.1097/DCR.0000000000000844] [Medline: 28481850]
44. Yates SW. Physician stress and burnout. Am J Med 2020 Feb;133(2):160-164 [FREE Full text] [doi: 10.1016/j.amjmed.2019.08.034] [Medline: 31520624]
45. Ueda M, Martins R, Hendrie P, McDonnell T, Crews J, Wong T, et al. Managing cancer care during the COVID-19 pandemic: agility and collaboration toward a common goal. J Natl Compr Canc Netw 2020 Mar 20:1-4. [doi: 10.6004/jnccn.2020.7560] [Medline: 32197238]
46. Livingston E, Desai A, Berkwits M. Sourcing personal protective equipment during the COVID-19 pandemic. JAMA 2020 May 19;323(19):1912-1914. [doi: 10.1001/jama.2020.5317] [Medline: 32221579]
47. Hollander JE, Carr BG. Virtually perfect? Telemedicine for Covid-19. N Engl J Med 2020 May 30;382(18):1679-1681. [doi: 10.1056/NEJMp2003539] [Medline: 32160451]

Abbreviations

EHR: electronic health record
HIT: health information technology
IT: information technology

Edited by G Eysenbach; submitted 04.03.21; peer-reviewed by YC Kato-Lin, KM Kuo; comments to author 24.03.21; revised version received 12.04.21; accepted 28.04.21; published 03.06.21.

Please cite as:
Esmaeilzadeh P, Mirzaei T
Using Electronic Health Records to Mitigate Workplace Burnout Among Clinicians During the COVID-19 Pandemic: Field Study in Iran
JMIR Med Inform 2021;9(6):e28497
URL: https://medinform.jmir.org/2021/6/e28497
doi:10.2196/28497
PMID:34033578

Original Paper

Analysis of Mental Health Disease Trends Using BeGraph Software in Spanish Health Care Centers: Case Study

Susel Góngora Alonso[1], MSc; Andrés de Bustos Molina[2,3], PhD; Beatriz Sainz-De-Abajo[1], PhD; Manuel Franco-Martín[4], MD; Isabel De la Torre Díez[1], PhD

[1]Department of Signal Theory and Communications, and Telematics Engineering, University of Valladolid, Valladolid, Spain

[2]Next Limit Technologies, Madrid, Spain

[3]Department of Technology-CIEMAT, Madrid, Spain

[4]Psychiatry Department, Rio Hortega University Hospital and Zamora Hospital, Valladolid, Zamora, Spain

Corresponding Author:
Isabel De la Torre Díez, PhD
Department of Signal Theory and Communications, and Telematics Engineering
University of Valladolid
Paseo de Belén, 15
Valladolid, 47011
Spain
Phone: 34 983423000 ext 3703
Fax: 34 983423667
Email: isator@tel.uva.es

Abstract

Background: In the era of big data, networks are becoming a popular factor in the field of data analysis. Networks are part of the main structure of BeGraph software, which is a 3D visualization application dedicated to the analysis of complex networks.

Objective: The main objective of this research was to visually analyze tendencies of mental health diseases in a region of Spain, using the BeGraph software, in order to make the most appropriate health-related decisions in each case.

Methods: For the study, a database was used with 13,531 records of patients with mental health disorders in three acute medical units from different health care complexes in a region of Spain. For the analysis, BeGraph software was applied. It is a web-based 3D visualization tool that allows the exploration and analysis of data through complex networks.

Results: The results obtained with the BeGraph software allowed us to determine the main disease in each of the health care complexes evaluated. We noted 6.50% (463/7118) of admissions involving unspecified paranoid schizophrenia at the University Clinic of Valladolid, 9.62% (397/4128) of admissions involving chronic paranoid schizophrenia with acute exacerbation at the Zamora Hospital, and 8.84% (202/2285) of admissions involving dysthymic disorder at the Rio Hortega Hospital in Valladolid.

Conclusions: The data analysis allowed us to focus on the main diseases detected in the health care complexes evaluated in order to analyze the behavior of disorders and help in diagnosis and treatment.

(JMIR Med Inform 2021;9(6):e15527) doi:10.2196/15527

KEYWORDS

BeGraph software; diseases; health care complexes; mental health; visualization

Introduction

Currently, the medical care provided in a wide range of clinical applications makes use of the latest technologies, multimedia data, multidimensional medical images, videos, and data based on sensors and texts [1]. In the area of mental health, algorithms and data mining techniques are used to extract large volumes of predictive data, build models based on large-scale medical information, and predict diagnoses [2]. Despite its advantages in the medical field, these techniques and algorithms pose a computational challenge. Complex data management and software design are necessary [3,4]. Medical applications linked to the network sciences, which are responsible for collecting, managing, analyzing, interpreting, and presenting data, allow the development of visualization tools that support diagnosis and medical decision making [5,6].

Today, we have web-based 3D visualization tools that allow the exploration and analysis of data [7,8]. BeGraph is a computing application, in the cloud, dedicated to the visualization and analysis of complex networks for companies and the scientific field. Networks are becoming popular in the field of data analysis and big data because they contain information that other data structures lack [9]. Characteristics, such as the phenomenon of the small world, the diffusion of events, and the grouping in clusters based on communities, are exclusive of networks. Additionally, networks are the basic data structure in BeGraph. A network is a set of N entities called nodes related by a set of L links that represent binary relationships. Nodes and links can have associated properties, either imported from a data source or calculated with BeGraph [10].

The BeGraph software uses different metrics that consist of mathematical operations based on the graphics theory used to characterize the network. These metrics provide information about the centrality of the nodes, busiest links, cohesion properties, grouping of networks, and topology [11,12].

The purpose of this research was to analyze the trend of mental health illnesses in a region of Spain through BeGraph in order to make health-related decisions, thus helping medical personnel to focus their research on preventing and improving the life quality of patients most affected by the most common diseases.

There are similar studies that show the viability of our research. In a previous study [1], the authors presented a 3D medical graphical avatar (MGA). It is a web-based personal health record visualization system designed to explore web-based delivery of a wide array of medical data types including multidimensional medical images, medical videos, text-based data, and spatial annotations. In a previous report [7], the authors developed an interactive 3D tool to facilitate the visualization and exploration of covariate distributions and imbalances across evidence for network meta-analysis (NMA). In another report [13], researchers developed 3D molecular dynamics visualization (MDV) that offers an immersive viewing experience of biomolecular systems.

Below, we provide a description of BeGraph, the methodology used, the results obtained in the study, and the discussion and conclusions of the investigation.

Methods

Database

For the study, the data set used had a total of 13,531 records of patients with mental health disorders. The data were provided by the following three acute care units of the care centers: University Clinic of Valladolid (7118 records), Zamora Hospital (4128 records), and Rio Hortega Hospital (2285 records). These health care complexes are located in the region of Castilla y Leon, Spain. The records of the acute units were used because they contain the largest amount of data. Moreover, patients with mental disorders are admitted to acute units at moments of greatest severity of their pathologies; therefore, suitable management of the acute phase is essential for disease stabilization. All the data included in the study were anonymous and followed the Ninth Revision, International Classification of Diseases (ICD-9), which the World Health Organization established to track mortality statistics, and they corresponded to the period between 2005 and 2015. The database collects information such as dates of hospital admission and departure, days of stay, year of registration, health care complex, gender, psychiatric diseases, other diseases not belonging to the area of mental health, and therapies corresponding to each medical diagnosis. Among all the data collected in the database for the study, the variables included were (1) the diagnoses of the mental health area, (2) the gender of the admitted patients, (3) the year of admission, and (4) the health care complex to which the patients belong.

The irrelevant variables for the study were eliminated, and these were (1) therapies, (2) other medical diagnoses that do not belong to the mental health area, (3) dates of admission, and (4) discharge from the hospital. Null values, double blanks, and special characters were also eliminated to avoid false categories.

BeGraph Software

BeGraph is a cloud platform for the visualization and analysis of complex networks. It is efficient and highly scalable, being able to represent, in 3D, networks of several million links. For this, it uses the layout algorithms implemented. It allows exploration and interaction with the network in real time, allowing the analyst to become familiar with the data and obtain qualitative information quickly and easily. The 3D visualization of the network is highly configurable in order to simultaneously represent various characteristics of the nodes or links (depending on shape, color, or size). In this way, researchers can visually identify the clusters or communities and patterns of the network [10].

BeGraph allows the loading of properties of nodes and links, which can influence the results of the metrics typical of graph theory. There are about 20 different metrics that characterize the network following criteria such as topology, centrality, cohesion, and clustering or division into communities. These metrics are characteristics of graph theory and provide information that conventional statistics do not provide. In the following section, we describe the database, metrics, and algorithms used in this study.

BeGraph Algorithms

As the BeGraph software is a cloud application, interaction with the user is done through a web browser. Input data are uploaded to the cloud (.csv files or remote databases), and the metrics are configured and executed. Several complex calculations can be made in parallel.

All these calculations take place on cloud servers transparent to the user. Additionally, the cloud server hardware can be configured to adapt to the size of the network for analysis. On the user's local computer, only the calculations related to the rendering processes of visualization take place. Depending on the number of nodes and links, these computations can be very expensive. It is necessary to have a powerful graphics processing unit (GPU) to visualize and navigate through large networks.

Once the network is loaded on the platform, it is necessary to use a layout algorithm to represent and visualize the network. The layout algorithms in networks have the nodes in the plane or space for their visualization. In the arrangement of each node (embedding), an attempt is made to preserve the distance between nodes in the network in the Euclidean space. Thus, the nodes that are neighbors or are a few jumps between them will be found nearby in the visualization. The layout algorithms preserve the connectivity and structure of network clusters, making them obvious to the observer. Among the available algorithms, we used the so-called Force Atlas that considers the network as a system of N bodies (nodes) on which the below forces act [14].

Repulsive force: It is assumed that all nodes are spheres with a charge of the same sign that are repelled according to Coulomb law. Therefore, for any pair of nodes (u and v) we will have the following:

where k_u and k_v are the degrees of the nodes and α is a constant adjusted by the user.

Attraction force: We consider that the nodes connected by a link are attracted with a force proportional to the distance that separates them as if they were joined by a spring (Hooke law) as follows:

where β is manually adjusted.

Gravity force: To avoid the not connected components of the network being separated too much, it is possible to adjust a small central force (according to the parameter λ) that attracts the nodes to the origin of coordinates as follows:

where d_{0u}^2 is the distance from the node to the origin of coordinates.

The combination of these three forces (with various acceleration algorithms, such as the Barnes-Hut approach to avoid calculating O [N2] repulsive forces) spatially provides the nodes in an intelligent and illustrative manner, revealing topological characteristics of the network.

Various metrics (graph theory) were used in this study. The first is *degree centrality*. It is the most basic measure of centrality of the network and is simply the number of neighbors of a node. It represents locally the importance of a node in the network, taking as a criterion the number of connections it has. The second is *eigenvector centrality*. It does a ranking of nodes in a nonlocal way, taking into account the degree of its neighbors, the neighbors of its neighbors, and so on. It reveals nodes that, without having a high degree, are important in the network because they are connected or close to other high-grade nodes. The third is *louvain communities*. Communities in a graph or network are the equivalent of clusters in other data structures. They are sets of nodes connected more densely to each other than to the rest of the network. As the problem of partitioning a graph is NP completeness (nondeterministic polynomial-time complete), community detection algorithms are limited by the size of the network, or they produce approximate results. We used the well-known algorithm of Louvain [12] for its efficiency and the quality of its results.

Results

In this section, we present the results using a database that included a total of 13,531 anonymous medical records. We used BeGraph software, which allows the user to upload, store, and share several networks in the cloud. Each network can be viewed and explored using a web browser, giving a general idea of its structure and topological properties [10].

To obtain the results shown below, we used different metrics and one of the BeGraph design algorithms (Force Atlas Layout) [14]. It is an algorithm based on a system directed by force and optimized to deal with large networks, and it is highly configurable. *Degree centrality* is responsible for identifying the most popular nodes in the network. The size of the node and the warmth of the color represent the degree of each node. *Eigenvector centrality* is responsible for measuring the centrality of the selected node, normalizing the components of the eigenvector to a maximum value of 1.

With the 3D visualization tool, we explored the data and represented the results. We seek to clarify aspects such as (1) networks of diseases prevalent in the 2005-2015 period of patient admission records and (2) networks of prevalent diseases depending on the patient's gender, with consideration of each of the health care complexes included in the study.

In Figure 1, the prevalent diseases of the University Clinic of Valladolid are shown during the period from 2005 to 2015.

Figure 1. Network with prevalent diseases-years in the University Clinic of Valladolid.

The nodes are identified with a property called NodeLabel, the ubiquitous node property used to identify the nodes in metrics and filters. The largest nodes represent the year and the smallest the diseases detected in that hospital. The links represent the binary relationships between them. In Figures 1, 2, and 3, it can be seen that there are diseases that were only registered in a given year and not present in the rest of the years. The nodes of the diseases that are interrelated with more than 1 year represent the same type of disease registered in several years.

In Figures 2 and 3, the networks with prevalent diseases in the Health Care Complex of Zamora and the Rio Hortega Hospital, respectively, are shown. The period analyzed is from 2005 to 2015 and visually reflects the relationship of the data, providing medical personnel with criteria for decision making.

Figure 2. Network with prevalent diseases-years in the Assistance Complex of Zamora.

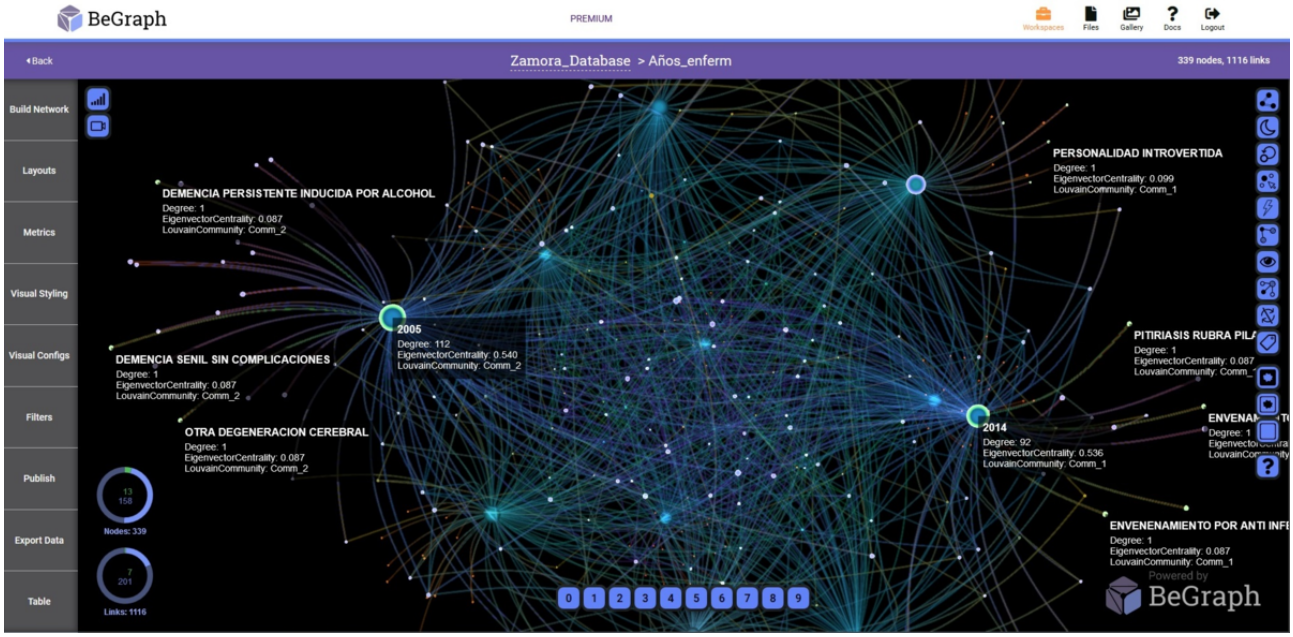
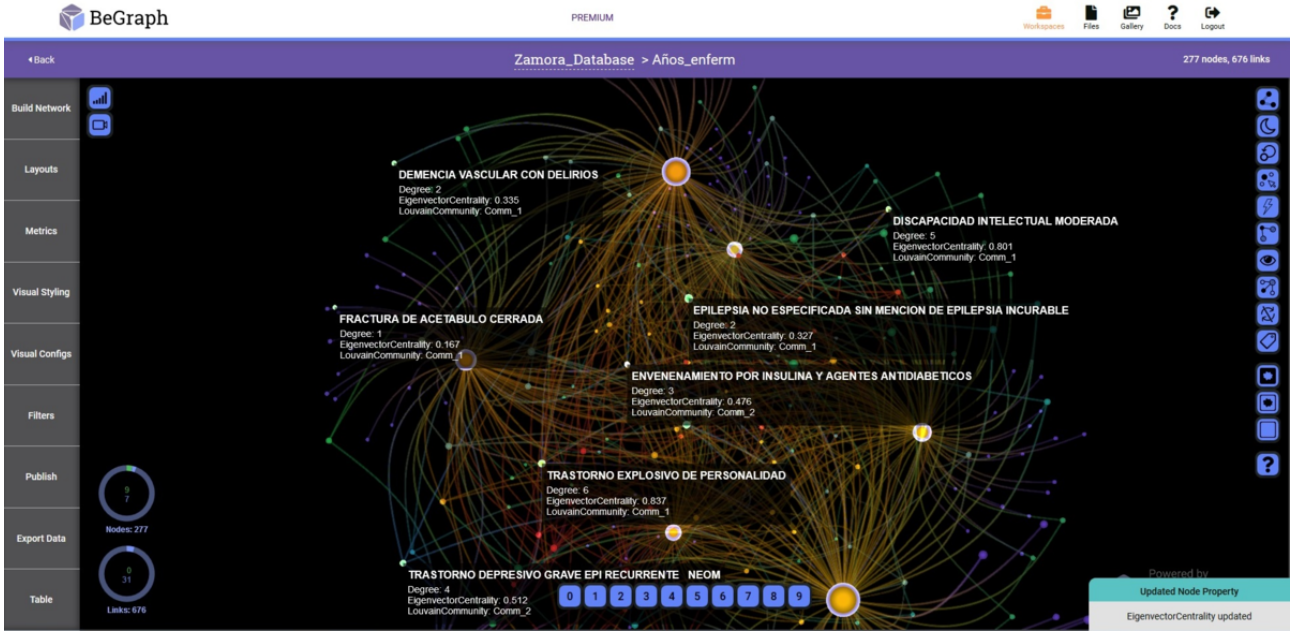


Figure 3. Network with prevalent diseases-years in the Rio Hortega Hospital of Valladolid.



The results of the main diseases presented by patient admission records with mental health diseases in the different health care complexes are shown in Table 1.

Table 1. Admission records according to mental health diseases for health care complexes in Valladolid and Zamora.

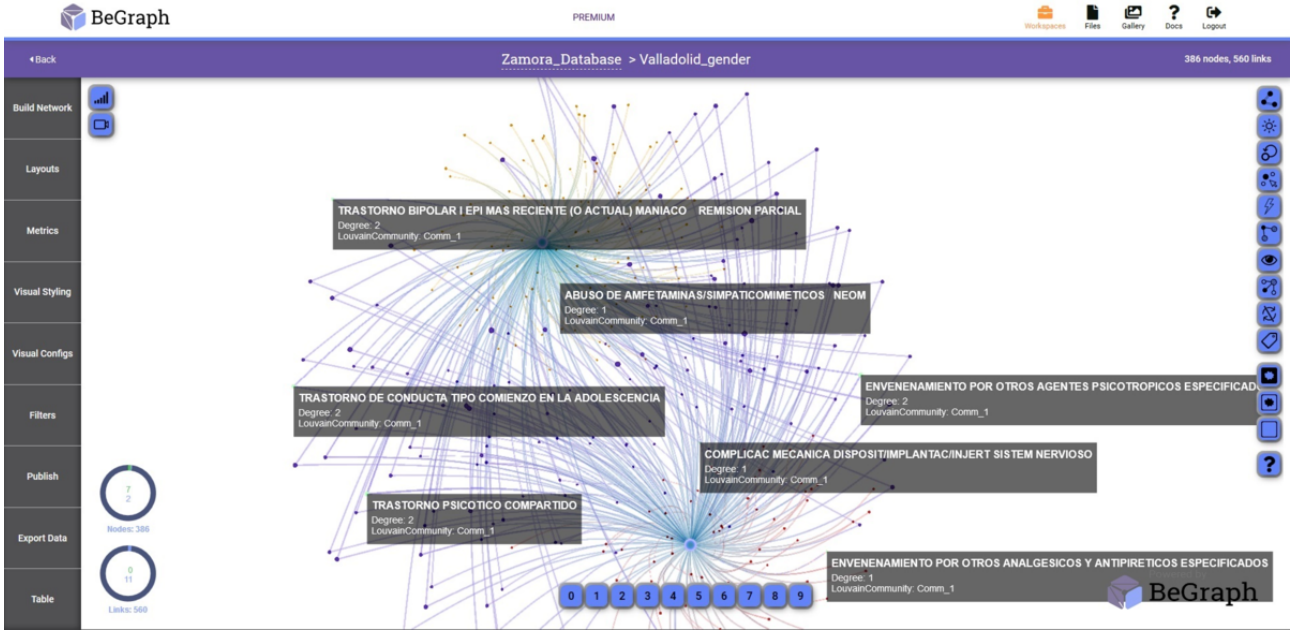
Main mental health diseases	Admission records, n (%)		
	University Clinic of Valladolid (N=7118)	Zamora Hospital (N=4128)	Rio Hortega Hospital of Valladolid (N=2285)
Unspecified paranoid schizophrenia	463 (6.50%)	10 (0.24%)	15 (0.66%)
Adaptation disorder with mixed disturbance of emotions and behavior	457 (6.42%)	7 (0.17%)	18 (0.79%)
Another alcohol dependence and neom	372 (5.22%)	N/A[a]	N/A
Mixed adaptation disorder of anxiety and depressed humor	356 (5.00%)	95 (2.30%)	36 (1.58%)
Unspecified psychosis	352 (4.94%)	45 (1.09%)	87 (3.81%)
Dysthymic disorder	289 (4.06%)	161 (3.90%)	202 (8.84%)
Chronic paranoid schizophrenia with acute exacerbation	3 (0.04%)	397 (9.62%)	108 (4.73%)
Chronic schizoaffective disorder/acute exacerbation	N/A	209 (5.06%)	74 (3.24%)
Another alcohol dependence and continuous neom	13 (0.18%)	173 (4.19%)	183 (8.01%)
Poisoning by tranquilizers based on benzodiazepine	113 (1.59%)	81 (1.96%)	110 (4.81%)
Delirious disorder	271 (3.81%)	114 (2.76%)	91 (3.98%)

[a]N/A: not applicable.

The percentages are obtained from the total records. The University Clinic of Valladolid had 7118 records, Zamora Hospital had 4128 records, and Rio Hortega Hospital had 2285 records.

Another parameter analyzed in our study was prevalent diseases depending on the patient's gender, as shown in Figures 4, 5, and 6.

Figure 4. Network with prevalent diseases-gender in the University Clinic of Valladolid.



In Figures 4 and 5, the main nodes represent the gender of the admission records. The upper node corresponds to men and the lower node corresponds to women. The small nodes represent the diseases detected in the University Clinic of Valladolid and Zamora Hospital, respectively. In the Rio Hortega Hospital, the upper node corresponds to women and the lower node corresponds to men (Figure 6). The diseases are visually shown only in men and women, and the common ones among them are displayed.

Tables 2 and 3 present a summary of the main diseases detected in each of the health care complexes, with consideration of gender.

Figure 5. Network with prevalent diseases-gender in the Healthcare Complex of Zamora.

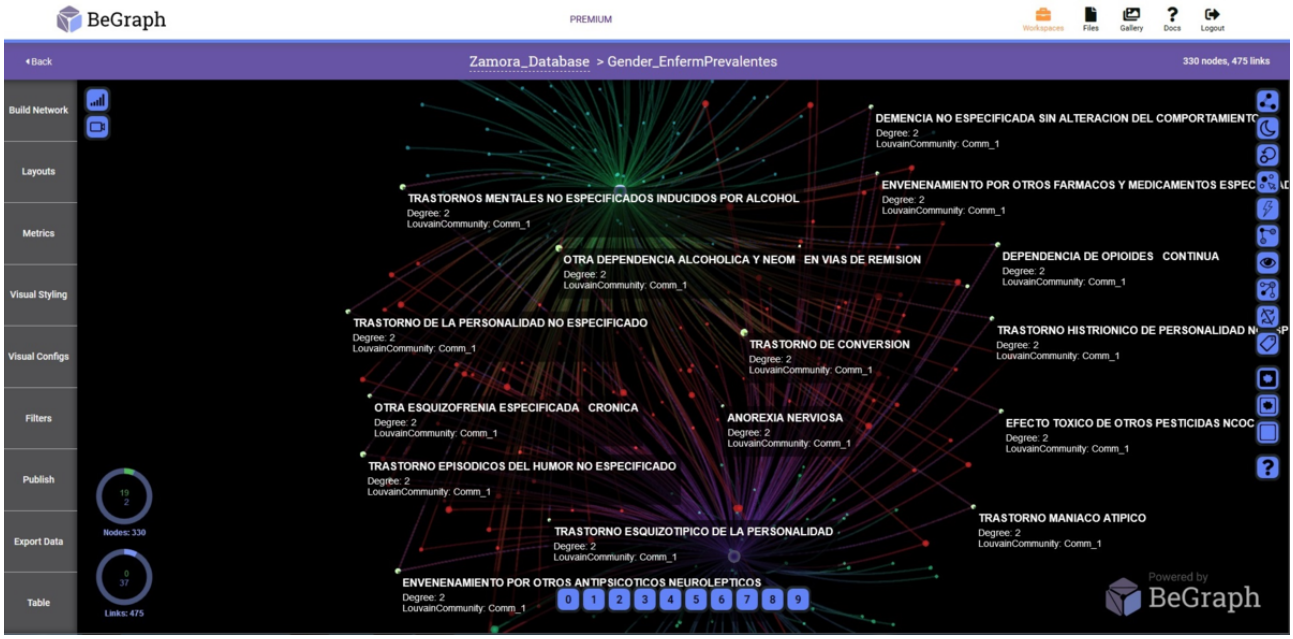


Figure 6. Network with prevalent diseases-gender in the Rio Hortega Hospital of Valladolid.

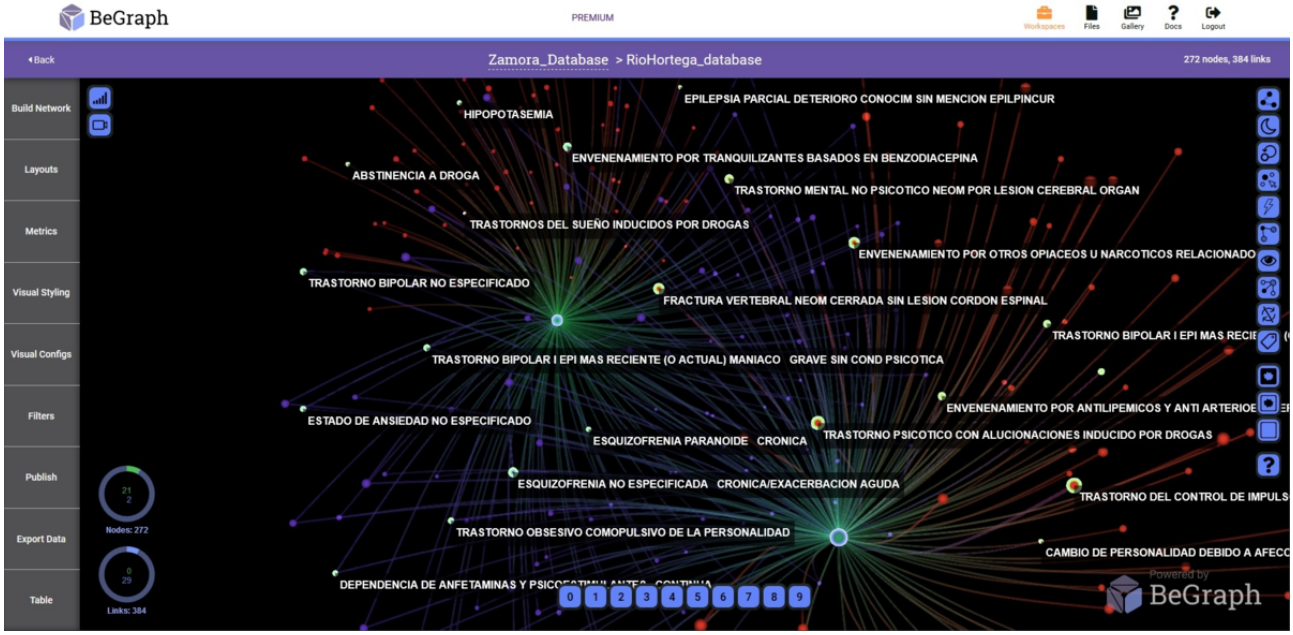


Table 2. Admission records according to mental health diseases in males for health care complexes in Valladolid and Zamora.

Main mental health diseases in males	Admission records, n (%)		
	University Clinic of Valladolid (N=3558)	Zamora Hospital (N=2282)	Rio Hortega Hospital of Valladolid (N=1150)
Unspecified paranoid schizophrenia	337 (9.47%)	6 (0.26%)	14 (1.22%)
Another alcohol dependence and neom	273 (7.67%)	N/A[a]	1 (0.09%)
Adaptation disorder with mixed disturbance of emotions and behavior	200 (5.62%)	N/A	10 (0.87%)
Chronic paranoid schizophrenia with acute exacerbation	3 (0.08%)	307 (13.45%)	74 (6.43%)
Another alcohol dependence and continuous neom	7 (0.20%)	139 (6.09%)	136 (11.83%)
Chronic schizoaffective disorder/acute exacerbation	N/A	82 (3.59%)	34 (2.96%)
Dysthymic disorder	64 (1.80%)	31 (1.36%)	55 (4.78%)

[a]N/A: not applicable.

Table 3. Admission records according to mental health diseases in females for health care complexes in Valladolid and Zamora.

Main mental health diseases in females	Admission records, n (%)		
	University Clinic of Valladolid (N=3560)	Zamora Hospital (N=1846)	Rio Hortega Hospital of Valladolid (N=1135)
Adaptation disorder with mixed disturbance of emotions and behavior	257 (7.22%)	7 (0.38%)	8 (0.70%)
Dysthymic disorder	225 (6.32%)	130 (7.04%)	147 (12.95%)
Mixed adaptation disorder of anxiety and depressed humor	202 (5.67%)	48 (2.60%)	20 (1.76%)
Chronic schizoaffective disorder/acute exacerbation	N/A[a]	127 (6.88%)	40 (3.52%)
Chronic paranoid schizophrenia with acute exacerbation	N/A	90 (4.88%)	34 (3.00%)
Poisoning by tranquilizers based on benzodiazepine	73 (2.05%)	57 (3.09%)	64 (5.64%)
Disorder or unspecified dissociative reaction	35 (0.98%)	16 (0.87%)	49 (4.32%)

[a]N/A: not applicable.

The results presented in Tables 2 and Table 3 show considerable differences in the types of disorders. Men have disorders related to schizophrenia and alcohol, while women have dysthymic and adaptive disorders.

To obtain the different networks shown in Figures 1-6, we used the layout algorithm called Force Atlas, which considers the network as a system of N nodes on which different forces act. In the algorithm configuration, we used 500 iterations, gravity of 1, jitter tolerance of 0.5, and mass range of 10, while in the configuration of the metrics, the *eigenvector centrality* used 100 iterations. These parameters allowed us to obtain the best visualizations of the networks formed by the data.

Discussion

In this study, we performed a visual analysis of mental health diseases prevalent in a region of Spain, using the BeGraph software, in order to make health-related decisions. The results show that BeGraph is a tool that allows visualization, in 3D, of the behavior of a network, offering complementary information to classical statistical analysis. As a novelty, we presented the analysis and visualization of medical data from the point of view of the theory of complex networks. Metrics based on networks or graphs provide new properties, which can be used to describe the data quantitatively or as additional properties in other machine learning algorithms. The visualization of the network provides an interactive method of scanning medical data as well as qualitative characterization quickly and easily.

In this work, we used a database of 13,531 medical records. The analysis allowed us to determine the main disease detected in each of the health care complexes. We noted 6.50% (463/7118) of admissions involving unspecified paranoid schizophrenia at the University Clinic of Valladolid, 9.62% (397/4128) of admissions involving chronic paranoid schizophrenia with acute exacerbation at the Zamora Hospital, and 8.84% (202/2285) of admissions involving dysthymic disorder at the Rio Hortega Hospital. In the health care complexes with the most admissions, such as the University Clinic of Valladolid and Zamora Hospital, the greatest mental health disorder was presented by patients with different types of schizophrenia. These data allowed us to focus on this disease to analyze its behavior and help in diagnosis and treatment. In Rio Hortega Hospital, dysthymic disorder represented the main disease. It was present in other health care complexes with similar amounts of admissions, but represented a lower value with regard to the main diseases detected in the University Clinic of Valladolid.

The analysis of the main diseases depended on gender. The results of the disorders were different in men and women. In the case of men, in the three health care complexes, the highest number of admissions detected was related to schizophrenia disorders and alcohol dependence. In women, the disorders were based on dysthymic disorders, adaptive disorders, schizoaffective disorders, and poisoning. Therefore, the results indicated that it is necessary to take gender into account in order to make decisions in diagnoses. The trends shown by these results provide a basis for future research regarding the prediction of mental health diseases and common patterns among patients with the same diseases. In this case, we focused on schizophrenia disorders taking into account gender, stay days, health care complex, age, and other diagnosed diseases. We propose the prediction of patient readmissions with schizophrenia using machine learning techniques.

Acknowledgments

We thank the Service of Psiquiatry of the Provincial Hospital of Zamora, Spain, for collaboration in this work. This research has been funded and supported by the Health Regional Service (GRS 1801/A/18).

Conflicts of Interest

None declared.

References

1. de Ridder M, Constantinescu L, Bi L, Jung Y, Kumar A, Kim J, et al. A web-based medical multimedia visualisation interface for personal health records. In: Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems. 2013 Presented at: 26th IEEE International Symposium on Computer-Based Medical Systems; June 20-22, 2013; Porto, Portugal p. 191-196. [doi: 10.1109/CBMS.2013.6627787]
2. Alonso SG, de la Torre-Díez I, Hamrioui S, López-Coronado M, Barreno DC, Nozaleda LM, et al. Data Mining Algorithms and Techniques in Mental Health: A Systematic Review. J Med Syst 2018 Jul 21;42(9):161. [doi: 10.1007/s10916-018-1018-2] [Medline: 30030644]
3. Yoo TS, Bliss D, Lowekamp BC, Chen DT, Murphy GE, Narayan K, et al. Visualizing Cells and Humans in 3D: Biomedical Image Analysis at Nanometer and Meter Scales. IEEE Comput. Grap. Appl 2012 Sep;32(5):39-49. [doi: 10.1109/mcg.2012.68]
4. Qi X, Mei G, Cuomo S, Xiao L. A network-based method with privacy-preserving for identifying influential providers in large healthcare service systems. Future Gener Comput Syst 2020 Aug;109:293-305 [FREE Full text] [doi: 10.1016/j.future.2020.04.004] [Medline: 32296253]
5. Chandra S, Dowling J, Engstrom C, Xia Y, Paproki A, Neubert A, et al. A lightweight rapid application development framework for biomedical image analysis. Comput Methods Programs Biomed 2018 Oct;164:193-205 [FREE Full text] [doi: 10.1016/j.cmpb.2018.07.011] [Medline: 30195427]

6. Mongeau R, Casu M, Pani L, Pillolla G, Lianas L, Giachetti A. Building a virtual archive using brain architecture and Web 3D to deliver neuropsychopharmacology content over the Internet. Comput Methods Programs Biomed 2008 May;90(2):124-136. [doi: 10.1016/j.cmpb.2007.12.007] [Medline: 18262677]
7. Batson S, Score R, Sutton AJ. Three-dimensional evidence network plot system: covariate imbalances and effects in network meta-analysis explored using a new software tool. J Clin Epidemiol 2017 Jun;86:182-195. [doi: 10.1016/j.jclinepi.2017.03.008] [Medline: 28344122]
8. Luković V, Ćuković S, Milošević D, Devedžić G. An ontology-based module of the information system ScolioMedIS for 3D digital diagnosis of adolescent scoliosis. Comput Methods Programs Biomed 2019 Sep;178:247-263 [FREE Full text] [doi: 10.1016/j.cmpb.2019.06.027] [Medline: 31416553]
9. Niyirora J, Aragones O. Network analysis of medical care services. Health Informatics J 2020 Sep 18;26(3):1631-1658 [FREE Full text] [doi: 10.1177/1460458219887047] [Medline: 31735109]
10. BeGraph. URL: https://begraph.net/ [accessed 2020-05-21]
11. Newman M. Networks: An Introduction. Oxford, England: Oxford University Press; 2010.
12. Blondel VD, Guillaume J, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J. Stat. Mech 2008 Oct 09;2008(10):P10008. [doi: 10.1088/1742-5468/2008/10/p10008]
13. Wiebrands M, Malajczuk C, Woods A, Rohl A, Mancera R. Molecular Dynamics Visualization (MDV): Stereoscopic 3D Display of Biomolecular Structure and Interactions Using the Unity Game Engine. J Integr Bioinform 2018 Jun 21;15(2):8 [FREE Full text] [doi: 10.1515/jib-2018-0010] [Medline: 29927749]
14. Jacomy M, Venturini T, Heymann S, Bastian M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. PLoS One 2014 Jun 10;9(6):e98679 [FREE Full text] [doi: 10.1371/journal.pone.0098679] [Medline: 24914678]

Edited by G Eysenbach, Q Zeng; submitted 17.07.19; peer-reviewed by M Sharma, A Khanna, M Syafrudin; comments to author 07.09.20; revised version received 17.09.20; accepted 12.04.21; published 16.06.21.

Please cite as:
Góngora Alonso S, de Bustos Molina A, Sainz-De-Abajo B, Franco-Martín M, De la Torre Díez I
Analysis of Mental Health Disease Trends Using BeGraph Software in Spanish Health Care Centers: Case Study
JMIR Med Inform 2021;9(6):e15527
URL: https://medinform.jmir.org/2021/6/e15527
doi:10.2196/15527
PMID:34132650

Original Paper

Smart Decentralization of Personal Health Records with Physician Apps and Helper Agents on Blockchain: Platform Design and Implementation Study

Hyeong-Joon Kim[1], MSc; Hye Hyeon Kim[1], PhD; Hosuk Ku[2], MD, MSc; Kyung Don Yoo[3], MD, PhD; Suehyun Lee[4], PhD; Ji In Park[5,6], MD, PhD; Hyo Jin Kim[7], MD, PhD; Kyeongmin Kim[8,9], MD, PhD; Moon Kyung Chung[1], BA; Kye Hwa Lee[1,10,11], MD, PhD; Ju Han Kim[1,2], MD, PhD

[1]Division of Biomedical Informatics, College of Medicine, Seoul National University, Seoul, Republic of Korea
[2]Division of Nephrology, Department of Internal Medicine, Inje University Seoul Paik Hospital, Seoul, Republic of Korea
[3]Division of Nephrology, Department of Internal Medicine, Ulsan University Hospital, Ulsan, Republic of Korea
[4]Department of Biomedical Informatics, College of Medicine, Konyang University, Daejeon, Republic of Korea
[5]Department of Internal Medicine, Kangwon National University Hospital, Chuncheon, Republic of Korea
[6]School of Medicine, Kangwon National University, Chuncheon, Republic of Korea
[7]Department of Internal Medicine and Biomedical Research Institute, Pusan National University Hospital, Busan, Republic of Korea
[8]Department of Internal Medicine, Eulji University Hospital, Daejeon, Republic of Korea
[9]College of Medicine, Eulji University, Daejeon, Republic of Korea
[10]Department of Information Medicine, Asan Medical Center, Seoul, Republic of Korea
[11]College of Medicine, University of Ulsan, Seoul, Republic of Korea

Corresponding Author:
Ju Han Kim, MD, PhD
Division of Biomedical Informatics
College of Medicine
Seoul National University
103, Daehak-ro
Jongno-gu
Seoul, 03080
Republic of Korea
Phone: 82 27408320
Fax: 82 27478928
Email: juhan@snu.ac.kr

Abstract

Background: The Health Avatar Platform provides a mobile health environment with interconnected patient Avatars, physician apps, and intelligent agents (termed *IoA*^{3}) for data privacy and participatory medicine; however, its fully decentralized architecture has come at the expense of decentralized data management and data provenance.

Objective: The introduction of blockchain and smart contract technologies to the legacy Health Avatar Platform with a clinical metadata registry remarkably strengthens decentralized health data integrity and immutable transaction traceability at the corresponding data-element level in a privacy-preserving fashion. A crypto-economy ecosystem was built to facilitate secure and traceable exchanges of sensitive health data.

Methods: The Health Avatar Platform decentralizes patient data in appropriate locations (ie, on patients' smartphones and on physicians' smart devices). We implemented an Ethereum-based hash chain for all transactions and smart contract–based processes to guarantee decentralized data integrity and to generate block data containing transaction metadata on-chain. Parameters of all types of data communications were enumerated and incorporated into 3 smart contracts, in this case, a health data transaction manager, a transaction status manager, and an application programming interface transaction manager. The actual decentralized health data are managed in an off-chain manner on appropriate smart devices and authenticated by hashed metadata on-chain.

Results: Metadata of each data transaction are captured in a Health Avatar Platform blockchain node by the smart contracts. We provide workflow diagrams each of the 3 use cases of data push (from a physician app or an intelligent agents to a patient

Avatar), data pull (request to a patient Avatar by other entities), and data backup transactions. Each transaction can be finely managed at the corresponding data-element level rather than at the resource or document levels. Hash-chained metadata support data element–level verification of data integrity in subsequent transactions. Smart contracts can incentivize transactions for data sharing and intelligent digital health care services.

Conclusions: Health Avatar Platform and interconnected patient Avatars, physician apps, and intelligent agents provide a decentralized blockchain ecosystem for health data that enables trusted and finely tuned data sharing and facilitates health value-creating transactions with smart contracts.

(JMIR Med Inform 2021;9(6):e26230) doi:10.2196/26230

KEYWORDS

personal health records; blockchain; mobile health; semantic interoperatbility; decentralized system; patient-centered system

Introduction

Personal health records are an electronic health information resource derived from multiple data sources that are integrated, managed, and controlled by individuals [1,2]. Through the use of a personal health record, a person can not only gain more knowledge about their current health conditions but can also receive assistance when seeking an appropriate treatment plan [3,4]. To maximize the utilization of personal health record, it is necessary to develop a system that allows patients to access, manage, and exchange their health information easily and safely and apply appropriate standards [5,6]. Recently, with the progress of Internet of Things (IoT) technology, patients can obtain various types of health-related information through a range of mobile devices or sensors [7]. The integrity of a personal health record demands both syntactic and semantic interoperability and the patient-centered health-data integration of various sources, including institutional electronic health records, IoT-enabled life logs, patient-reported outcome measures, and personal genomic data.

There is a need for an electronic environment for patient personal health records to interact with both physician apps and third-party artificial intelligence service agents. The Health Avatar Platform (HAP) began as decentralized health data management platform supporting patient-centered health data integration on a mobile smartphone app (a patient Avatar). HAP allows patients to store and manage their health data received from various health care institutions with syntactic and semantic interoperability. Once authorized and registered, third-party agents or distributed artificial intelligence services can access patient-centric health data on patient Avatars through HAP RESTful (representational state transfer) application programming interfaces (APIs) [8,9]. Electronic health record data can be pushed to an institutional gateway server (XNetHub) and then pulled by physician apps (XNet). Syntactic interoperability is supported by standard messaging protocols such as HL7 FHIR (Fast Healthcare Interoperability Resources), HL7 Continuity of Care Document, and ASTM Continuity of Care Record protocols. Semantic interoperability of electronic health record data from different hospitals is secured by predefined, preregistered, and postexpandable common data elements that full comply with ISO/IEC 11179 metadata registry international standards [8].

HAP enables peer-to-peer bidirectional communications among patient Avatars, third-party intelligent agents, and physician apps (termed *IoA*^{3}). DialysisNet (a physician app) and Avatar Beans (a patient Avatar) were the first and most successful apps for chronic kidney disease and hemodialysis patient management. The Avatar Beans app is downloadable from Google Play [9] and Apple App Store [10]. As of December 1, 2020, these apps were used by 22 nephrologists connecting 14 teaching hospitals using different electronic health records and 245 patients in South Korea. Kim et al [11] successfully conducted a multicenter cohort study for evaluating the treatment patterns of renal anemia of patients undergoing hemodialysis using DialysisNet. DialysisNet demonstrated seamless connectivity and semantic interoperability of standardized electronic clinical data capture among heterogeneous electronic health record systems. RehabilitationNet and Avatar Fit were launched (in 2020) as a second wave for an industrial-accident hospital network.

As public ledger technology [12,13], blockchain records transaction data between participants on a network in tamper-resistant storage [14]. Smart contracts can be developed and deployed by Ethereum [15,16], allowing contracts or interactions between participants in the blockchain network to be represented in Turing-complete language and automatically executed [17]. Many medical institutions and health care vendors have been conducting research on blockchain methods to build a system that is more transparent, traceable, verifiable, and irreversible with transactions occurring in conventional information systems [18-24]. The main objective of most of these studies was to share and exchange medical information mainly generated by providers securely. Others have applied a blockchain network for personal health record management [23,25-31]; these studies focused on a mobile health platform for patients to manage patient-centered health information.

Serving solely as an intermediary, HAP does not store any health data but securely relays authenticated and authorized data transmissions in a fully decentralized fashion among mobile devices and servers of interconnected patient Avatars, physician apps, and intelligent agents. In other words, even before the introduction of blockchain, HAP has already been a fully decentralized blockchain-friendly electronic or personal health record management platform. HAP is not vendor- or provider-centric but patient-centric. Because HAP is a mobile device–based health data integration or exchange platform for patients (ie, Avatars) and physicians (ie, apps) with no central storage, there are no privacy risks (eg, unauthorized access) as there are in centralized management systems [32-34]. HAP

supports iPads for physician apps for security reasons and both iOS and Android smartphones for patient Avatars for broader acceptance. Therefore, the introduction of blockchain technology to HAP and interconnected patient Avatars, physician apps, and intelligent agents systems may be a natural evolutionary path for better decentralization of digital health care. It resolves the old problems of (1) managing redundancy and integrity in decentralized health data among interconnected patient Avatars, physician apps, and intelligent agents devices; (2) verifying data authenticity and provenance; and (3) protecting data security from interference, forgery, and tampering by means of immutability. Blockchain technology guarantees better trust with regard to these functionalities [14]. Furthermore, the introduction of smart contract technology enables (4) smart access control down to each data element level from the current document or resource level, (5) smart data sharing at each data element level, and (6) a crypto-economy ecosystem for correctly incentivizing health care behaviors while also ensuring data privacy protection.

This paper describes (1) HAP and interconnected patient Avatars, physician apps, and intelligent agents system architecture for decentralized health data management by means of hash chain, RESTful API, and smart contract–based processes of authorized (2) data pushes to patient Avatars and to physician apps by each other, (3) data pulls from Avatars and apps upon a request from an intelligent agent for the purpose of decision support, and (4) data backup into a secure backup storage. A physician can prescribe a scheduled questionnaire to a patient and collect patient-reported outcome measures by combining these processes. Moreover, while standard messaging protocols such as HL7 FHIR, and HL7 Continuity of Care Document or ASTM Continuity of Care Record allow resource-level bulk queries, each common data element–level detailed query for data push and pull instances is supported by HAP interconnected patient Avatars, physician apps, and intelligent agents implementations by means of smart contracts. Each step in the processes can be systematically incentivized by the crypto-economy to facilitate data transactions and healthy behaviors in the HAP interconnected patient Avatars, physician apps, and intelligent agents ecosystem.

Methods

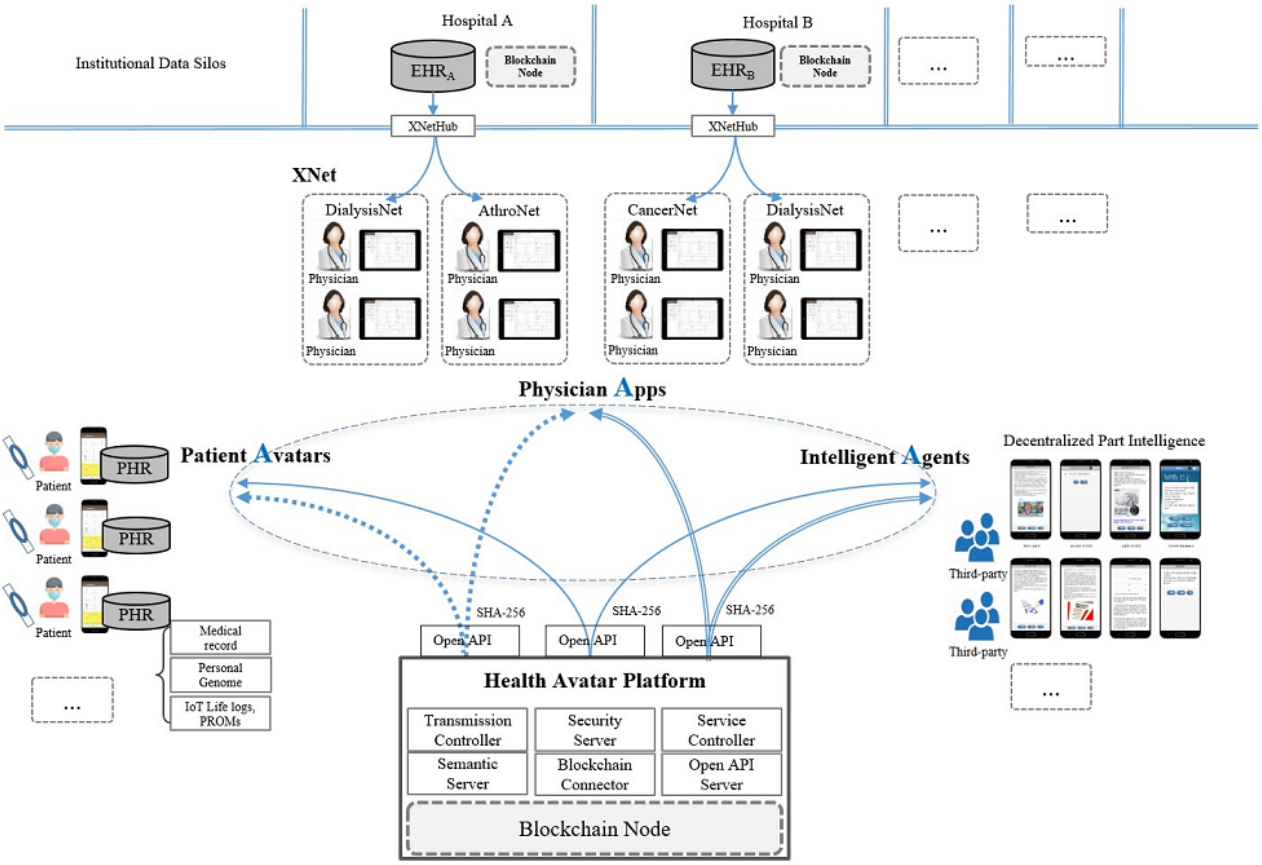
Health Avatar Platform Architecture and Data Communication

Avatars, apps, and agents of interconnected patient Avatars, physician apps, and intelligent agents represent patients, physicians, and third-party digital health care service providers, respectively. HAP has no central health data storage and performs decentralized data management (Figure 1). Patient data from many electronic health record systems are extracted, transformed via metadata validations, loaded onto a gateway XNetHub, and then synchronized with physician apps. Relevant personal health data are pushed from physician apps to patient Avatars, permitting an institutional data policy to be implemented. Smartphone-enabled patient Avatars are at the center of patient-centered data integration by collecting and integrating fragmented health data from multiple health care institutions. Patients can also generate patient-reported outcome measures that can be stored and sent to appropriate physician apps. Intelligent agents can request that physician apps or patient Avatars transmit health data via HAP RESTful APIs to provide digital health care services that are certified by and registered to the platform. Agents' recommendations and analysis results can also be sent to Avatars and XNets. The HAP server authenticates and authorizes instances of data access and transmission by Avatars, apps, and agents (or interconnected patient Avatars, physician apps, and intelligent agents) via smart contracts.

Fully decentralized health data management enabling strong data privacy is the hallmark of the HAP and interconnected patient avatar, physician app, and intelligent agent system. Each data element resides in its proper location (ie, a patient's data in the corresponding patient Avatar, a physician's data in the physician app, and a service provider's data in a third-party agent). Data redundancy is inevitable when a patient sends a copy of their patient-reported outcome measure to a physician and when a physician sends a copy of electronic health record data such as laboratory results and medications, to a patient. Intelligent agents can also receive health data and send (expert system) recommendations for clinical decision support to physicians as well as directly to patients. Previously, provenances of redundant decentralized health data were managed by a legacy HAP system. Introducing a hash chain for each data transaction among interconnected patient Avatar, physician app, and intelligent agent entities ensures better data provenance.

HAP provides a mobile platform for highly interconnected personal health records connecting many health care institutions, patients, and decentralized artificial intelligence agents. Semantic interoperability during data exchanges is achieved by fully curated and registered clinical common data elements supporting the ISO/IEC 11179 Metadata Registry standard. Electronic health record data are automatically transformed into common data elements at the time of extraction, transformation, and loading into the XNetHub metadata registry server. The metadata registry documents the standardization and registration of metadata to make the data understandable and shareable. HL7 FHIR, HL7 Continuity of Care Document, and ASTM Continuity of Care Record standards are supported (Multimedia Appendix 1) and semantically enriched at each common data element level by the metadata registry server for each of the specific apps. HAP has been used by 2 real-world practices in Korea—DialysisNet for chronic kidney disease and RehabilitationNet for neuromusculoskeletal disability management.

Figure 1. *IoA*^{3} (internet of Avatars, apps, and agents) system architecture. All data communications between entities are authenticated and hash-audited by the Health Avatar Platform to ensure data provenance. Blockchain nodes are distributed among the participating hospitals. API: application programming interface; EHR: electronic health record; PHR: personal health record; PROM: patient-reported outcome measure.



Application of a Blockchain Network

Though HAP has already been used for decentralized health data management in clinical practices, it is challenging for the legacy HAP system to verify whether or not the data on a terminal device such as a patient's smartphone have been compromised. We implemented an Ethereum-based hash chain as a tamper-proof and traceable modular storage approach to guarantee data integrity among terminal devices by storing a hash for each data transmission and applying them to verify data authenticity between originals and the copies of transmitted data. Thus, all data that have ever been transmitted through HAP can be correctly verified by HAP hash audits without risk to privacy (such as those that arise from capturing sensitive health data in a central storage). A patient's own patient-reported outcome measures from wearable devices or self-reporting forms can be verified for data provenance when patients send these records to themselves for digital signing or to another entity through hash auditing.

Health data are stored and managed off-chain in a decentralized fashion. The platform serves as a relay server that only stores the hash values on-chain of all data transactions for verification, data provenance, and auditing for tamper-proof data privacy. Two modules, called Blockchain Monitor and Node Manager, were newly added to the legacy HAP for creating block data in Ethereum (Multimedia Appendix 2). Introducing a blockchain technology to a legacy health care information system requires careful consideration, even with a popular and technically mature solution [35]. While Bitcoin is considered to be the most secure and representative platform, Ethereum is one of the most popular and robust platforms (1) allowing smart contract to be executed on-chain, (2) providing both permissioned and permission-less blockchain network, and (3) supporting a protocol-based crypto-economy environment, which is essential in incentivizing highly regulated but heterogeneous interactions [36,37]. The Ethereum network's proof-of-authority consensus algorithm was used [38]. Unlike a permission-less blockchain, the proof-of-authority algorithm can manage participants in the blockchain network. This algorithm also allows participants or initial nodes in the blockchain network to act as block generators for nodes that are new in the network. Because health-related information can be categorized as sensitive personal information, unauthorized access by other users should be prevented. If the permission-less blockchain approach is applied to a mobile health system, patient information is bound to be passed on to unreliable anonymous nodes. The application of permissioned blockchain is more appropriate to curb the potential for the occurrence one such breach. In this study, this permissioned blockchain was designed to allow only preconsulted care providers or health organizations to participate as nodes. Additionally, proof-of-authority Ethereum can create blocks more rapidly than permission-less blockchain networks method such as proof of work [12,39].

Off-Chain Data Management

To capture transaction hash logs in the blockchain, smart contracts that can be executed on an Ethereum virtual machine are required. Table 1 summarizes the parameters that can be extracted from data communications for each step. We designed smart contracts based on these investigated parameters. By designing these contracts, we intended to manage transaction metadata on the blockchain for data provenance while also managing patients' health data off-chain safely in their proper locations(ie, patient smartphones, physicians' smart pads, and agents' servers) for strong privacy protection. In addition, it is necessary to design a process that executes additionally implemented contracts in the legacy data communication process and delivers the required parameters; therefore, we compared different blockchain architectures (Multimedia Appendix 3).

We implemented Go-Ethereum blockchain with the smart contracts (Table 1) on CentOS (version 7.2; Linux) along with initial 3 proof-of-authority blockchain nodes of DialysisNet on HAP. We set the block generation cycle of the nodes to 5 seconds. For the purpose of performance evaluation, we tested the use case scenarios using sample data sets including data elements for simulated hemodialysis patients' vital signs, laboratory results, and medications.

Table 1. Parameters delivered in data transmission scenarios. Parameters are considered as metadata for transmitted health data and must be stored and managed in the blockchain.

Scenario and steps	Departure	Destination	Name of parameter	Data type	Description
Agent or app sends data to Avatar					
1	Physician app	Patient Avatar	senderID	string	Unique identifier of the data sender (app)
			receiverID	string	Unique identifier of the data receiver (Avatar)
			dataSegment	JSON[a]	Sent data segment by the sender
			timestamp	datetime	Timestamp for data transmission
Agent or app requests data to Avatar					
1	Agent or physician app	Patient Avatar	API[b]	string	API syntax including requests for detailed data query
			senderID	string	Unique identifier of the data sender (Avatar)
			receiverID	string	Unique identifier of the data receiver (agent or app)
			timestamp	datetime	Timestamp for data transmission
2	Patient Avatar	Agent or physician app	dataSegment	JSON	Sent data segment by the sender.
Agent sends data to app					
1	Agent	Physician app	senderID	string	Unique identifier of the data sender (agent)
			receiverID	string	Unique identifier of the data receiver (app)
			timestamp	datetime	Timestamp for data transmission
			dataSegment	JSON	Sent data segment by the sender
Agent requests data to app					
1	Agent	Physician app	API	string	API syntax including requests for detailed data query
			senderID	string	Unique identifier of the data sender (app)
			receiverID	string	Unique identifier of the data receiver (agent)
			timestamp	datetime	Timestamp for data transmission
2	Physician app	Agent	dataSegment	JSON	Sent data segment by the sender

[a]JSON: JavaScript object notation.

[b]API: application programming interface.

Results

Smart Contracts and Use Cases

Each Ethereum node stores and manages transaction metadata during the course of all data exchanges on the HAP interconnected patient Avatars, physician apps, and intelligent agents. SC-1, as the health data transaction manager, stores *senderAddr* (the account address of the sender of the data segment), *receiverAddr* (the account address of the receiver), *HashedDS* (the hash value of the data segment through the

Secure Hash Algorithm-256 function), and *HashSeq* (a unique key for the transaction), which can also be used as a foreign key between contracts. SC-2, as the health data transaction status manager, manages the status of data transactions to be saved by SC-1. Finally, SC-3, as the HAP API transaction manager, was developed to manage information related to an agent's personal health record data requests. SC-3 manages the hash value of the requested API (Table 2).

Patient data are located in their smartphones (Avatar), physician's data for their patients are located in their smart Pads (XNet), agent's data for its customer are located in its server, and the health care institution's data are located in its electronic health record or other production servers. Thus, data are primarily stored and managed off-chain. All data transmission logs to proper receivers are on-chain through the HAP hash-and-relay server with a proper rationale and at a proper time (Figure 1). Node Manager manages information about each node that makes up the blockchain network. Information on personal health record data transactions are transmitted or requested to be traced to Blockchain Monitor (Multimedia Appendix 2). Blockchain verifies personal health record data managed off-chain in HAP or stores transaction metadata for verification. All transaction can be properly incentivized.

Table 2. Smart contracts (SC-1, SC-2, and SC-3) and variables in each contract.

Smart contract and variable	Data type	Description
SC-1: Health data transaction manager		
senderAddr	address	Address of the health data sender's Ether account
receiverAddr	address	Address of the health data receiver's Ether account
HashedDS	string	Hashed string value of data segment
HashSeq	uint256	Unique sequence for identification of the *HashedDS* value
SC-2: Health data transaction status manager		
contractAddr	address	Address of the smart contract account
HashSeq	uint256	Unique sequence for identification of the *HashedDS* value
status	string	Status of health data transaction. (eg, "waiting," "complete")
SC-3: HAP[a] API[b] transaction manager		
hashedAPI	string	Hashed string value of agent API syntax.
HashSeq	uint256	Unique sequence for identification of the *HashedDS* value

[a]HAP: Health Avatar Platform.

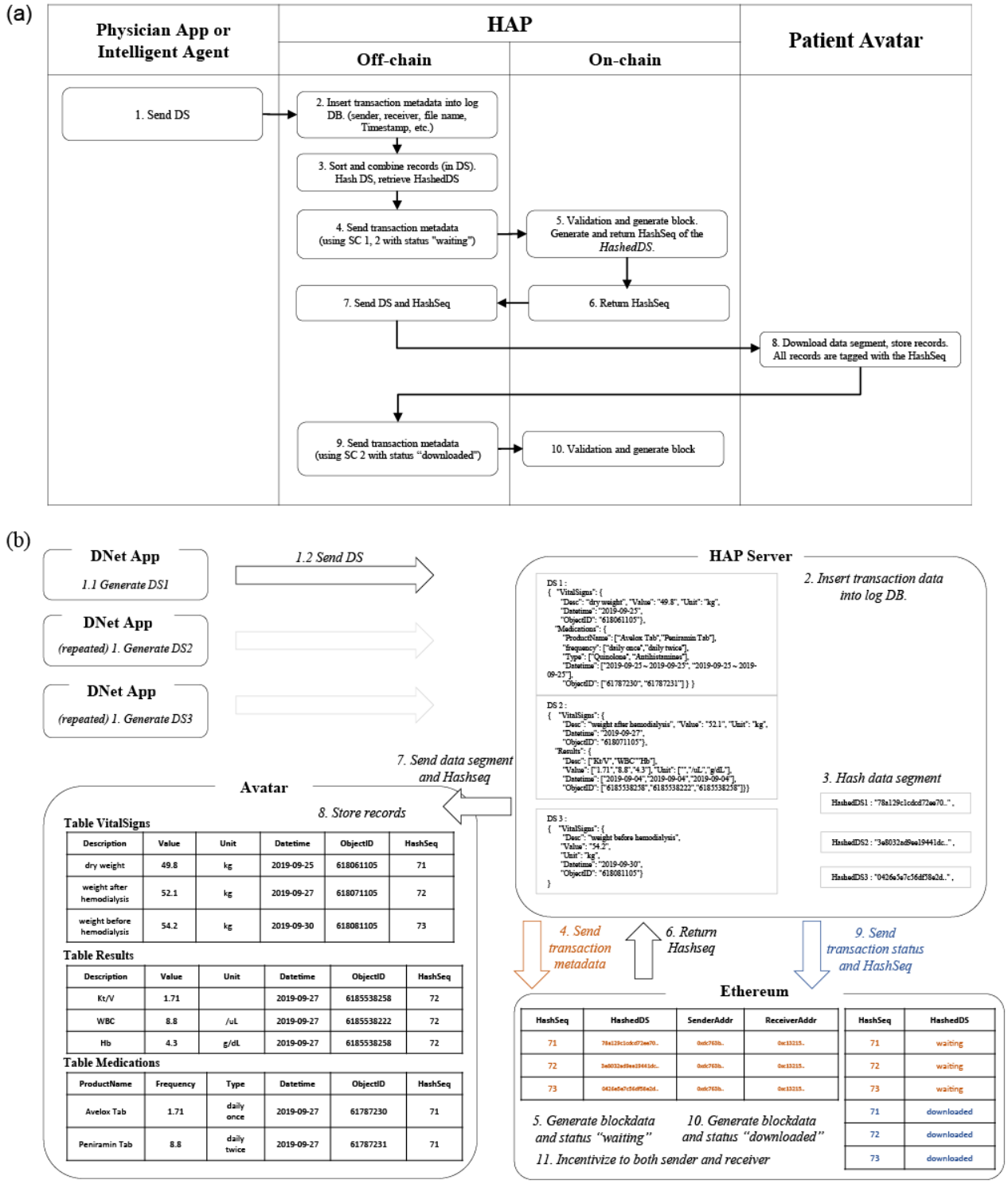
[b]API: application programming interface.

Push: Updating via the Physician App

A data segment has one or more data elements with values (sample data sets can be found in Multimedia Appendix 3:Figures S1-S3). When a physician app initiates a data transmission of vital signs (or laboratory results or medications) to a patient Avatar, the HAP relay server saves the time-stamped logs of the sender, recipient, and file name, sorts the data elements in the data segment, and extracts *HashedDS* from the data segment via Secure Hash Algorithm-256 (Figure 2a). HAP then transfers the extracted *HashedDS* and blockchain account addresses of the sender and receiver to smart contract SC-1 health data transaction manager for execution (Figure 2b). The blockchain node creates transaction metadata as block data with SC-1 health data transaction manager executed. By executing SC-2 health data transaction status manager, block data are created and tagged with the "waiting" status for the relevant data transaction. If block creation is successful, a *HashSeq* value is created by the blockchain and SC-1 returns this *HashSeq*. *HashSeq* is a sequence number created by SC-1 health data transaction manager to serve as a unique identifier corresponding to the *HashedDS* value. Through SC-1 health data transaction manager, *HashedDS* is mapped to this *HashSeq* and stored in the blockchain. Because *HashSeq* can function as a foreign key in the 3 smart contracts, the metadata for the exchanged data segment can be managed by normalizing relative to each contract.

When *HashSeq* is returned from the blockchain, the patient Avatar is enabled by the server with a push message to receive the data segment and *HashSeq*. The patient Avatar stores all records included in the downloaded data segment file in the personal health record database. These records are tagged with *HashSeq* and updated in the personal health record (or the patient Avatar database). When the personal health record update process is completed, the patient Avatar transmits its status information to HAP, indicating that data downloading is complete. The HAP server then updates the status information of the data segment to "complete" through SC-2 health data transaction status manager to record the completion of the patient's health data transmission and the personal health record update on the blockchain. Until the data transmission process is completed, metadata pertaining to the 3 data segments transmitted are maintained in blockchain storage (Figure 2).

Figure 2. Process of transmitting health data from a physician App or third-party Agent to the patient Avatar. Health data transaction hash logs are generated and updated via smart contracts in Ethereum blockchain. Steps of three separate data transmissions from a physician App to the patient Avatar for PHR update are demonstrated as (a) a workflow diagram and (b) detailed illustration. SC : Smart Contract; DS : Data Segment; DB: database; DNet: DialysisNet; HAP: Health Avatar Platform; PHR: personal health record.



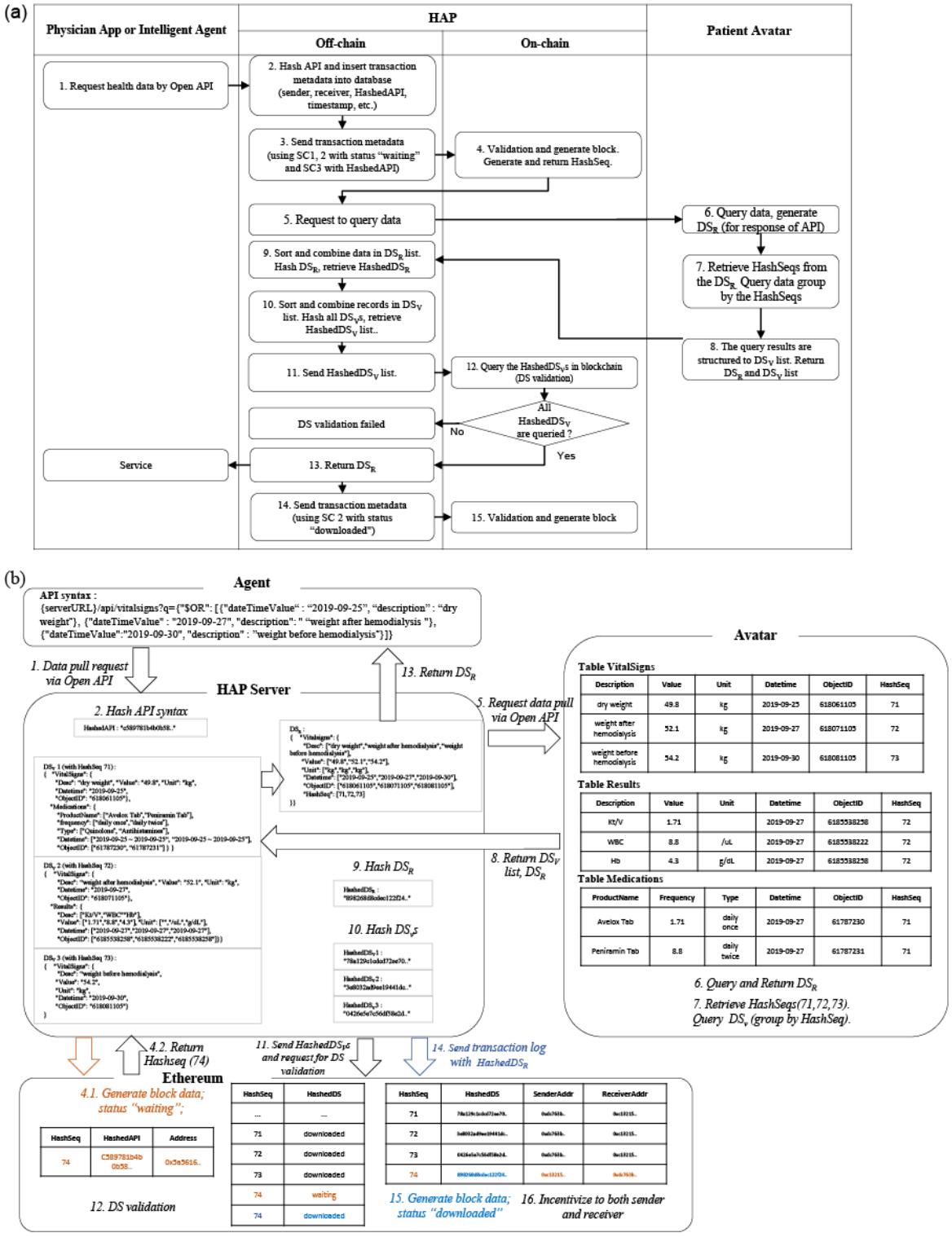
Pull: Requesting and Receiving Data

The HAP server relays the request (Figure 3) through the proper API to the patient Avatar (Multimedia Appendix 4: Figure S4). After authorization and authentication, the Avatar responds to a proper and trustworthy request by returning 2 types of data segments: data segments for the response (DSR) and data segments for validation (DSV), as the query response for the API request and the query result for data segment validation, respectively. Data segment validation is a process that checks whether there is a modulated record in the data segment before returning the query result to the requester. An example transaction scenario (Figure 3) is an intelligent agent requesting a patient Avatar for personal health record data including "dry weight," "weight after hemodialysis," and "weight before hemodialysis" through open API complying with the HAP

RESTful API syntax. The "/api/vitalSigns/" part of the API syntax refers to the database table *VitalSigns*, and the part next to *q* is the query string. The query string requests the dry weight measurement of "2019-09-25," weight after hemodialysis of "2019-09-27," and weight before hemodialysis of "2019-09-30." After performing authentication and authorization for the requesting agent, HAP transmits the data request to the corresponding patient Avatar. In response to the request, the patient's Avatar queries 2 types of data segments: DSV and DSR. First, DSR is extracted from the patient Avatar's personal health record database. Avatars' personal health record tables are devised on a data model that conforms to ASTM Continuity of Care Record and HL7 Continuity of Care Document standards. Health records previously delivered in the same data segment are tagged with the same *HashSeq* but are separately stored in 3 different tables (Figure 2b and Figure 3b) according to the data model. The corresponding data segments are pulled and processed to compile the query result for the requested API syntax and then returned to the requester.

Data segment validation, a process of verifying whether or not the transmitted DSR has been tampered with, is performed before the queried DSR is returned to the agent. A query using *HashSeq* included in the DSR is executed in the Avatar, resulting in a list of DSVs. Each DSV is bound to the *HashSeq* and regenerates *HashedDS* by sorting and hashing the records (or data elements) included in each DSV. If all regenerated *HashedDS*s are successfully retrieved from the blockchain, the DSV corresponding to the *HashedDS* has not been compromised. This also means that the records included in DSR have not been modified. Upon successful validation, information about the agent's data request by API is inserted into the block by the SC-3 HAP API transaction manager. This creates a block to update the status of the personal health record transaction to "complete" through the SC-2 health data transaction status manager. When transaction data generated by the agent API are created as block data, the server returns the DSR to the requesting agent. It was demonstrated that a smart contract in collaboration with data elements provided by metadata registry enable detailed data element–level query and access control beyond the resource-level query enabled by HL7 FHIR and other messaging standards. An authorized agent can provide highly personalized health care services without requesting an unnecessary amount of data beyond its declared capability and beyond what is authorization by HAP.

Figure 3. Process of requesting patient data and receiving data by an intelligent agent or a physician app: (a) workflow diagram and (b) detailed example of data flow initiated by an intelligent agent (or a physician app) requesting patient data stored in a patient Avatar for the purpose of providing clinical recommendations via Open API. DSV: data segment for validation; DSR: data segment for response; HAP: Health Avatar Platform; SC: smart contract.

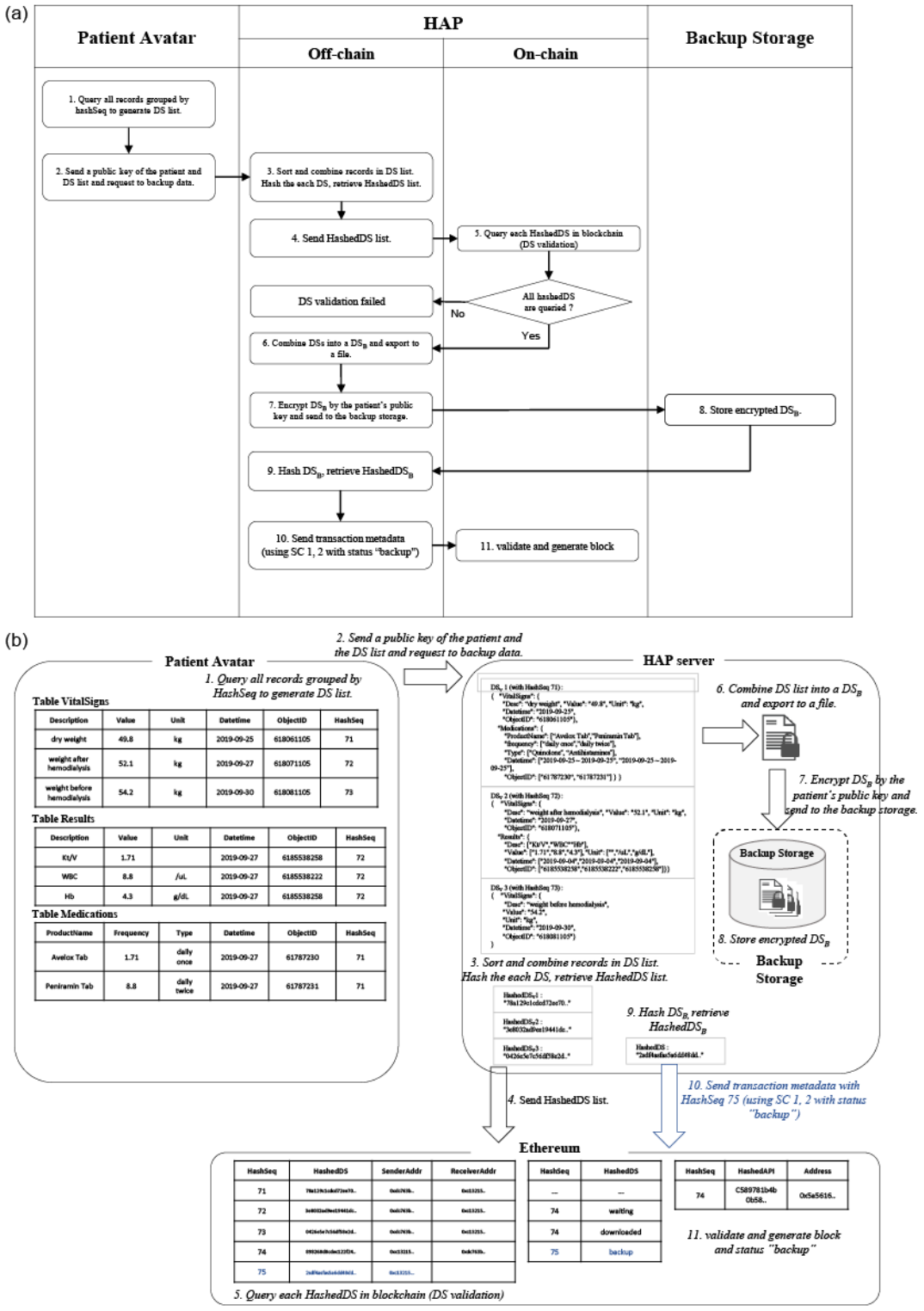


Data Backup Process

For the purpose of strong data privacy protection, HAP does not store any health data that are transmitted through the server; however, data backups are necessary for many purposes under strict patient control (Figure 4, Multimedia Appendix 4: Figure S5). When a patient initiates the backup process, the patient Avatar queries all personal health record data except for the patient identifier and transmits it to the HAP API with a backup request. The API allows one to output the data set grouped by *HashSeq*. Data segments transmitted through HAP will be hashed, and each *HashedDS* is created for validation. If validated passes, it means that the data segments to be backed up have not been altered and they are sent to the backup storage. Before the data segments are saved into backup storage, the verified

segments are integrated into a data segment file. Using the patient's public key, the file is encrypted by the RSA (Rivest–Shamir–Adleman [40]) method and stored in the backup storage. When the encrypted file is stored in storage, the server creates *HashedDS* after hashing the integrated data segment of the file. To record the completion of the transaction in storage, the *HashedDS* "2sdf4asfas5a6dd48dd..." is connected to *HashSeq* 75 and stored. The blockchain stores the status of the transaction as "backup."

Figure 4. Data backup process: (a) workflow diagram and (b) use-case illustration initiated by an entity. DS: data segment; DSB: data segment for backup; HAP: Health Avatar Platform; SC: smart contract.



Discussion

Principal Findings

The legacy HAP successfully performed decentralized health data management. From a data management perspective, decentralized management of personal health record with a patient's smartphone app is less efficient than a centralized approach; however, in terms of privacy protection and patient empowerment, decentralization is better for creating a highly interconnected mobile health ecosystem. We built a decentralized system and performed real-world clinical-practice validation with DialysisNet and RehabilitationNet. The platform successfully prevented data reuse and personal information leakages based on the trust of the system. The introduction of the blockchain and smart contracts significantly improved the efficiency and effectiveness of our decentralized health data management method. The adoption of blockchain to the legacy HAP system inevitably incurs overhead (Multimedia Appendix 5); however, we observed that the overhead, 32.58 ms on average, added to the legacy system was minimized by introducing asynchronous blockchain connection. The platform demonstrated that blockchain is a suitable software tool that safely and efficiently performs the required data verification and decentralized data backup processes.

HAP provides semantic interoperability for all data exchanges in the system. ASTM Continuity of Care Record and HL7 Continuity of Care Document standards were applied as a syntactic backbone required for HAP data management; however, syntactic standards alone are insufficient for a unified specification (eg, data type, format) for all data exchanges on the platform. Thus, we installed XNetHub in each health care institution (Figure 1) and metadata registry to ensure semantic interoperability among different electronic health records. HAP XNetHub supports HL7 FHIR, allowing resource-level health data queries. HAP treats a message (or a health record) as a collection of data segments composed of data elements, which are defined and managed by a metadata server in compliance with ISO/IEC 11179 metadata registry standards and provide thousands of expert curated common data elements required for hemodialysis patient management. One unique advantage of our work is that the blockchain-enabled HAP allows data segment–/data element–level querying and health data processing that are fully authenticated and audit trailed by the enabling technologies such as the immutable hash and sub–hash management schemes (Figures 2-4) with smart contracts (Table 2). In contrast, HL7 FHIR's resource level data management does not allow granularity that is fine enough for health data querying or processing.

The introduction of metadata registry on top of these syntactic standards with predefined, preregistered, and postexpandable common data elements, highly enriched in semantics by means of standard vocabulary and ontology mappings, further improves semantic queries to each data element value level. Furthermore, we demonstrated that data segment– and data element–level data verifications were enabled by this architecture. A metadata registry improves the semantic interoperability of health data exchanges [8,41,42]. The semantic layer allows patients to integrate their health records from multiple health care institutions. Physicians can consolidate the health records of patients from different institutions. Moreover, third parties can have an integrated view of patients' health records through HAP RESTful APIs.

Blockchain and smart contract technologies were used in this platform to enhance the security of patient-centered personal health record transactions and the efficiency of decentralized data management. Additionally, for exchanges of patient data that may occur on the platform, HAP can provide incentives for data sharing to parties with whom the data are being exchanged. Many health care systems adopting fee-for-service reimbursement mechanism mainly reward highly materialized clinical services, such as medications, laboratory testing, or interventions, but lack sufficient reward systems for education, exercise, prevention, or long-term management that are more relevant for chronic conditions, which are ever increasing. Given all of these advantages, the HAP interconnected patient Avatars, physician apps, and intelligent agents system can become an ecosystem that promotes the reliable sharing of health data performed with patient empowerment.

Limitations

Due to the features of the proof-of-authority consensus algorithm, a delay during block generation equal to the setting in the genesis block occurs; however, in this prototype system, the block data are generated through an asynchronous on-chain process apart from off-chain transactions for health data, meaning that there are no delays in off-chain data transactions. Another challenge arises when verifying the patient Avatar's personal health record data (through data backup or data query processes)—when a large message is exchanged, the speed of data verification and the return of the verification result may be slower. Accordingly, it may be necessary in the future to calculate the data alignment method included in the DSV and the appropriate time required during the process of hashing the data segment. For this process, a trade-off study on the time required for data processing and the size of the transmitted data segment may be required.

Conclusions

We designed and built an ecosystem that provides efficient and effective decentralized health data management and exchange operations by applying a prototype blockchain and smart contract to a patient device–based personal health record system. It was demonstrated that health data access control and authenticity verification of personal health record data are enabled not only at the overall personal health record or resource level but also at granular data element and data value levels.

Acknowledgments

This study was funded by the Korean Health Technology Research and Development Project by the Ministry of Health and Welfare in the Republic of Korea (HI18C2386).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Transmitted data segment.
[PNG File , 57 KB - medinform_v9i6e26230_app1.png]

Multimedia Appendix 2
Health Avatar Platform participant service modules.
[PNG File , 162 KB - medinform_v9i6e26230_app2.png]

Multimedia Appendix 3
Comparison of blockchain-based health information systems.
[DOCX File , 18 KB - medinform_v9i6e26230_app3.docx]

Multimedia Appendix 4
Data segments with demo data sets.
[DOCX File , 22 KB - medinform_v9i6e26230_app4.docx]

Multimedia Appendix 5
Performance evaluation.
[DOCX File , 142 KB - medinform_v9i6e26230_app5.docx]

References

1. Horowitz J, Mon D, Bernstein B, Bell K. Defining key health information technology terms. Office of the National Coordinator for Health Information Technology. 2008. URL: https://s3.amazonaws.com/rdcms-himss/files/production/public/HIMSSorg/Content/files/Code%205%20Defining%20Key%20Health%20Information%20Technology%20Terms.pdf [accessed 2021-05-21]
2. Tang PC, Ash JS, Bates DW, Overhage JM, Sands DZ. Personal health records: definitions, benefits, and strategies for overcoming barriers to adoption. J Am Med Inform Assoc 2006;13(2):121-126 [FREE Full text] [doi: 10.1197/jamia.M2025] [Medline: 16357345]
3. Graetz I, Huang J, Brand RJ, Hsu J, Yamin CK, Reed ME. Bridging the digital divide: mobile access to personal health records among patients with diabetes. Am J Manag Care 2018 Jan;24(1):43-48 [FREE Full text] [Medline: 29350505]
4. Kahn JS, Aulakh V, Bosworth A. What it takes: characteristics of the ideal personal health record. Health Aff (Millwood) 2009;28(2):369-376. [doi: 10.1377/hlthaff.28.2.369] [Medline: 19275992]
5. Hsieh G, Chen RJ. Design for a secure interoperable cloud-based personal health record service. 2012 Presented at: 4th IEEE International Conference on Cloud Computing Technology and Science; December 3-6; Taipei, Taiwan p. 472-479. [doi: 10.1109/CLOUDCOM.2012.6427582]
6. Liu L, Shih P, Hayes G. Barriers to the adoption and use of personal health record systems. In: Proceedings of the 2011 iConference. 2011 Feb 08 Presented at: iConference; February 8-11; Seattle, Washington p. 363-370. [doi: 10.1145/1940761.1940811]
7. Tyagi S, Agarwal A, Maheshwari P. A conceptual framework for IoT-based healthcare system using cloud computing. 2016 Presented at: 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence); January 14-15; Noida, India p. 503-507. [doi: 10.1109/confluence.2016.7508172]
8. Park YR, Kim JH. Metadata registry and management system based on ISO 11179 for Cancer Clinical Trials Information System. AMIA Annu Symp Proc 2006:1056 [FREE Full text] [Medline: 17238675]
9. Avatar Beans. Google Play Store. URL: https://play.google.com/store/apps/details?id=org.snubi.avatarbeans [accessed 2021-02-01]
10. Avatar Beans. Apple App Store. URL: https://apps.apple.com/kr/app/아바타빈즈/id1356376744 [accessed 2021-02-01]
11. Kim H, Park J, Yoo K, Kim Y, Baek H, Kim SH, et al. Real-world treatment patterns of renal anemia in hemodialysis patients: a multicenter cohort study performed using DialysisNet (RRAHD study). Medicine (Baltimore) 2020 Jan;99(2):e18749 [FREE Full text] [doi: 10.1097/MD.0000000000018749] [Medline: 31914095]

12. Wright CS. Bitcoin: a peer-to-peer electronic cash system. SSRN 2019:1-9. [doi: 10.2139/ssrn.3440802]
13. Zheng Z, Xie S, Dai HN, Chen X, Wang H. Blockchain challenges and opportunities: a survey. IJWGS 2018;14(4):352. [doi: 10.1504/ijwgs.2018.095647]
14. Crosby M, Pattanayak P, Verma S, Kalyanaraman V. Blockchain technology: beyond bitcoin. Applied Innovation. 2016 Jun. URL: https://j2-capital.com/wp-content/uploads/2017/11/AIR-2016-Blockchain.pdf [accessed 2021-05-21]
15. Buterin V. A next-generation smart contract and decentralized application platform. Blockchain Lab. 2014. URL: https://blockchainlab.com/pdf/Ethereum_white_paper-a_next_generation_smart_contract_and_decentralized_application_platform-vitalik-buterin.pdf [accessed 2021-05-27]
16. Christidis K, Devetsikiotis M. Blockchains and smart contracts for the internet of things. IEEE Access 2016;4:2292-2303. [doi: 10.1109/ACCESS.2016.2566339]
17. Bartoletti M, Pompianu L. An empirical analysis of smart contracts: platforms, applications, and design patterns. In: Brenner M, editor. Financial Cryptography and Data Security. Cham: Springer; 2017:494-509.
18. Azaria A, Ekblaw A, Vieira T, Lippman A. MedRec: using blockchain for medical data access and permission management. In: Proceedings of the 2nd International Conference on Open and Big Data. 2016 Presented at: 2nd International Conference on Open and Big Data; August 22-24; Vienna, Austria p. 25-30. [doi: 10.1109/OBD.2016.11]
19. Dagher GG, Mohler J, Milojkovic M, Marella PB. Ancile: privacy-preserving framework for access control and interoperability of electronic health records using blockchain technology. Sustain Cities Soc 2018 May;39:283-297. [doi: 10.1016/j.scs.2018.02.014]
20. Xia Q, Sifah E, Smahi A, Amofa S, Zhang X. BBDS: blockchain-based data sharing for electronic medical records in cloud environments. Information 2017 Apr 17;8(2):44. [doi: 10.3390/info8020044]
21. Xia Q, Sifah EB, Asamoah KO, Gao J, Du X, Guizani M. MeDShare: trust-less medical data sharing among cloud service providers via blockchain. IEEE Access 2017;5:14757-14767. [doi: 10.1109/access.2017.2730843]
22. Dubovitskaya A, Xu Z, Ryu S, Schumacher M, Wang F. Secure and trustable electronic medical records sharing using blockchain. AMIA Annu Symp Proc 2017;2017:650-659 [FREE Full text] [Medline: 29854130]
23. Motohashi T, Hirano T, Okumura K, Kashiyama M, Ichikawa D, Ueno T. Secure and scalable mhealth data management using blockchain combined with client hashchain: system design and validation. J Med Internet Res 2019 May 16;21(5):e13385 [FREE Full text] [doi: 10.2196/13385] [Medline: 31099337]
24. Vazirani AA, O'Donoghue O, Brindley D, Meinert E. Blockchain vehicles for efficient medical record management. NPJ Digit Med 2020 Jan 06;3(1):1 [FREE Full text] [doi: 10.1038/s41746-019-0211-0] [Medline: 31934645]
25. Hylock RH, Zeng X. A blockchain framework for patient-centered health records and exchange (HealthChain): evaluation and proof-of-concept study. J Med Internet Res 2019 Aug 31;21(8):e13592 [FREE Full text] [doi: 10.2196/13592] [Medline: 31471959]
26. Roehrs A, da Costa CA, da Rosa Righi R. OmniPHR: A distributed architecture model to integrate personal health records. J Biomed Inform 2017 Jul;71:70-81 [FREE Full text] [doi: 10.1016/j.jbi.2017.05.012] [Medline: 28545835]
27. Rajput AR, Li Q, Taleby Ahvanooey M, Masood I. EACMS: emergency access control management system for personal health record based on blockchain. IEEE Access 2019;7:84304-84317. [doi: 10.1109/access.2019.2917976]
28. Thwin TT, Vasupongayya S. Blockchain-based access control model to preserve privacy for personal health record systems. Secur Commun Netw 2019 Jun 25;2019:1-15. [doi: 10.1155/2019/8315614]
29. Wang S, Zhang Y, Zhang Y. A blockchain-based framework for data sharing with fine-grained access control in decentralized storage systems. IEEE Access 2018;6:38437-38450. [doi: 10.1109/access.2018.2851611]
30. Lee H, Kung H, Udayasankaran JG, Kijsanayotin B, B Marcelo A, Chao LR, et al. An architecture and management platform for blockchain-based personal health record exchange: development and usability study. J Med Internet Res 2020 Jun 09;22(6):e16748 [FREE Full text] [doi: 10.2196/16748] [Medline: 32515743]
31. Rahmadika S, Rhee K. Blockchain technology for providing an architecture model of decentralized personal health information. Int J Eng Bus Manag 2018 Aug;10:184797901879058. [doi: 10.1177/1847979018790589]
32. Senor IC, Aleman JLF, Toval A. Personal health records: new means to safely handle health data? Computer 2012 Nov;45(11):27-33. [doi: 10.1109/mc.2012.285]
33. Li J. Ensuring privacy in a personal health record system. Computer 2015 Feb;48(2):24-31. [doi: 10.1109/mc.2015.43]
34. Liu J, Huang X, Liu JK. Secure Secure sharing of personal health records in cloud computing: ciphertext-policy attribute-based signcryption. Future Gener Comput Syst 2015 Nov;52:67-76. [doi: 10.1016/j.future.2014.10.014]
35. Kuo T, Zavaleta Rojas H, Ohno-Machado L. Comparison of blockchain platforms: a systematic review and healthcare examples. J Am Med Inform Assoc 2019 May 01;26(5):462-478 [FREE Full text] [doi: 10.1093/jamia/ocy185] [Medline: 30907419]
36. Macdonald M, Liu-Thorrold L, Julien R. The blockchain: a comparison of platforms and their uses beyond bitcoin. Research Gate. 2017. URL: https://www.researchgate.net/publication/313249614_The_Blockchain_A_Comparison_of_Platforms_and_Their_Uses_Beyond_Bitcoin [accessed 2021-05-25]
37. Chowdhury MJM, Ferdous MS, Biswas K, Chowdhury N, Kayes ASM, Alazab M, et al. A comparative analysis of distributed ledger technology platforms. IEEE Access 2019;7:167930-167943. [doi: 10.1109/access.2019.2953729]

38. Proof-of-Authority Chains - Wiki. openethereum. URL: https://openethereum.github.io/Proof-of-Authority-Chains [accessed 2021-01-01]
39. Wood G. Ethereum: a secure decentralised generalised transaction ledger. Ethereum. 2014. URL: https://ethereum.github.io/yellowpaper/paper.pdf [accessed 2021-05-27]
40. Rivest RL, Shamir A, Adleman L. A method for obtaining digital signatures and public-key cryptosystems. Commun ACM 1978 Feb;21(2):120-126. [doi: 10.1145/359340.359342]
41. Park YR, Yoon YJ, Kim HH, Kim JH. Establishing semantic interoperability of biomedical metadata registries using extended semantic relationships. Stud Health Technol Inform 2013;192:618-621. [Medline: 23920630]
42. Sinaci AA, Laleci Erturkmen GB. A federated semantic metadata registry framework for enabling interoperability across clinical research and care domains. J Biomed Inform 2013 Oct;46(5):784-794 [FREE Full text] [doi: 10.1016/j.jbi.2013.05.009] [Medline: 23751263]

Abbreviations

API: application programming interface
DSR: data segment for response
DSV: data segment for validation
FHIR: Fast Healthcare Interoperability Resources
HAP: Health Avatar Platform
IoT: Internet of Things
ISO/IEC: International Organization for Standardization/International Electrotechnical Commission
SC: smart contract
XNet: physician apps
XNetHub: institutional gateway server

Edited by G Eysenbach, R Kukafka; submitted 03.12.20; peer-reviewed by TT Kuo, D Kim, C Reis; comments to author 31.12.20; revised version received 12.02.21; accepted 03.04.21; published 07.06.21.

Please cite as:
Kim HJ, Kim HH, Ku H, Yoo KD, Lee S, Park JI, Kim HJ, Kim K, Chung MK, Lee KH, Kim JH
Smart Decentralization of Personal Health Records with Physician Apps and Helper Agents on Blockchain: Platform Design and Implementation Study
JMIR Med Inform 2021;9(6):e26230
URL: https://medinform.jmir.org/2021/6/e26230
doi:10.2196/26230
PMID:34096877

Original Paper

A Word Pair Dataset for Semantic Similarity and Relatedness in Korean Medical Vocabulary: Reference Development and Validation

Yunjin Yum[1,2], MS; Jeong Moon Lee[2], BS; Moon Joung Jang[2], BS; Yoojoong Kim[2], PhD; Jong-Ho Kim[2,3], PhD; Seongtae Kim[4], BA; Unsub Shin[4], BA; Sanghoun Song[4*], PhD; Hyung Joon Joo[3,5,6*], MD, PhD

[1]Department of Biostatistics, Korea University College of Medicine, Seoul, Republic of Korea

[2]Korea University Research Institute for Medical Bigdata Science, Korea University, Seoul, Republic of Korea

[3]Department of Cardiology, Cardiovascular Center, Korea University College of Medicine, Seoul, Republic of Korea

[4]Department of Linguistics, Korea University, Seoul, Republic of Korea

[5]Korea University Research Institute for Medical Bigdata Science, Korea University Anam Hospital, Seoul, Republic of Korea

[6]Department of Medical Informatics, Korea University College of Medicine, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Hyung Joon Joo, MD, PhD
Korea University Research Institute for Medical Bigdata Science
Korea University Anam Hospital
145 Anam-ro, Seoungbuk-gu
Seoul, 02841
Republic of Korea
Phone: 82 010 3476 0526
Email: drjoohj@gmail.com

Abstract

Background: The fact that medical terms require special expertise and are becoming increasingly complex makes it difficult to employ natural language processing techniques in medical informatics. Several human-validated reference standards for medical terms have been developed to evaluate word embedding models using the semantic similarity and relatedness of medical word pairs. However, there are very few reference standards in non-English languages. In addition, because the existing reference standards were developed a long time ago, there is a need to develop an updated standard to represent recent findings in medical sciences.

Objective: We propose a new Korean word pair reference set to verify embedding models.

Methods: From January 2010 to December 2020, 518 medical textbooks, 72,844 health information news, and 15,698 medical research articles were collected, and the top 10,000 medical terms were selected to develop medical word pairs. Attending physicians (n=16) participated in the verification of the developed set with 607 word pairs.

Results: The proportion of word pairs answered by all participants was 90.8% (551/607) for the similarity task and 86.5% (525/605) for the relatedness task. The similarity and relatedness of the word pair showed a high correlation (ρ=0.70, P<.001). The intraclass correlation coefficients to assess the interrater agreements of the word pair sets were 0.47 on the similarity task and 0.53 on the relatedness task. The final reference standard was 604 word pairs for the similarity task and 599 word pairs for relatedness, excluding word pairs with answers corresponding to outliers and word pairs that were answered by less than 50% of all the respondents. When FastText models were applied to the final reference standard word pair sets, the embedding models learning medical documents had a higher correlation between the calculated cosine similarity scores compared to human-judged similarity and relatedness scores (namu, ρ=0.12 vs with medical text for the similarity task, ρ=0.47; namu, ρ=0.02 vs with medical text for the relatedness task, ρ=0.30).

Conclusions: Korean medical word pair reference standard sets for semantic similarity and relatedness were developed based on medical documents from the past 10 years. It is expected that our word pair reference sets will be actively utilized in the development of medical and multilingual natural language processing technology in the future.

(JMIR Med Inform 2021;9(6):e29667) doi:10.2196/29667

KEYWORDS

medical word pair; similarity; relatedness; word embedding; fastText; Korean

Introduction

The rapid development of natural language processing (NLP) technology in tandem with advances in artificial intelligence and deep learning have greatly influenced our day-to-day life. In the field of medical service, there are high expectations that NLP will be able to improve patient-physician communication and provide basic medical information support for patients in the blind spot of medical care. Several commercial chatbot applications are now available on the web or mobile environments [1]. For example, OneRemission provides useful information for patients with cancer. Babylon Health also provides symptom-based medical consultation services.

However, medical terms require special expertise and are challenging to decipher not only for ordinary people but also, at times, for medical experts. This raises 2 fundamental requirements with regard to medical NLP technology. First, as medical terms appear sparsely in plain text, a vast amount of medical-specific text data is required to improve the current medical NLP technology. Second, as medical terms often convey a wide range of meanings in different contexts, it is crucial to build a semantic network based on a comprehensive set of medical terminologies via word embedding. For example, BioWordVec is a set of biomedical word embeddings involving 2,324,849 words from Medical Subject Headings (MeSH) and PubMed [2]. BioConceptVec was created by learning domain-specific vector representations of biological concepts (eg, genes, drugs, and proteins) using large-scale biomedical corpora [3]. BioBERT, a biomedical-specific language model, was constructed using approximately 18 billion words from PubMed abstracts and PubMed Central full-text articles [4]. Although English is the main language being used in the field of medical NLP, multilingual approaches involving other languages (eg, Chinese, German, French, Italian, Japanese, Korean) are also being investigated [5,6]. Technical validation of language embedding models is also highly important in the field of medical NLP. Only a few standard datasets have so far been introduced. University of Minnesota Semantic Relatedness Set (UMNSRS) datasets (566 pairs for similarity and 587 pairs for relatedness) were developed by involving 8 medical residents [7]. Previously, a Mayo semantic relatedness set of 101 medical term pairs evaluated by 13 medical coding experts was proposed [8]. Both of the aforementioned datasets were created more than a decade ago. Many important medical discoveries have been made over the past decade, and as a result, the medical procedures have also undergone changes. Therefore, the erstwhile reference standards do not necessarily involve the current knowledge of medical science. This raises the necessity of updating the standard datasets that are being used for NLP model validation.

This study aimed to propose a new standard word pair set especially for Korean medical terms. We generated a large amount of text data using Korean medical terms through academic papers, websites, and textbooks and selected words in consideration of the frequency of appearance of each word. Concept ratings for the new standard word pair set were set by highly qualified attending physicians from a tertiary hospital. Finally, we evaluated the developed word embedding model to demonstrate its feasibility.

Methods

Data Acquisition

We collected 3 types of Korean medical documents. Medical research articles were selected as high-quality documents at the professional level, health information news articles were selected as popular and general documents, and medical textbooks were selected as intermediate level but very high-quality documents. First, regarding medical textbooks, 2 Korean publishing companies provided textbooks for the present study. Each publisher had a classification system for the subject of books and textbooks in the 54 subfields (eg, internal medicine, emergency medicine, orthopedics, dentistry, pharmacy, public health) of the medical field that were selected. Finally, 518 text files from Korean medical textbooks published from 2010 to 2020 were used. Second, regarding health information news, NAVER, a widely used internet portal site in Korea, distributes news articles from general newspapers, internet newspapers, and broadcasting stations. We collected all the news articles in a section called “Health Information” under the “News” section of the NAVER portal. Finally, 72,844 health information news articles published from January 1, 2010 to December 31, 2020 in NAVER were collected through internet crawling. Third, regarding medical research articles, 72 journals (eg, Journal of The Korean Society of Integrative Medicine, Journal of The Korean Society of Emergency Medicine) published in the Korean language were selected from the journals listed in the Korean Studies Information Service System (KISS). Finally, 15,698 medical research articles published from 2010 to 2020 were collected. The text files were parsed and modified for further investigation. Consequently, we were able to build a large corpus comprising 191 million tokens in terms of morphemes (129 million tokens in terms of Korean words).

Pair Set Development

The top 10,000 nouns were selected in order of occurrence frequency in the corpus. Those terms were then categorized by 2 experienced medical vocabulary experts (a certified health information manager and a medical physician) into the following 5 categories: “symptom and sign,” “diagnosis,” “medication,” “operation and procedure,” and “not applicable.” Because the 10,000 nouns included nonmedical terms, the “not applicable” items were manually filtered out. This process left 1214 medical terms in total (625 diagnoses, 277 symptoms and signs, 177 medications, 135 operations and procedures). After the medical term selection, 607 medical term pair sets were manually developed considering the distribution of similarity and

relatedness for each pair set. The similarity and relatedness for each pair set were categorized into 4 groups (very dissimilar/unrelated, somewhat dissimilar/unrelated, somewhat similar/related, and very similar/related). The order of presentation of the pair sets and the order of the terms in each pair set were randomized for each participant.

Human Validation

For the validation of the medical term pair set, 16 attending physicians (2 cardiologists, 1 gastroenterologist, 3 nephrologists, 1 endocrinologist, 1 oncologist, 1 infectious medicine doctor, 1 family medicine doctor, 2 pediatricians, 1 psychiatrist, 1 emergency medicine doctor, 1 radiologist, and 1 general surgeon) at Korea University Anam Hospital were recruited. The study protocol was approved by the Institutional Review Board of Korea University Anam Hospital (IRB No. 2021AN0059). Written informed consent was obtained from all participants at enrollment. Our study complied with the principles of the Declaration of Helsinki.

The participants were randomized into 1 of 2 tasks (similarity or relatedness); 8 participants were assigned to the similarity evaluation group, and the other 8 were assigned to the relatedness evaluation group. The tasks were explained to the participants with examples (eg, “myocardial infarction” and “chest pain” are related but not similar; “cardiovascular disease” and “coronary artery disease” are similar). The 8 participants were separated into quiet rooms, where they sat in front of a 15-inch laptop. Relatedness and similarity evaluations for each medical term pair set were performed on a laptop using a toolkit used in psychological experiments known as OpenSesame [9]. The evaluations were performed using a 10-point scale, ranging from 1 to 10: the greater the value, the higher the similarity or relatedness [10]. In the case of a word pair set that was difficult to answer, the participants were asked to skip it by entering “x.” In the case of word pair sets from different semantic domains, there was a possibility that the participants may unknowingly answer the relatedness rather than the similarity, especially owing to time constraints. This could lead to a somewhat low degree of correspondence in terms of scoring in the similarity task. Therefore, to minimize the potential error and bias, no time limit was given for answering. However, the subjects were instructed to answer the word pairs in an intuitive manner. To maintain the degree of concentration on the task, after each 200 word pairs, the participant was allowed to autonomously take a break of 3-5 minutes. To minimize practice effects, a practice session consisting of 15-24 word pairs not included in the main word pair sets was provided before the main evaluation (Table 1).

Table 1. Examples of the practice session for human validation in which word pairs of the practice session included medical as well as general terms. Term 1 and Term 2 were presented to the participants. However, their anticipated similarity and relatedness categories were kept hidden.

Term 1	Term 2	Anticipated category
책방 (bookstore)	서점 (bookshop)	Similarity: high
학교 (school)	경찰서 (police station)	Similarity: middle
까치 (magpie)	중국어 (the Chinese language)	Similarity: low
친구 (friend)	사람 (human)	Relatedness: high
겨울 (winter)	난로 (heater)	Relatedness: middle
핸드폰 (cell phone)	미술 (art)	Relatedness: low
심혈관질환 (cardiovascular disease)	관상동맥질환 (coronary artery disease)	Similarity: high
암성통증 (cancer pain)	월경통 (menstrual pain)	Similarity: middle
좌골신경통 (sciatica)	간성혼수 (hepatic coma)	Similarity: low
심근경색 (myocardial infarction)	흉통 (chest pain)	Relatedness: high
세티리진 (cetirizine)	구강건조 (dry mouth)	Relatedness: middle
백반증 (vitiligo)	라미 (lamisil)	Relatedness: low

Embedding Model Validation

We applied 2 unsupervised word embedding models (Word2Vec [11] and FastText [12]) to the Korean medical word pair sets, and the results were compared with those of the human evaluation. A Korean medicine–focused corpus with 129 million words (aforementioned) was used for model training. The preprocessing of the obtained corpus was twofold. First, the entire corpus data were segmented using the Korean Sentence Splitter (KSS) 2.2.0.2. Second, the Mecab-ko tagger 0.4.0 was used to convert the sentences into morphological tokens [13]. Then, the models were built using the Gensim Python library [14]. The details related to the tuning of the model hyperparameters are listed in Table 2.

The cosine distance between the embedded concepts of word pairs was calculated and compared with that of the human evaluation.

Table 2. Hyperparameters of Word2Vec and FastText.

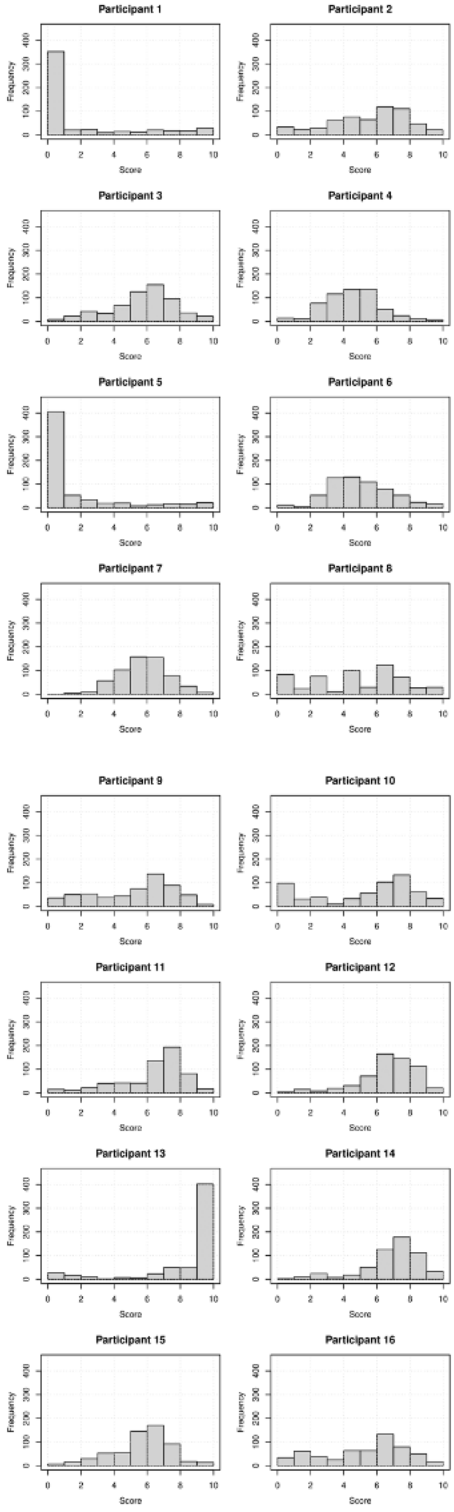
Parameter name	Specified argument
Dimension size	300
Window size	5
Negative sampling ratio	10%
Minimum frequency	10
Workers	3
Batch words	10,000
Alpha	0.25%
Epochs	20

Statistical Analyses

The word pair set was organized in a variety of ways, ranging from nonsimilar/related to closely similar/related. Although the participants received instruction, the distribution of the scores of some participants was found to be highly skewed (Figure 1). The scores of the participants whose absolute skewness value of the score distribution was 1.5 or higher were excluded. The scores of 6 participants in the similarity task and 7 participants in the relatedness task were finally included for further analyses.

The intraclass correlation coefficient (ICC) was calculated to measure the interrater reliability. Among the models defined by Shrout and Fleiss [15], the ICC(2,1) was used. This model is a two-way random effects model based on a single rater used for generalizing reliability results. The definition of consistency was chosen to consider the scores of the subjects that were correlated in word pairs. To compare the distribution between the original and modified data, the Kolmogorov-Smirnov statistic was used. Spearman rank correlation was used to determine the relationship between similarity and relatedness and to compare the magnitude of the automated measures of relatedness and similarity representing the human annotated scores.

Figure 1. Score distribution plots of the participants (n=16 attending physicians from a tertiary hospital); the distributions of the scores of participants 1, 5, and 13 were absolutely skewed and were therefore excluded from further analyses.



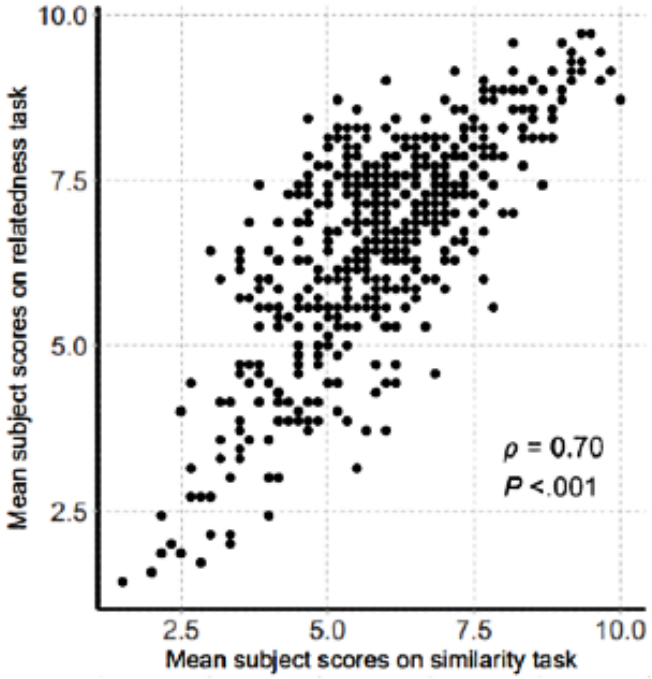
Results

Words in the standard reference dataset should be sufficiently representative of medical terms, and if so, they should show a high response rate. The previous UMNSRS dataset showed 81.1% (587/724) and 78.2% (566/724) response rates on the relatedness and similarity tasks, respectively [7]. This study revealed that the word pair sets scored by all the assigned participants on the relatedness task were 86.5% (525/607) of the 607 word pair sets, and on the similarity task, these were 90.8% (551/607) of the 607 word pair sets. This indicates that the words in the present dataset have some degree of representation as medical terms. The most common scoring failure cases were word pair sets including “diagnosis” category words on both relatedness and similarity tasks (99/153, 64.7% and 51/77, 66% of all failures for relatedness and similarity

tasks, respectively). This failure to score can be attributed to the fact that diagnostic approaches are rapidly developing and becoming increasingly specialized. The highest failure rate of scoring for the diagnosis-related word pairs in this study is believed to reflect the rapid pace of development in the medical field. Although some word pairs were not scored, all 607 word pair sets were scored by more than 4 (4/8, 50%) participants in each task group. The average response time per word pair set was 4.4 (SD 2.9) seconds on the relatedness task and 3.0 (SD 2.4) seconds on the similarity task. In the relatedness and similarity tasks, 71.8% (3050/4249) and 90.2% (3284/3642), respectively, of all the responses were completed within 5 seconds. This indicates that most of the responses were answered in an intuitive manner.

The scores between relatedness and similarity were highly correlated (ρ=0.70, *P*<.001; Figure 2). Mean score was higher in the relatedness task as compared to the similarity task (6.6, SD 1.6 vs 5.9, SD 1.5; *P*<.001). Further, compared to relatedness, it was difficult to assess the similarity between medical terms that were from different semantic domains (eg, “myocardial infarction” [disease domain] and “percutaneous coronary intervention” [procedure domain]). Therefore, several word pairs, which can be considered as somewhat similar or even highly related, were determined as dissimilar (upper left side of the plot in Figure 1).

Figure 2. Scatter plot of the correlation between the similarity and relatedness tasks.



To use the present word pair sets as a reference standard, the consistency of the scores must be evaluated. The interrater agreements on the word pair sets were in an acceptable range (ICC=0.53 and 0.47, in the relatedness and similarity tasks, respectively, *P*<.001). Previously, word pairs in the UMNSRS sets showed different patterns of disagreement based on their domains [7]. The word pairs from the same domain (Drug-Drug) showed higher ICC compared to the other domain categories in the UMNSRS sets. Similarly, the disagreements related to scoring on the word pair sets were not uniformly distributed in this study. In the similarity task, interrater agreement of the word pair sets consisting of the same domain was higher than that of the word pair sets consisting of the different domains. However, in the relatedness task, word pair sets consisting of the different domains had higher interrater agreement (Table 3). To qualify the reliability and consistency of the scores, the word pairs that were scored by more than half of the participants and the SDs of the scores higher than any values above 1.5 × interquartile ranges were removed from the original word pair sets. After removing 3 word pairs in the similarity task and 8 word pairs in the relatedness task, 604 word pairs in the similarity task and 599 word pairs in the relatedness task were included in the final standard reference Korean medical word pair set (Multimedia Appendix 1 and Multimedia Appendix 2). There was no difference in the interrater agreements between the word pair sets in the final word pair sets and the original word pair sets. The distribution of scores was also similar between the original and final word pair sets in both tasks (Kolmogorov-Smirnov statistic for both tasks, *P*>.90).

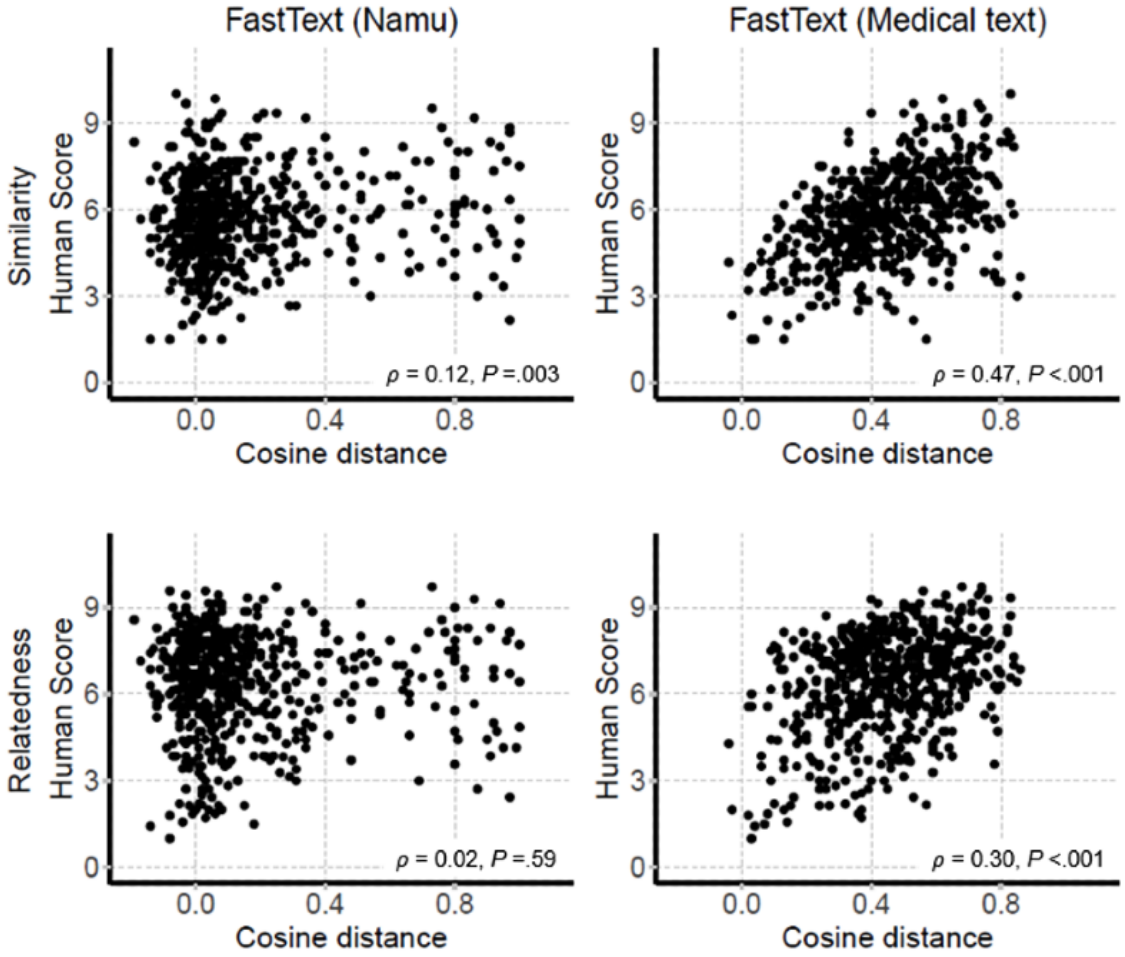
Table 3. Interrater agreement (using the intraclass correlation coefficient) on word pair-sets grouped by the semantic domain types.

Task	Word pairs of the same domain	Word pairs of different domains
Similarity task		
Original word pair set	0.49[a]	0.41[b]
Final word pair set after modification	0.49[c]	0.42[d]
Relatedness task		
Original word pair set	0.52[a]	0.57[b]
Final word pair set after modification	0.51[e]	0.57[f]

[a]n=409.
[b]n=198.
[c]n=408.
[d]n=196.
[e]n=407.
[f]n=192.

To explore the application and efficacy of the final word pair sets, they were applied to the 2 most used unsupervised word embedding models (Word2Vec and FastText). All the word pairs were successfully retrieved by the FastText trained with namu and medical text, and 13 (13/607, 2.1%) word pairs with namu and 145 (145/607, 23.9%) word pairs with medical text were successfully retrieved by Word2Vec. Because more than half of the word pairs were excluded from the Word2Vec model, we only focused on the FastText model. The correlations between cosine distance and human evaluation were higher in the models trained using medical text than in those trained using namu (Figure 3). These results suggest that collection and application of Korean medical texts are key to the development of NLP technology in the Korean medical field.

Figure 3. Correlation between cosine distance from FastText and the human evaluations from 13 attending physicians.



Discussion

In this study, a standard Korean medical word pair set is proposed, with reference values for the similarity and relatedness of word pairs. The novel contributions of this study are as follows: (1) The proposed word pair set used contemporary medical words from various resources, such as textbooks, academic journals, news, and social media; (2) the similarity and relatedness were assessed by attending physicians with at least 5 years of experience in various fields; (3) the developed set is the Korean version and also the first non-English standard reference dataset for semantic similarity and relatedness; (4) when this set was tested for different embedding models, it was found that the more medical-specific the embedding models, the higher the similarity scores provided by the physicians. Therefore, the present word pair set can be successfully applied to evaluate how well the embedding models can represent the medical concepts.

Comparison With Prior Work

Several reference standards exist for estimating semantic relatedness. MayoSRS and MiniMayoSRS consist of 101 and 29 clinical word pairs, respectively, whose relatedness was estimated by 9 medical coders and 3 physicians [8,10]. UMNSRS-Similarity and UMNSRS-related datasets, which were developed in 2010, consist of 566 and 587 word pairs of unified medical language system (UMLS) concepts [7], respectively. Their corresponding similarity and relatedness scores were manually assessed by 8 medical residents. Several studies have adopted these datasets [16,17]. The other reference standard involves the random selection of word pairs from the standardized Medical Dictionary for Regulatory Activities queries [18]. Semantic similarity and relatedness were automatically calculated based on the previously proposed statistical formulas (eg, Resnik [19] and Lin [20]) and the probability sources from the Adverse Event Reporting System database. UMLS and MedDRA are standardized medical vocabularies with organized code systems. However, they are limited because they do not reflect the terms used in the real world. The text sources of this study included textbooks, academic journals, news, and social media articles. Moreover, medical terms that appeared frequently were selected to better reflect the reality of actual use in the present era. In addition, the word pair set of this study include newer terms, such as “Middle East Respiratory Syndrome (MERS),” “Apixaban,” and “Keytruda.”

Korean Translation Version of UMNSRS Word Pair Sets

Before a reference standard word pair set for a local language of Korean was proposed in this study, we considered using a word pair set of an English reference standard, such as the UMNSRS data set, translated into a Korean version. In a preliminary study with 12 health information managers certified by the Korean government (similar to the registered health information administrators or technicians in the United States), the average answer rate for the UMNSRS word pair sets translated into Korean was only 66.7% for the similarity set and 66.5% for the relatedness set. The average answer rates of the same participants for the present word pair sets were increased to 81.5% for the similarity set and 82.1% for the relatedness set. There could be 2 underlying reasons for the difference in response rates: the translation task and the medical environment.

First, the translation process inevitably causes some data loss, even across Indo-European languages. In a French study, the UMNSRS word pair sets were translated into French, and only 73% of the similarity set and 71% of the relatedness set were translated and used [21]. In a comparable Spanish study, only 65% of the relatedness set and 67% of the similarity set were automatically rendered into Spanish because of regional differences in medical protocols and commercial drug names [22]. These reports imply that reference standards need to be developed specifically for each language.

Second, differences in cultural and medical backgrounds may influence the response rate. The correlations between the scores given for the Korean translation and for the original UMNSRS word-sets were modest (ρ=0.71 for the similarity and 0.71 for relatedness; Figure 4). When the embedding models in the present study were applied to the Korean translation version of UMNSRS word pair sets, the correlations between the cosine distance of the embedding models and the human evaluation scores of 12 health information managers on Korean-translated UMNSRS word pair sets were similar for the similarity task and slightly higher for the relatedness task compared to the correlations between the cosine distance of the embedding models and the scores of 13 attending physicians on the reference standard word pair sets from the present study (Figure 5). Notably, human evaluations were performed in different groups (health information managers and attending physicians). In word pair scoring, experienced physicians could be more affected by the complexity and various possibilities of medicine. This can be attributed to the relatively inconsistent results for the tasks, particularly the relatedness task. Furthermore, the semantic relation of word pairs in the present study could be more complex and specialized compared to those of UMNSRS word pair sets. This is bolstered by the fact that the correlation between the cosine distance of the embedding model trained with only namu wiki and the scores of physicians on the reference standard word pair sets of the present study was extremely low for the relatedness task (ρ=0.02). It is also noteworthy that the word embedding model trained with the medical texts shows better performance on both the word pair sets (Korean translated UMNSRS word pair sets and the word pair sets of the present study). The results thus obtained suggest that it is important to secure large medical texts for better performance in word embedding techniques.

Figure 4. Correlation between the reference scores for the original University of Minnesota Semantic Relatedness Set (UMNSRS) word pair sets (English version) and the scores from 12 health information managers for the Korean translation version.

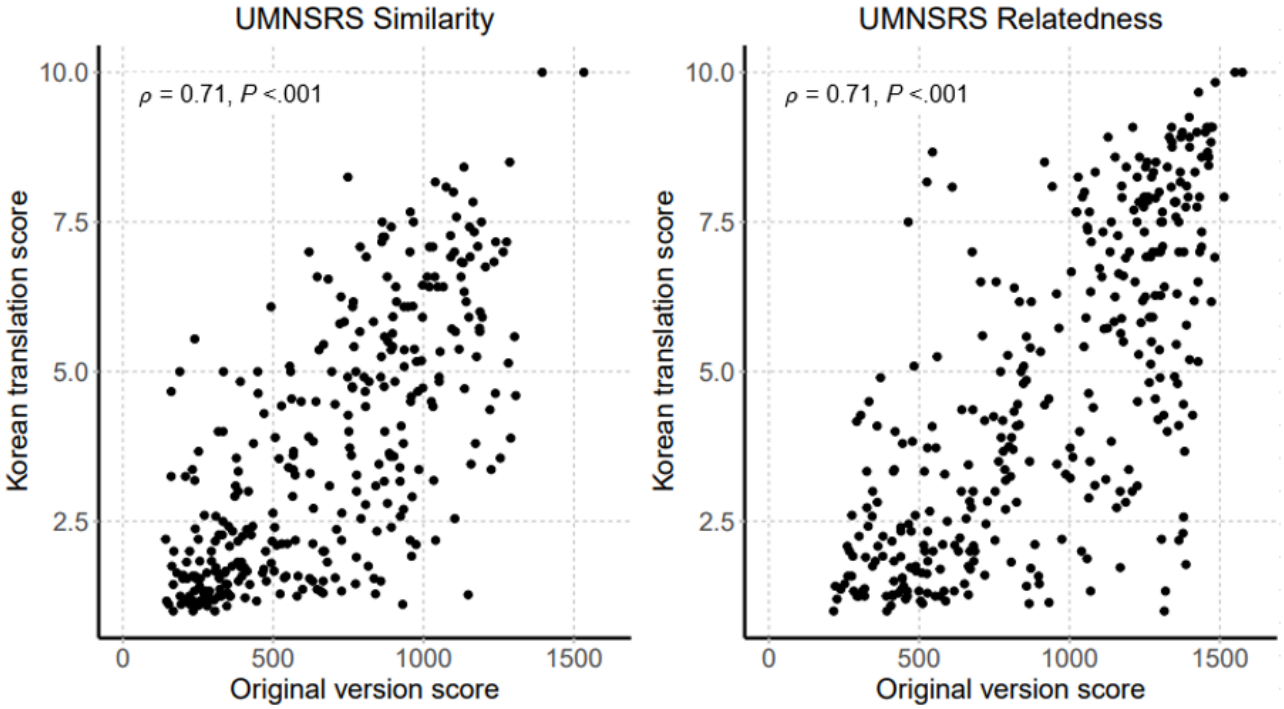
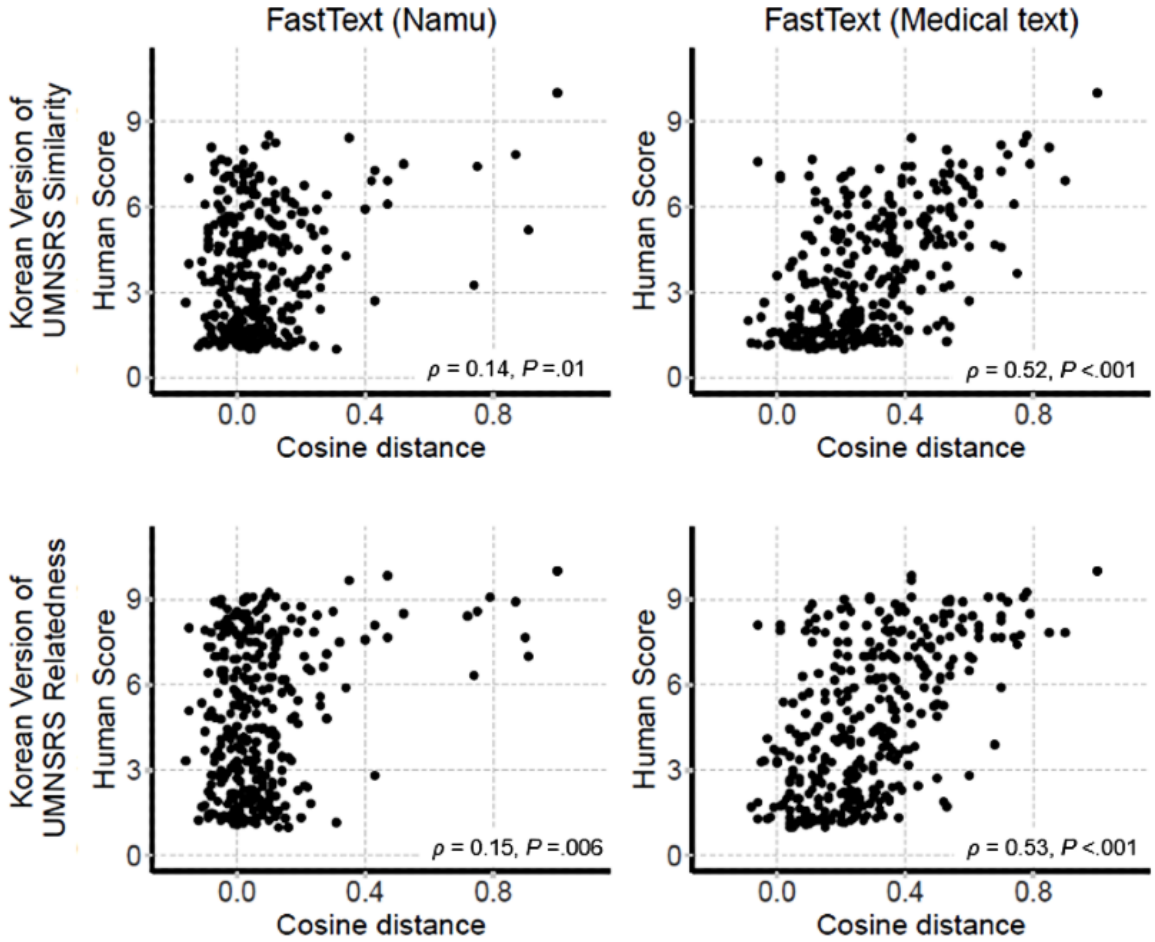


Figure 5. Correlation between the cosine distance of FastText embedding models and the human evaluations by 12 health information managers of the Korean version of the University of Minnesota Semantic Relatedness Set (UMNSRS) word pair sets.



Considered together, it is expected that the Korean word pair sets proposed in the present study can be used as one of the global reference standards through appropriate translation along with the UMNSRS word pair sets. We acknowledge that the UMNSRS word pair sets have the potential to become a global reference standard though appropriate translation.

Multilingual NLP

Recently, a variety of NLP technologies for diverse human languages have been studied. However, there are only a few studies on multilingual NLP in the medical field [23-25]. Some studies built their own corpus and compared the performance of their embedding models [25,26]. One study adopted a metathesaurus (eg, UMLS) that included multiple languages [25]. Using a reference standard for each language to evaluate multilingual NLP models is not preferred because of the following 2 reasons. First, because the translation task itself is neither straightforward nor bias-free, we cannot necessarily rely on translated data. Second, to the best of our knowledge, no word pair dataset has yet been built for use as a non-English reference standard.

An ideal reference standard pertains to the underlying linguistic structures of an individual language, as well as the semantics of medical concepts. It has been noted that recent multilingual NLP architectures involve some elementary cross-linguistic knowledge [26,27]. However, linguistically naive NLP models often do not consider the typological differences in different languages (eg, morphological features) because they are overfitted to a specific type of language variation. Thus, linguistically naive multilingual models fail even in a word-by-word evaluation [28]. Furthermore, a multilingual NLP system is prone to biases because large English corpora are heavily weighted against low-resource languages, such as Korean corpora. This data scarcity problem undermines the reliability of word embeddings trained on medical texts and causes the model to perform inconsistently across different languages.

Limitations

The present study has several limitations. First, the most noticeable discrepancy between human cognition and embedding models is based on semantics. For example, 이부프로펜 (ibuprofen) is one of the most popular nonsteroidal anti-inflammatory drugs. The participants in our study could easily recognize the word 발열 (fever) and 이부프로펜 (ibuprofen). However, this drug was not considered when building our corpus owing to the lack of sufficient reasons. Instead, the word 이부프로펜 (ibuprofen) was usually identified by its effects, such as 해열 (anti-fever) and 소염 (anti-inflammation). As a result, the cosine similarities between 이부프로펜 (ibuprofen)-발열 (fever) was 0.18, which is considerably lower than 이부프로펜 (ibuprofen)-해열 (anti-fever), at 0.53, and 이부프로펜 (ibuprofen)-소염 (anti-inflammation), at 0.61. This implies that word embedding models are not good at finding similarities between words with low distributional and high lexical analogy. Second, capturing intuitive judgments by medical experts was not directly correlated to the automatic prediction of medical labels, such as the names of treatments and drugs. Thus, the Korean version of similarity and relatedness data in this study regards that word embeddings represent human-like knowledge of Korean medical terms, whereas they are general predictors of performance in real-world applications. This distinction between intrinsic and extrinsic evaluation also corresponds to the standard methodology for assessing unsupervised word embeddings trained on general domain corpora [29]. Third, there was bias in the preliminary results from the Korean translation of UMNSRS word sets presented in the Discussion section. The medical specialty and domain knowledge of the participating group between the Korean translation and the original version were not met. Further, no experiment was conducted involving multilingual participants. Although these results are insufficient to reveal the limitations of translation, and since the current translation technology is rapidly developing, limitations related to translation in specialized fields, such as medical terms, still exist. In this respect, we believe that developing reference standards for various languages can help accelerate the development of medical NLP technologies in the future.

Conclusions

The word pair reference standard is an important tool for evaluating the performance of NLP techniques, such as word embedding models. It is difficult to evaluate the semantic relevance of medical terms, as such evaluations require knowledge and expertise. In this study, 604 word pairs were proposed for similarity evaluation and 599 word pairs for relatedness evaluation as Korean reference standard word pair sets for use in the medical domain. This study is the first step toward the development of a word pair reference standard using various resources, including textbooks, news, and academic journals published in a non-English language, Korean, and is expected to facilitate the further acceleration of medical NLP techniques for different languages.

Data Availability

The final word pair sets for the similarity and the relatedness are available at https://github.com/KU-RIAS/.

Acknowledgments

The authors would like to thank the students who completed the biodata engineering training program for their assistance. This research was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant numbers HI19C0201 and HI19C0832).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Medical word pair set (similarity, Korean).
[XLSX File (Microsoft Excel File), 38 KB - medinform_v9i6e29667_app1.xlsx]

Multimedia Appendix 2
Medical word pair set (relatedness, Korean).
[XLSX File (Microsoft Excel File), 37 KB - medinform_v9i6e29667_app2.xlsx]

References

1. Safi Z, Abd-Alrazaq A, Khalifa M, Househ M. Technical Aspects of Developing Chatbots for Medical Applications: Scoping Review. J Med Internet Res 2020 Dec 18;22(12):e19127 [FREE Full text] [doi: 10.2196/19127] [Medline: 33337337]
2. Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. Sci Data 2019 May 10;6(1):52 [FREE Full text] [doi: 10.1038/s41597-019-0055-0] [Medline: 31076572]
3. Chen Q, Lee K, Yan S, Kim S, Wei C, Lu Z. BioConceptVec: Creating and evaluating literature-based biomedical concept embeddings on a large scale. PLoS Comput Biol 2020 Apr;16(4):e1007617 [FREE Full text] [doi: 10.1371/journal.pcbi.1007617] [Medline: 32324731]
4. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
5. Wajsbürt P, Sarfati A, Tannier X. Medical concept normalization in French using multilingual terminologies and contextual embeddings. J Biomed Inform 2021 Feb;114:103684. [doi: 10.1016/j.jbi.2021.103684] [Medline: 33450387]
6. Grabar N, Grouin C, Section Editors for the IMIA Yearbook Section on Natural Language Processing. A Year of Papers Using Biomedical Texts: Findings from the Section on Natural Language Processing of the IMIA Yearbook. Yearb Med Inform 2019 Aug;28(1):218-222 [FREE Full text] [doi: 10.1055/s-0039-1677937] [Medline: 31419835]
7. Pakhomov S, McInnes B, Adam T, Liu Y, Pedersen T, Melton GB. Semantic Similarity and Relatedness between Clinical Terms: An Experimental Study. AMIA Annu Symp Proc 2010 Nov 13;2010:572-576 [FREE Full text] [Medline: 21347043]
8. Pakhomov SVS, Pedersen T, McInnes B, Melton GB, Ruggieri A, Chute CG. Towards a framework for developing semantic relatedness reference standards. J Biomed Inform 2011 Apr;44(2):251-265 [FREE Full text] [doi: 10.1016/j.jbi.2010.10.004] [Medline: 21044697]
9. Mathôt S, Schreij D, Theeuwes J. OpenSesame: an open-source, graphical experiment builder for the social sciences. Behav Res Methods 2012 Jun;44(2):314-324 [FREE Full text] [doi: 10.3758/s13428-011-0168-7] [Medline: 22083660]
10. Pedersen T, Pakhomov SVS, Patwardhan S, Chute CG. Measures of semantic similarity and relatedness in the biomedical domain. J Biomed Inform 2007 Jun;40(3):288-299 [FREE Full text] [doi: 10.1016/j.jbi.2006.06.004] [Medline: 16875881]
11. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed Representations of Words and Phrases and their Compositionality. Cornell University. 2013. URL: https://arxiv.org/abs/1310.4546 [accessed 2021-06-10]
12. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching Word Vectors with Subword Information. Cornell University. 2017. URL: https://arxiv.org/abs/1607.04606 [accessed 2021-06-10]
13. Eunjeonhan project. URL: http://eunjeon.blogspot.com/ [accessed 2021-06-10]
14. Rehurek R, Sojka P. Software Framework for Topic Modelling with Large Corpora. 2010 Presented at: Workshop On New Challenges For NLP Frameworks 2010; May 22, 2010; Valletta, Malta.
15. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979 Mar;86(2):420-428. [doi: 10.1037//0033-2909.86.2.420] [Medline: 18839484]
16. Jiang S, Wu W, Tomita N, Ganoe C, Hassanpour S. Multi-Ontology Refined Embeddings (MORE): A hybrid multi-ontology and corpus-based semantic representation model for biomedical concepts. J Biomed Inform 2020 Nov;111:103581. [doi: 10.1016/j.jbi.2020.103581] [Medline: 33010425]
17. Pakhomov SVS, Finley G, McEwan R, Wang Y, Melton GB. Corpus domain effects on distributional semantic modeling of medical terms. Bioinformatics 2016 Dec 01;32(23):3635-3644 [FREE Full text] [doi: 10.1093/bioinformatics/btw529] [Medline: 27531100]
18. Bill RW, Liu Y, McInnes BT, Melton GB, Pedersen T, Pakhomov S. Evaluating semantic relatedness and similarity measures with Standardized MedDRA Queries. AMIA Annu Symp Proc 2012;2012:43-50 [FREE Full text] [Medline: 23304271]
19. Resnik P. Using information content to evaluate semantic similarity in a taxonomy. 1995 Presented at: 14th International Joint Conference on Artificial Intelligence (IJCAI); August 20-25,1995; Montreal, Quebec, Canada URL: https://arxiv.org/abs/cmp-lg/9511007
20. Lin D. An Information-Theoretic Definition of Similarity. 1998 Presented at: 15th International Conference on Machine Learning; July 24-27, 1998; Madison, WI URL: http://www.mathcs.emory.edu/~choi/courses/reading/lin-98a.pdf
21. Dynomant E, Lelong R, Dahamna B, Massonnaud C, Kerdelhué G, Grosjean J, et al. Word Embedding for the French Natural Language in Health Care: Comparative Study. JMIR Med Inform 2019 Jul 29;7(3):e12310 [FREE Full text] [doi: 10.2196/12310] [Medline: 31359873]

22. Soares F, Villegas M, Gonzalez-Agirre A, Krallinger M, Armengol-Estapé J. Medical word embeddings for Spanish: development and evaluation. 2019 Presented at: 2nd Clinical Natural Language Processing Workshop; June 2019; Minneapolis, MN. [doi: 10.18653/v1/w19-1916]
23. Huang EW, Wang S, Lee DJ, Zhang R, Liu B, Zhou X, et al. Framing Electronic Medical Records as Polylingual Documents in Query Expansion. AMIA Annu Symp Proc 2017;2017:940-949 [FREE Full text] [Medline: 29854161]
24. Wajsbürt P, Sarfati A, Tannier X. Medical concept normalization in French using multilingual terminologies and contextual embeddings. J Biomed Inform 2021 Feb;114:103684. [doi: 10.1016/j.jbi.2021.103684] [Medline: 33450387]
25. Zhang ZC, Zhang MY, Zhou T, Qiu YL. Pre-trained language model augmented adversarial training network for Chinese clinical event detection. Math Biosci Eng 2020 Mar 24;17(4):2825-2841 [FREE Full text] [doi: 10.3934/mbe.2020157] [Medline: 32987500]
26. Pires T, Schlinger E, Garrette D. How Multilingual is Multilingual BERT? 2019 Presented at: 57th Annual Meeting of the Association for Computational Linguistics; July 2019; Florence, Italy. [doi: 10.18653/v1/p19-1493]
27. Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, et al. Unsupervised Cross-lingual Representation Learning at Scale. 2020 Presented at: 58th Annual Meeting of the Association for Computational Linguistics; July 5-10, 2020; Online. [doi: 10.18653/v1/2020.acl-main.747]
28. Bender EM. Linguistically Naïve != Language Independent: Why NLP Needs Linguistic Typology. 2009 Presented at: EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?; March 2009; Athens, Greece URL: https://www.aclweb.org/anthology/W09-0106.pdf [doi: 10.3115/1642038.1642044]
29. Schnabel T, Labutov I, Mimno D, Joachims T. Evaluation methods for unsupervised word embeddings. 2015 Presented at: Conference on Empirical Methods in Natural Language Processing; September 2015; Lisbon, Portugal. [doi: 10.18653/v1/d15-1036]

Abbreviations

ICC: intraclass correlation coefficient
KHIDI: Korea Health Industry Development Institute
KISS: Korean Studies Information Service System
KSS: Korean Sentence Splitter
MERS: Middle East Respiratory Syndrome
MeSH: Medical Subject Headings
NLP: natural language processing
UMLS: unified medical language system
UMNSRS: University of Minnesota Semantic Relatedness Set

Edited by G Eysenbach; submitted 16.04.21; peer-reviewed by M Rodrigues; comments to author 07.05.21; revised version received 08.05.21; accepted 16.05.21; published 24.06.21.

Please cite as:
Yum Y, Lee JM, Jang MJ, Kim Y, Kim JH, Kim S, Shin U, Song S, Joo HJ
A Word Pair Dataset for Semantic Similarity and Relatedness in Korean Medical Vocabulary: Reference Development and Validation
JMIR Med Inform 2021;9(6):e29667
URL: https://medinform.jmir.org/2021/6/e29667/
doi:10.2196/29667
PMID:34185005

Original Paper

A National, Semantic-Driven, Three-Pillar Strategy to Enable Health Data Secondary Usage Interoperability for Research Within the Swiss Personalized Health Network: Methodological Study

Christophe Gaudet-Blavignac[1,2], BSc, MSc; Jean Louis Raisaro[3,4], PhD; Vasundra Touré[5], PhD; Sabine Österle[5], PhD; Katrin Crameri[5], MPH, PhD; Christian Lovis[1,2], MD, MPH, FACMI

[1]Division of Medical Information Sciences, Geneva University Hospitals, Geneva, Switzerland

[2]Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland

[3]Data Science Group, Division of Information Systems, Lausanne University Hospital, Lausanne, Switzerland

[4]Precision Medicine Unit, Department of Laboratories, Lausanne University Hospital, Lausanne, Switzerland

[5]Personalized Health Informatics Group, SIB Swiss Institute of Bioinformatics, Basel, Switzerland

Corresponding Author:
Christophe Gaudet-Blavignac, BSc, MSc
Division of Medical Information Sciences
Geneva University Hospitals
Rue Gabrielle-Perret-Gentil 4
Geneva, 1205
Switzerland
Phone: 41 223726201
Email: christophe.gaudet-blavignac@hcuge.ch

Abstract

Background: Interoperability is a well-known challenge in medical informatics. Current trends in interoperability have moved from a data model technocentric approach to sustainable semantics, formal descriptive languages, and processes. Despite many initiatives and investments for decades, the interoperability challenge remains crucial. The need for data sharing for most purposes ranging from patient care to secondary uses, such as public health, research, and quality assessment, faces unmet problems.

Objective: This work was performed in the context of a large Swiss Federal initiative aiming at building a national infrastructure for reusing consented data acquired in the health care and research system to enable research in the field of personalized medicine in Switzerland. The initiative is the Swiss Personalized Health Network (SPHN). This initiative is providing funding to foster use and exchange of health-related data for research. As part of the initiative, a national strategy to enable a semantically interoperable clinical data landscape was developed and implemented.

Methods: A deep analysis of various approaches to address interoperability was performed at the start, including large frameworks in health care, such as Health Level Seven (HL7) and Integrating Healthcare Enterprise (IHE), and in several domains, such as regulatory agencies (eg, Clinical Data Interchange Standards Consortium [CDISC]) and research communities (eg, Observational Medical Outcome Partnership [OMOP]), to identify bottlenecks and assess sustainability. Based on this research, a strategy composed of three pillars was designed. It has strong multidimensional semantics, descriptive formal language for exchanges, and as many data models as needed to comply with the needs of various communities.

Results: This strategy has been implemented stepwise in Switzerland since the middle of 2019 and has been adopted by all university hospitals and high research organizations. The initiative is coordinated by a central organization, the SPHN Data Coordination Center of the SIB Swiss Institute of Bioinformatics. The semantics is mapped by domain experts on various existing standards, such as Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), Logical Observation Identifiers Names and Codes (LOINC), and International Classification of Diseases (ICD). The resource description framework (RDF) is used for storing and transporting data, and to integrate information from different sources and standards. Data transformers based on SPARQL query language are implemented to convert RDF representations to the numerous data models required by the research community or bridge with other systems, such as electronic case report forms.

Conclusions: The SPHN strategy successfully implemented existing standards in a pragmatic and applicable way. It did not try to build any new standards but used existing ones in a nondogmatic way. It has now been funded for another 4 years, bringing

the Swiss landscape into a new dimension to support research in the field of personalized medicine and large interoperable clinical data.

(JMIR Med Inform 2021;9(6):e27591) doi:10.2196/27591

KEYWORDS

interoperability; clinical data reuse; personalized medicine

Introduction

Background

Interoperability is a well-known challenge in medical informatics and is one of the main obstacles preventing data-driven medicine from realizing its full potential. Efforts to classify and express meaning in health care are as old as the International Classification of Diseases (ICD) [1]. Organizations, such as Health Level Seven International (established in 1987) [2] and SNOMED International, which maintains and releases the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [3], are dedicated to promoting interoperability in health care. Moreover, multiple national and international programs are seeking to promote interoperability. Examples of major initiatives designed to tackle the interoperability challenge in health care include the Meaningful Use program under the Health Information Technology for Economic and Clinical Health Act [4] in the United States and the Integrating the Healthcare Enterprise initiative [5], with more than 175 member organizations worldwide.

Semantic Interoperability

Semantic interoperability usually involves controlled vocabularies. In the medical field, the equivalent is part of the culture, but is named differently, such as scales, scores, and classifications. These involve organization of medical knowledge into a finite set of classes. They are used in daily practice to evaluate, describe, and prognose many situations or conditions. For example, medical scales or scores are narrow-scope classifications used in everyday medical practice. The Glasgow Coma Scale [6] and the Apgar score [7] are examples to describe the level of consciousness of patients and the health of newborns, respectively. In some cases, there are several of them for a specific condition with different perspectives, such as for heart failure [8]. Clinicians commonly use dozens of scores and scales in their daily practice, and there are numerous applications that combine and facilitate their use [9,10].

More extensive medical classifications, such as the 10th revision of the International Classification of Diseases (ICD-10) [11] and the Logical Observation Identifiers Names and Codes (LOINC) [12], are large systems that attempt to organize broader areas of medical knowledge, such as diseases and causes of death (ICD-10), or health measurements, observations, and documents (LOINC).

They can be articulated into larger representations (meta-organizations) that consolidate several classifications, ontologies, terminologies, etc. The Unified Medical Language System (UMLS) Metathesaurus [13,14], for example, combines several classifications having different purposes, such as diagnosis encoding and literature indexing. SNOMED CT is another example, which combines 19 top-level hierarchies into one representation.

Specific classifications are characterized by a partitioning of the knowledge represented according to a specific purpose, usually the intention for which the classification has been designed. Thus, SNOMED CT is historically dedicated to pathology and was extended later with clinical codes. ICD-9 and 10 are well adapted to represent diagnosis and morbidity causes, while LOINC is mostly used to represent laboratory analytical and preanalytical characteristics. Drugs are often handled using Global Trade Item Number (GTIN) for logistical needs and Anatomical Therapeutic Chemical (ATC) classifications for order entry decision support [15,16], while adverse drug reactions are reported in MedDRA [17].

Challenges for Semantic Interoperability

As a result of having specific classifications well designed for specific purposes, they are usually not well adapted to express other types of knowledge or different organizations (partitioning) of that knowledge. SNOMED CT is able to represent almost any pathological test result and has been used to represent free text, but it fails to express some types of concrete values [18,19]. ICD-10 can be used to assign a code to any disease, but its mono-hierarchical structure prevents meaningful information reuse (eg, it is not possible to easily extract all codes representing infectious diseases). Finally, GTIN identifies commercial drug products, but it does not efficiently represent active substances, while ATC expresses only substances, but not the products. Classifications are tools used to represent the meaning of the data, but they always carry an intent, and none can be used for every purpose.

Data Organization

Data are usually organized with data models, and the first and most simple is the text or tabular file that is still widely used, notably in clinical research settings. The serialization of data in comma-separated value (CSV) files can be expanded into more complex representations. Data models structure the data into entities and relationships that fit a given purpose. These have existed in health care for a long time, and some of them are widely used. Health Level 7 (HL7) version 2, which is the most widely implemented standard for health care in the world [20], is linked to the Reference Information Model (RIM), a data model designed to be the backbone of HL7, with the following three main classes: Act (representing something that has happened or will happen), Entity (any living or nonliving thing), and Role (a competency expressed by an Entity). These three classes will then be used to build an event using a "connector" named "Participation" that allows building of complex nested structures [21]. Finally, as for controlled

vocabularies, data models can be articulated in meta-models, such as the bridge recently created between the Observational Medical Outcomes Partnership (OMOP) and the Informatics for Integrating Biology and the Bedside (i2b2) [22] data models.

Challenges of Data Interoperability

The structure of each data model depends on the goal of the standard and on the community that will use the data. For example, the RIM was primarily targeted at electronic health record (EHR) interoperability, while the Common Data Model of OMOP specifically targeted clinical research [23]. The data model of i2b2 [24] is designed to integrate genetic and phenotypic data, while the Clinical Data Interchange Standards Consortium (CDISC) operational data model [25] is required for drug regulatory constraints by the United States Food and Drug Administration. The openEHR project is built around another paradigm and is composed of archetypes that are small domain models aimed at providing a specific piece of information. The definition of archetypes and templates of archetypes are very flexible and can solve numerous interoperability challenges; however, it still requires adopting the reference model for the storage of data [26,27]. The design of these models is based on specific goals, and there is no one-size-fits-all data model that can serve every purpose.

The Swiss Personalized Health Network

The Swiss Personalized Health Network (SPHN) aims to leverage research in the field of personalized health in Switzerland by building a nationally coordinated infrastructure network that supports exchange and reuse of health-related data produced by the health care system and in biomedical and clinical research settings [28]. This national initiative was launched in 2017, with funding of up to CHF 137 million (US $153 million) assured until 2024 [29,30]. In essence, the goal of the SPHN is to connect the Swiss health care system, the research community, regulatory agencies, and eventually industrial partners involved in personalized medicine. Consequently, the SPHN is at the interface between three communities and must overcome the multiple challenges of exchanging data in a secure, interoperable, and meaningful manner.

Objectives

The challenges of interoperability described above have been the focus of active research in recent decades. Every year, new standards appear with the goal of addressing the remaining challenges. Interestingly, each of these new standards solves some problems but also generates new ones.

As opposed to conventional approaches, which are aimed at mapping data to one common standard and are in practice only effective for specific use cases, our interoperability strategy uses existing standards in a purpose-specific and complementary manner without depending on any particular one, thus providing great flexibility and sustainability. As such, it enables data interoperability between various communities, each of which has different needs or follows different requirements with regard to the type of data model to be used.

Vocabulary

Interoperability is by essence an interdisciplinary process. Therefore, the vocabulary used to describe its components can vary. This section aims to define the words used in this work and their meaning. *Data model* is an abstract model that organizes elements of data in structures. *Data model-independent* is used to describe a system that does not depend on a predefined data model. *Encoding* is the action of expressing something with a specific coding system. For example, encoding a concept into a terminology means linking this concept to the elements of the targeted terminology that adequately represent it. *Interoperability* is the ability of two different entities to connect, share, understand, and use data in their processes. *Semantics* is the encoding of meaning into one or more knowledge representations (KRs). *Knowledge representation* is organization of knowledge into a list of elements, such as controlled vocabularies, terminologies, classifications, taxonomies, ontologies, thesauri, and coding systems.

Methods

Overview

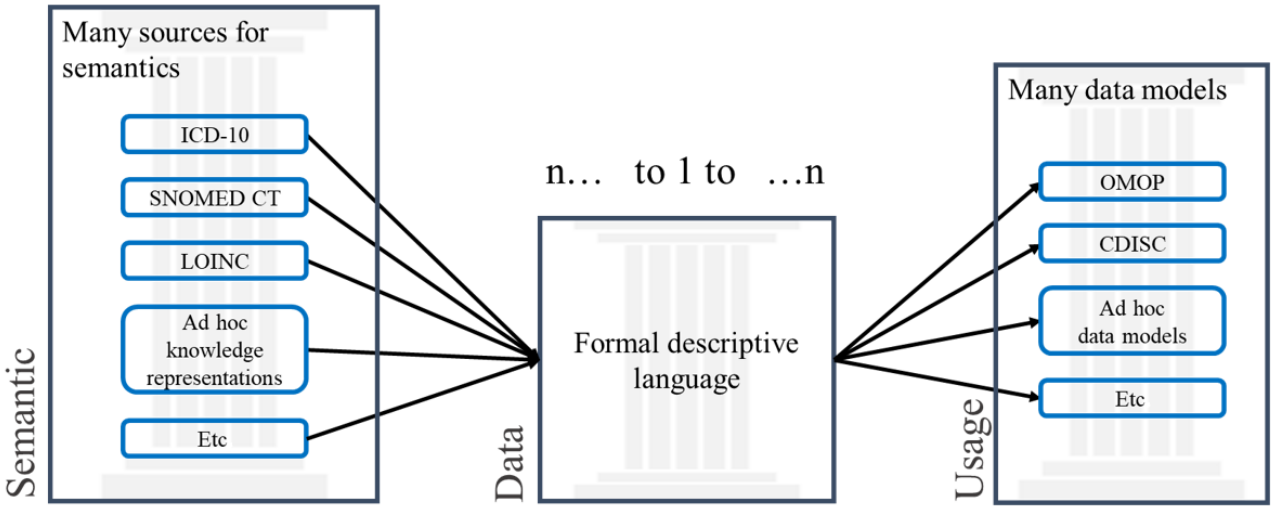
Based on the lessons learned from previous attempts, this work addresses the interoperability challenge adopting a semantic-driven data model–independent framework based on the following three pillars (Figure 1):

1. A multidimensional encoding of the concepts. Only the required concepts (variables) are encoded in any KR system. This decision is completely agnostic, so that several international standards can be used at any time, according to the needs.
2. Resource description framework (RDF)-based storage and transport of the instances of these concepts when used to express clinical data. The RDF is well suited for a federated national exchange format. As it is a formal descriptive language, it is very scalable to any future needs not yet known.
3. Conversion of the RDF to any target data model that is needed for a specific research community or usage, according to the needs of the users.

This ends up with the first two pillars being completely data model independent. Only at the third pillar will the data be available in any required model, such as CDISC and OMOP, according to needs. We thus considered this strategy “semantic agnostic” and “model independent.”

This strategy is being implemented stepwise since January 2019. This paper focuses on the strategy. The deployment and societal challenges will be discussed in a further publication.

Figure 1. The three pillars of the proposed data interoperability strategy. CDISC: Clinical Data Interchange Standards Consortium; Etc: et cetera; ICD-10: International Classification of Diseases; LOINC: Logical Observation Identifiers Names and Codes; OMOP: Observational Medical Outcomes Partnership; SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms.



Integrative and Usability-Focused Semantic Approach

As stated in the Introduction, it is illusory to believe that different communities will adopt a single standard for the sake of mutual compatibility. Therefore, this strategy does not enforce a specific KR to express meaning. The goal is to enable the use of an adequate KR, based on the purpose and context of use, without imposing any specific one. However, the presence of a semantic definition of the data is crucial and must be the central axis of the strategy.

The first pillar of our approach consists, therefore, of developing a semantic framework comprising a set of concept definitions relying on existing KRs or new ones if needed. The concept definition must be adapted to the granularity required by the use case. Each concept can be encoded into as many KRs as required. For instance, the concept “Heart Rate” can include encoding into SNOMED CT and LOINC. The power of representation and usability is prioritized over conceptualization. It is thus possible to express the meaning of the data without enforcing a specific KR. Finally, instantiations of the concepts can use an adequate KR, depending on the context. Axioms of the first pillar are summarized in Textbox 1.

Textbox 1. Axioms of the first pillar of the strategy.

Axioms

- Framework composed of a set of concept definitions.
- Semantic encoding using a knowledge representation.
- Multiencoding of a concept in several knowledge representations allowed.
- Selection of concepts defined by use cases.
- Combination and extension of concepts allowed.

Descriptive Formalism for Transfer and Storage

Transport and storage of information are essentially the same. Since transport is a “moving storage” and storage is a “nonmoving transport,” they can be regarded as a single challenge. The data and concept landscapes in health care are constantly evolving with new elements to be exchanged. To best answer this need for sustainability, scalability, and plasticity, the strategy is based on the use of a descriptive formalism (eg, the RDF, the Arden syntax, and the Web Ontology Language [OWL]) [31-33]. These languages offer flexible storage and transport of information (be it data, semantics, processes, or rules). This differs from a data model–based approach, as it does not constrain data to fit a specific format but only describes the data and its semantics in an intuitive and unconstrained way as it is collected at the source. Our approach allows the use of different formalisms when needed. For example, RDF can be used to store and transport the data, and the Arden syntax can be used to describe rules, such as automatic alert and clinical decision support systems [32]. Similarly, other formalisms can be used for other types of information and purposes (eg, Guidelines Interchange Format for guidelines [34] and Java Business Process Model for workflows [35]). Textbox 2 summarizes the approach for the second pillar.

Textbox 2. Axioms of the second pillar of the strategy.

Axioms

- Common approach for storage and transport.
- No a priori definition of a data model.
- Use of descriptive formalisms to describe the data encoded in the first pillar.
- Choice of the formalism depending on the use case.

Purpose-Specific Transformation to Data Models

The final building block of our strategy is the transformation of data from a flexible representation, based on formal descriptive languages, to a more rigid but application-oriented representation, such as relational data models. The goal is to provide a way of efficiently sharing data between different communities used to working with their own data models. As mentioned above, no common data model can be adopted by all communities, and mappings across data models are often partial because of incompatible information representations.

As a result of the first and second pillars, it is possible to create ad hoc conversions based on users' needs. In particular, the use of a data model–independent formalism to store data enables the implementation of one-to-many mappings to any target data model. For example, existing work has already proposed the transformation of RDF resources into customized relational data models [36] or standard common data models, such as i2b2 [37,38] and OMOP [39]. This approach addresses the complexity of the current many-to-many mappings and will enable the sharing of data with any community, provided that the mapping is done while keeping the data unchanged. Textbox 3 summarizes the approach.

Textbox 3. Axioms of the third pillar of the strategy.

Axioms

- Ad hoc conversions from the descriptive formalism of the second pillar to data models.
- Building of a reusable one-to-many mapping catalog.
- Selection of the targeted data models based on use cases.

Results

The SPHN

The proposed interoperability strategy was implemented to serve the data-sharing needs of the SPHN. The projects supported are all large multicentric projects, multihospitals, multiresearch centers, and data-driven research related to personalized medicine [28]. They vary in terms of not only methodology and research questions, but also the clinical data concepts requested from the data providers involved. The projects are designed to generalize the use of the Swiss General Consent, improve clinical data management systems on care providers, build a national data interoperability landscape for research, and leverage research organizations.

The defined approach was implemented by every university hospital and high research organization of Switzerland as the national standard for sharing clinical data. Twelve driver projects were funded and used the approach for their data needs.

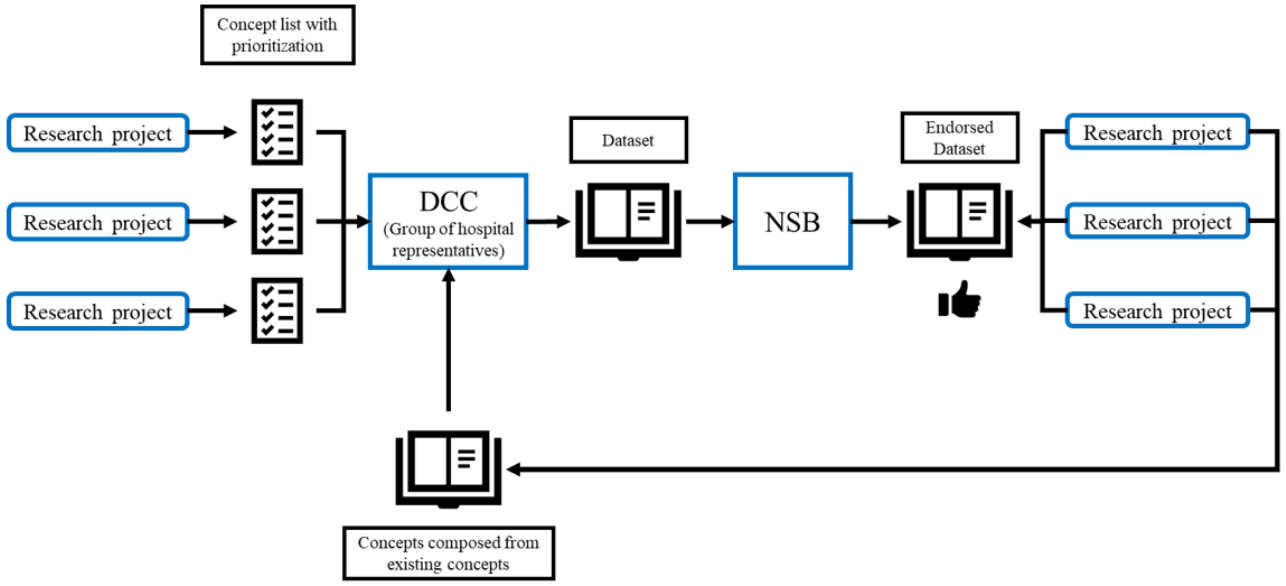
Organization

In the implementation of the first pillar, a semantic framework has been built and maintained by the SPHN Data Coordination Center (DCC). The DCC is the central hub for data interoperability in the SPHN and part of the SIB Swiss Institute for Bioinformatics. Its mandate is to coordinate the development of the specification of the structure and semantics of the SPHN data set, which describes the type of data that is available and potentially shareable within the network (hereafter referred to as the SPHN semantic data set). A full description of the DCC is available on the SPHN website [40].

First Pillar

The content of the SPHN semantic data set is defined by leveraging domain knowledge from the Swiss clinical research community. Every research project provides the list of variables they need to the group in charge of aligning the semantics. This group includes domain experts and clinical semantics specialists. This SPHN semantic data set is periodically reviewed and extended according to experience obtained in projects by extracting and using the data and, of course, the new needs of research projects. There is a validation process that ends up in the publication of a new release of the core list of concepts endorsed by the SPHN National Steering Board (NSB). After official release, the new concepts are used by university hospitals for interoperable data exchange. The steps involved in this process are shown schematically in Figure 2. The complete structure of the SPHN is beyond the scope of this article and openly available in published reports [30].

Figure 2. Flowchart of the validation process. DCC: Data Coordination Center; NSB: National Steering Board.



The concept list is evolving, such that each element contains, in addition to semantics, management metadata, such as unique ID, a name, a description, and several fields for versioning. All data transfer for SPHN projects should comply with these concepts once enforced by the NSB. Examples of the encoding of these concepts with SNOMED CT and LOINC are shown in Table 1, in which the code is linked to the row where relevant and applicable. As more use cases arise, new encodings can be created.

The DCC has the task of exploring common international KR when validating new concepts, so as to select the most appropriate one. A KR for a concept is chosen taking into consideration not only its capacity to represent the concept correctly and unambiguously but also the ability of hospitals to comply with it and the research project to use it. Currently, more than 300 concepts are being used, which can describe demographics, laboratory analysis and results, drugs and prescriptions, clinical and physiological variables, etc [41].

Table 1. Examples of encoded concepts used to describe a temperature measurement.

Concept name	SNOMED CT[a]	LOINC[b]
Temperature	386725007 |Body temperature (observable entity)|	8310-5 Body temperature
Unit	767524001 |Unit of measure (qualifier value)|	N/A[c]
Body site	123037004 |Body structure (body structure)|	39111-0 Body site

[a]SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms.

[b]LOINC: Logical Observation Identifiers Names and Codes.

[c]N/A: not applicable.

Second Pillar

The data storage and transport step of the SPHN was implemented using RDF as proposed by the World Wide Web Consortium [42]. RDF allows to map instances of real data originating from a clinical database with the conceptual framework defined in the first pillar. The RDF allows to build a labeled directed multigraph, where nodes and edges are identified by uniform resource identifiers. The basic entity in the RDF graph is known as a “triple” and is composed of a subject, a predicate, and an object. Several triples compose a graph. Since the RDF does not depend on a specific semantic standard, it allows for the use of different ontologies and value sets, as required by the strategy. The reasons for choosing RDF technologies are summarized in Textbox 4 [43-47].

Textbox 4. Reasons for choosing the resource description framework.

Reasons

- Flexibility to represent complex knowledge with simple statements (ie, triples of information).
- Scalability to other fields (eg, the resource description framework [RDF] has been adopted by systems biology and molecular biology for specific data representation [43,44]).
- Advanced query system (ie, with the SPARQL language).
- Existing tools in a rich community to create, maintain, validate, explore, and visualize RDF representation (eg, Protégé and WebVOWL [45-47]).

A set of rules and conventions has been defined to guide the creation of an SPHN RDF schema, that is, how RDF classes and properties required to generate instances (RDF resources) for storing hospitals' data should be created [48,49]. Particularly, such rules stipulate (1) how concepts defined in the SPHN semantic data set should be converted into RDF classes or RDF properties and (2) how concepts that are not semantically linked to each other by composition should be linked to encapsulate contextual information provided at the time of data capture.

Swiss hospitals' clinical research data warehouses are primarily based on relational database management systems. To transform data from a relational model representation into a graph representation based on RDF, extract, transform, and load (ETL) pipelines have been implemented by data providers' informatics teams. They typically include an RDF transformer step where raw data from the EHR is converted and loaded into a triple store. Then, data can be extracted and serialized into RDF files for each specific project.

Third Pillar

Converters are used to transform the RDF data into purpose-specific data models, serializing the RDF data into other common formats such as XML, JSON, JSON-LD, and TSV/CSV. For example, SPARQL queries have been implemented to convert data into flat files that can be processed by research-enabling software or machine learning pipelines [49].

Discussion

Overview

While the proposed data interoperability strategy offers a number of advantages in terms of flexibility and extensibility over more conventional approaches based on common data models, several challenges had to be addressed to allow effective implementation.

Granularity Challenges

Finding the right representation for a concept is not trivial. Data can be represented in many ways (eg, "arm circumference" defined as a concept or a "circumference" concept connected with a "body site" concept taking the value "arm"), and agreeing on a common way to represent data is a challenging process. While both of these representations may be correct, interoperability is not always ensured if both are used, even though an international KR is used. This difference in the level of granularity also influences the way the user can query the data. When only one level of granularity is used in a specific data set, querying for relevant information is trivial. The user simply queries for the data of interest using the relevant defined concepts. However, if the data set comes from two different sources with different levels of granularity for the same type of information, either the querying needs to be adapted so that it can recognize both patterns or mapping must be performed beforehand to ensure that the results obtained are complete. Within the SPHN community, the granularity challenge has been addressed in the following two complementary ways: (1) when possible, a specific level is agreed by consensus and (2) in all other situations, all levels are encoded using a KR (for example SNOMED CT), allowing to query at different levels of granularity.

Different Needs

Defining a common concept for different use cases proved to be complex when creating the semantic framework. Depending on the project, needs may vary widely. For example, one project may require the temperature of a patient, without any information on the site or the device used to measure it, while another project may require the exact device and site for the temperature. This problem is addressed by representing the meaning strongly, therefore allowing the different concepts to be represented. Thus, it is possible to express temperature and many additional (present and future) concepts, and associate them freely. This is a major advantage when compared to any formal data model. When a concept requires further specification, it can be combined with other existing concepts (eg, body site and device) or extended by new project-specific properties.

Implementation Challenges

The process of clinical data acquisition passes through numerous filters before it ends up in a data warehouse for further usage. From acquisition of the data through questionnaires, formularies, texts, devices, etc in many different systems to the warehouse, several ETL processes usually will be required, resulting in loss of information. Therefore, the granularity and precision of the back-office semantic linkage can only represent the information richness known at that time. For example, the status "covid positive" cannot be coded in LOINC as this would require knowing the analytical method used by the laboratory. During that process, similar data in the data warehouse might originate from different contexts, which are not represented in the data warehouse. This is true within a care provider organization and is amplified when aggregating data originating from different care facilities and sources. These challenges were addressed in the strategy in several manners. The semantic framework with clear definitions of the concepts and their encoding in KR limited the ambiguity when creating the ETL procedures in the hospital. Second, the task of mapping the raw data to SPHN concepts was performed in each hospital by people knowing the internal data acquisition processes. Finally, the possibility to include relevant KR depending on the use case allowed the inclusion of relevant classifications used directly in care facilities, such as clinical, logistic, and billing classifications.

Resource Challenges

The creation, evolution, and management of these semantic descriptions raise several challenges, notably scalability and coherence. Since the data sets rely on multiple external standards, there is versioning required, especially because the data considered can cover decades. The same is true for the maintenance of KR created in the project and for the infrastructure and human resources that will handle the storage and transport layers in hospitals. Most hospitals did not know RDF before the SPHN strategy. Competencies had to be built internally to ensure the sustainability of local solutions. Adoption by care facilities has thus been a critical factor to

improve successful and sustainable implementation, with development of strategies for internal added value.

Competencies and Educational Challenges

The introduction of several new approaches in care facilities (semantic-centered data handling, formal descriptive language for storage and transport, and relegating data models to the end of the data pipeline) has been a huge challenge and still encounters resistances in the information technology (IT) community. Dedicated efforts in building several working groups for semantics, RDF, and data model bridging involving numerous hospital representatives have been important to handle this challenge. This was managed by the DCC, which gathered representatives from all stakeholders. The task of identifying the list of variables to be exchanged and their prioritization was given to the research projects.

The semantic framework is bound to evolve as the user base grows, and this evolution must follow the needs of projects without compromising the strategy. This will only be possible if the strategy is well understood both centrally and at the hospital level by specialists in medical informatics within IT departments. A strong effort is therefore currently underway within the SPHN to disseminate the strategy via the publication of strategic papers, webinars, and courses given to members of the SPHN community [50-52].

Conclusions

The main contribution of this work involves a new strategy for enabling nationwide intercommunity health data interoperability. The proposed strategy relies on the development of a semantic-based framework, which is designed to not replace existing standards but use them in a synergistic, pragmatic, and purpose-specific way. As the framework is built on the compositionality principle, it offers high flexibility and sustainability. The use of formal descriptive languages, such as RDF, as a data storage and transport layer ensures strong scalability to new needs. At the final stage, building specific bridges to fulfill the many data models used in research or required to comply with regulatory frameworks has proven successful and has been an important asset to ensure continuity of existing processes.

The wide adoption of the proposed strategy by every university hospital and high research organization in Switzerland as the national standard for sharing clinical data marks an important transition to an interoperable landscape for personalized health in Switzerland.

Acknowledgments

We would like to acknowledge the Swiss Personalized Health Network (SPHN) Clinical Data Semantic Interoperability, and the Hospital IT Working Groups and the resource description framework (RDF) Task Force, as well as all five university hospitals for their contribution to the implementation of the strategy. This work was funded by the Swiss Government through the SPHN Initiative.

Conflicts of Interest

CL is the Editor-in-Chief of this journal (JMIR Medical Informatics).

References

1. International Classification of Diseases, 11th Revision (ICD-11). World Health Organization. URL: http://www.who.int/classifications/icd/en/ [accessed 2021-06-09]
2. Health Level Seven International-Homepage. HL7 International. URL: http://www.hl7.org/ [accessed 2021-06-09]
3. Bhattacharyya SB. SNOMED CT History and IHTSDO. In: Introduction to SNOMED CT. Singapore: Springer; 2016.
4. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. N Engl J Med 2010 Aug 05;363(6):501-504. [doi: 10.1056/NEJMp1006114] [Medline: 20647183]
5. Integrating the Healthcare Enterprise (IHE). IHE International. URL: https://www.ihe.net/ [accessed 2021-06-09]
6. Sternbach GL. The Glasgow Coma Scale. The Journal of Emergency Medicine 2000 Jul;19(1):67-71. [doi: 10.1016/s0736-4679(00)00182-7] [Medline: 10863122]
7. American Academy of Pediatrics, Committee on Fetus and Newborn, American College of Obstetricians and Gynecologists and Committee on Obstetric Practice. The Apgar score. Pediatrics 2006 Apr;117(4):1444-1447. [doi: 10.1542/peds.2006-0325] [Medline: 16585348]
8. What Are the Classifications of Heart Failure? Heart Failure. 2019. URL: https://heart-failure.net/classification [accessed 2021-06-09]
9. Schoonjans F, Zalata A, Depuydt CE, Comhaire FH. MedCalc: a new computer program for medical statistics. Comput Methods Programs Biomed 1995 Dec;48(3):257-262. [doi: 10.1016/0169-2607(95)01703-8] [Medline: 8925653]
10. Elovic A, Pourmand A. MDCalc Medical Calculator App Review. J Digit Imaging 2019 Oct;32(5):682-684 [FREE Full text] [doi: 10.1007/s10278-019-00218-y] [Medline: 31025219]
11. ICD-10 Version:2019. World Health Organization. URL: https://icd.who.int/browse10/2019/en [accessed 2021-06-09]
12. McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clin Chem 2003 Apr;49(4):624-633. [doi: 10.1373/49.4.624] [Medline: 12651816]

13. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 01;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
14. The Unified Medical Language System (UMLS). National Library of Medicine. URL: https://www.nlm.nih.gov/research/umls/new_users/online_learning/OVR_001.html [accessed 2021-06-09]
15. GTIN Definition: Information. GTIN. URL: https://www.gtin.info/ [accessed 2021-06-09]
16. WHO Collaborating Centre for Drug Statistics Methodology-Home. WHOCC. URL: https://www.whocc.no/ [accessed 2021-06-09]
17. MedDRA. URL: https://www.meddra.org/ [accessed 2021-06-09]
18. Planned transition to concrete domains. SNOMED. 2020. URL: https://www.snomed.org/news-and-events/articles/planned-transition-concrete-domains [accessed 2021-06-09]
19. Gaudet-Blavignac C, Foufi V, Bjelogrlic M, Lovis C. Use of the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) for Processing Free Text in Health Care: Systematic Scoping Review. J Med Internet Res 2021 Jan 26;23(1):e24594 [FREE Full text] [doi: 10.2196/24594] [Medline: 33496673]
20. Benson T, Grieve G. Implementing Terminologies. In: Principles of Health Interoperability. Health Information Technology Standards. Cham: Springer; 2016:189-219.
21. Benson T, Grieve G. The HL7 v3 RIM. In: Principles of Health Interoperability. Health Information Technology Standards. Cham: Springer; 2016:243-264.
22. Klann JG, Joss MAH, Embree K, Murphy SN. Data model harmonization for the All Of Us Research Program: Transforming i2b2 data into the OMOP common data model. PLoS One 2019;14(2):e0212463 [FREE Full text] [doi: 10.1371/journal.pone.0212463] [Medline: 30779778]
23. OMOP Common Data Model. OHDSI. URL: https://www.ohdsi.org/data-standardization/the-common-data-model/ [accessed 2021-06-09]
24. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010;17(2):124-130 [FREE Full text] [doi: 10.1136/jamia.2009.000893] [Medline: 20190053]
25. Hume S, Aerts J, Sarnikar S, Huser V. Current applications and future directions for the CDISC Operational Data Model standard: A methodological review. J Biomed Inform 2016 Apr;60:352-362 [FREE Full text] [doi: 10.1016/j.jbi.2016.02.016] [Medline: 26944737]
26. Kalra D, Beale T, Heard S. The openEHR Foundation. Stud Health Technol Inform 2005;115:153-173. [Medline: 16160223]
27. Thomas B. Archetypes: Constraint-based Domain Models for Futureproof Information Systems. CiteSeerX. 2000. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.21.1158 [accessed 2021-06-09]
28. Swiss Personalized Health Network (SPHN). URL: https://sphn.ch [accessed 2021-06-09]
29. First review report of the International Advisory Board. SPHN. 2019. URL: https://sphn.ch/2019/12/20/iab-report/ [accessed 2021-06-09]
30. Swiss Personalized Health Network. Report from the National Steering Board 2016–2019. Zenodo. URL: https://zenodo.org/record/4044123 [accessed 2021-06-09]
31. Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C. URL: https://www.w3.org/TR/rdf-concepts/ [accessed 2021-06-09]
32. Samwald M, Fehre K, de Bruin J, Adlassnig K. The Arden Syntax standard for clinical decision support: experiences and directions. J Biomed Inform 2012 Aug;45(4):711-718 [FREE Full text] [doi: 10.1016/j.jbi.2012.02.001] [Medline: 22342733]
33. OWL Web Ontology Language Semantics and Abstract Syntax. W3C. URL: https://www.w3.org/TR/2004/REC-owl-semantics-20040210/ [accessed 2021-06-09]
34. Peleg M, Boxwala AA, Ogunyemi O, Zeng Q, Tu S, Lacson R, et al. GLIF3: the evolution of a guideline representation format. Proc AMIA Symp 2000:645-649 [FREE Full text] [Medline: 11079963]
35. jBPM. URL: https://www.jbpm.org/ [accessed 2021-06-09]
36. Mate S, Köpcke F, Toddenroth D, Martin M, Prokosch H, Bürkle T, et al. Ontology-based data integration between clinical and research systems. PLoS One 2015;10(1):e0116656 [FREE Full text] [doi: 10.1371/journal.pone.0116656] [Medline: 25588043]
37. Stöhr MR, Majeed RW, Günther A. Metadata Import from RDF to i2b2. Stud Health Technol Inform 2018;253:40-44. [Medline: 30147037]
38. Solbrig HR, Hong N, Murphy SN, Jiang G. Automated Population of an i2b2 Clinical Data Warehouse using FHIR. AMIA Annu Symp Proc 2018;2018:979-988 [FREE Full text] [Medline: 30815141]
39. Pacaci A, Gonul S, Sinaci AA, Yuksel M, Laleci Erturkmen GB. A Semantic Transformation Methodology for the Secondary Use of Observational Healthcare Data in Postmarketing Safety Studies. Front Pharmacol 2018 Apr 30;9:435 [FREE Full text] [doi: 10.3389/fphar.2018.00435] [Medline: 29760661]
40. Data Coordination Center (DCC). SPHN. URL: https://sphn.ch/network/data-coordination-center/ [accessed 2021-06-09]
41. SPHN Dataset Release. SPHN. URL: https://sphn.ch/document/sphn-dataset/ [accessed 2021-06-09]
42. Decker S, Melnik S, van Harmelen F, Fensel D, Klein M, Broekstra J, et al. The Semantic Web: the roles of XML and RDF. IEEE Internet Comput 2000 Sep;4(5):63-73. [doi: 10.1109/4236.877487]

43. Demir E, Cary MP, Paley S, Fukuda K, Lemer C, Vastrik I, et al. The BioPAX community standard for pathway data sharing. Nat Biotechnol 2010 Sep;28(9):935-942 [FREE Full text] [doi: 10.1038/nbt.1666] [Medline: 20829833]
44. Antezana E, Blondé W, Egaña M, Rutherford A, Stevens R, De Baets B, et al. BioGateway: a semantic systems biology tool for the life sciences. BMC Bioinformatics 2009 Oct 01;10 Suppl 10:S11 [FREE Full text] [doi: 10.1186/1471-2105-10-S10-S11] [Medline: 19796395]
45. Lohmann S, Link V, Marbach E, Negru S. WebVOWL: Web-based Visualization of Ontologies. In: Lambrix P, Hyvönen E, Blomqvist E, Presutti V, Qi G, Sattler U, et al, editors. Knowledge Engineering and Knowledge Management. EKAW 2014. Lecture Notes in Computer Science, vol 8982. Cham: Springer; 2015:154-158.
46. Musen MA, Protégé Team. The Protégé Project: A Look Back and a Look Forward. AI Matters 2015 Jun;1(4):4-12 [FREE Full text] [doi: 10.1145/2757001.2757003] [Medline: 27239556]
47. Protégé. URL: https://protege.stanford.edu/ [accessed 2021-06-09]
48. SPHN RDF Schema. SPHN. URL: https://sphn-semantic-framework.readthedocs.io/en/latest/sphn_framework/sphnrdfschema.html#technical-specification [accessed 2021-06-09]
49. SPHN RDF quality control. GitLab. URL: https://git.dcc.sib.swiss/sphn-semantic-framework/sphn-rdf-quality-control [accessed 2021-06-09]
50. SPHN Webinar Series. SPHN. URL: https://sphn.ch/services/seminars-training/ [accessed 2021-06-09]
51. Clinical Data Semantics Interoperability Working Group Strategy. SPHN. URL: https://sphn.ch/document/csi_wg_strategy/ [accessed 2021-06-09]
52. Fact sheet Semantic Strategy. SPHN. URL: https://sphn.ch/document/fact-sheet-semantic-strategy/ [accessed 2021-06-09]

Abbreviations

ATC: Anatomical Therapeutic Chemical
CDISC: Clinical Data Interchange Standards Consortium
DCC: Data Coordination Center
EHR: electronic health record
ETL: extract, transform, and load
GTIN: Global Trade Item Number
HL7: Health Level 7
i2b2: Informatics for Integrating Biology and the Bedside
ICD: International Classification of Diseases
IT: information technology
KR: knowledge representation
LOINC: Logical Observation Identifiers Names and Codes
NSB: National Steering Board
OMOP: Observational Medical Outcomes Partnership
RDF: resource description framework
RIM: Reference Information Model
SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms
SPHN: Swiss Personalized Health Network

Edited by CL Parra-Calderón, G Eysenbach; submitted 29.01.21; peer-reviewed by G Jiang, A Rector, S Nelson; comments to author 22.02.21; revised version received 27.04.21; accepted 19.05.21; published 24.06.21.

Please cite as:
Gaudet-Blavignac C, Raisaro JL, Touré V, Österle S, Crameri K, Lovis C
A National, Semantic-Driven, Three-Pillar Strategy to Enable Health Data Secondary Usage Interoperability for Research Within the Swiss Personalized Health Network: Methodological Study
JMIR Med Inform 2021;9(6):e27591
URL: https://medinform.jmir.org/2021/6/e27591/
doi:10.2196/27591
PMID:34185008

Original Paper

Unsupervised Machine Learning for Identifying Challenging Behavior Profiles to Explore Cluster-Based Treatment Efficacy in Children With Autism Spectrum Disorder: Retrospective Data Analysis Study

Julie Gardner-Hoag[1], BSc, MSc; Marlena Novack[2], BA; Chelsea Parlett-Pelleriti[3], BA, MSc; Elizabeth Stevens[3*], BSc, MSc, PhD; Dennis Dixon[2*], BA, PhD; Erik Linstead[3*], BSc, MSc, PhD

[1]Schmid College of Science and Technology, Chapman University, Orange, CA, United States

[2]Center for Autism and Related Disorders, Woodland Hills, CA, United States

[3]Fowler School of Engineering, Chapman University, Orange, CA, United States

[*]these authors contributed equally

Corresponding Author:
Erik Linstead, BSc, MSc, PhD
Fowler School of Engineering
Chapman University
One University Drive
Orange, CA, 92866
United States
Phone: 1 714 289 3159
Email: linstead@chapman.edu

Abstract

Background: Challenging behaviors are prevalent among individuals with autism spectrum disorder; however, research exploring the impact of challenging behaviors on treatment response is lacking.

Objective: The purpose of this study was to identify types of autism spectrum disorder based on engagement in different challenging behaviors and evaluate differences in treatment response between groups.

Methods: Retrospective data on challenging behaviors and treatment progress for 854 children with autism spectrum disorder were analyzed. Participants were clustered based on 8 observed challenging behaviors using *k* means, and multiple linear regression was performed to test interactions between skill mastery and treatment hours, cluster assignment, and gender.

Results: Seven clusters were identified, which demonstrated a single dominant challenging behavior. For some clusters, significant differences in treatment response were found. Specifically, a cluster characterized by low levels of stereotypy was found to have significantly higher levels of skill mastery than clusters characterized by self-injurious behavior and aggression (P<.003).

Conclusions: These findings have implications on the treatment of individuals with autism spectrum disorder. Self-injurious behavior and aggression were prevalent among participants with the worst treatment response, thus interventions targeting these challenging behaviors may be worth prioritizing. Furthermore, the use of unsupervised machine learning models to identify types of autism spectrum disorder shows promise.

(JMIR Med Inform 2021;9(6):e27793) doi:10.2196/27793

KEYWORDS

autism spectrum disorder; challenging behaviors; unsupervised machine learning; subtypes; treatment response; autism; treatment; behavior; machine learning; impact; efficacy; disorder; engagement; retrospective

Introduction

Autism spectrum disorder is a neurodevelopmental disorder characterized by deficits in social communication and social interaction, as well as the presence of restricted, repetitive patterns of behavior, interests, and activities [1]. With the exception of restricted, repetitive behaviors (eg, stereotypy, perseveration), challenging behaviors are not classified as a core symptom of autism spectrum disorder; however, these behaviors are prevalent among individuals with autism spectrum disorder. As many as 94% of children with autism spectrum disorder engage in some type of challenging behavior, often including stereotypy (eg, self-stimulatory or persistent repetitive motor or vocal behavior), aggression, tantrums, and self-injurious behavior [2,3]. Challenging behaviors may pose risk of injury to the individual or others and may inhibit learning opportunities and social interactions [4]. Furthermore, challenging behaviors may negatively impact family functioning and contribute to caregiver stress [5,6].

Various risk factors for engagement in challenging behaviors have been investigated in individuals with autism spectrum disorder. Symptom severity has been found to predict challenging behaviors, with greater symptom severity associated with engagement in higher numbers of challenging behaviors at stronger intensities [2,3]. Intellectual functioning has also been linked to challenging behaviors in individuals with autism spectrum disorder, with greater deficits in intellectual functioning predicting greater frequencies of stereotypy [7,8], aggression [8], and self-injurious behavior [8,9]. In addition, deficits in adaptive skills [10,11] and expressive language skills [11] have been associated with engagement in challenging behaviors in individuals with autism spectrum disorder, but studies [8-12] that investigated the relationship between gender and challenging behaviors found no significant differences in engagement in challenging behaviors between boys and girls with autism spectrum disorder.

Applied behavior analysis interventions, which involve the application of principles and procedures of learning and motivation to alter behavior [13,14], may be used to reduce challenging behaviors and increase appropriate behaviors in individuals with autism spectrum disorder. Specific challenging behaviors that are commonly addressed in treatment include stereotypy, noncompliance, and aggression [15]. Outcome studies for children with autism spectrum disorder have not often included challenging behaviors as an outcome measure [4,16]. Several group design studies [17-19] have found evidence to support the use of caregiver training to manage challenging behaviors. Furthermore, there is an abundance of single-individual research evaluating the effectiveness of behavioral interventions for challenging behaviors in individuals with autism spectrum disorder, and reviews of this research have found behavioral interventions, particularly those implementing pretreatment functional assessments, to be effective in reducing challenging behaviors [20-22].

Applied behavior analysis–based therapy is considered to be well-established for the treatment of autism spectrum disorder [23,24]. While ample research demonstrates the effectiveness of applied behavior analysis–based treatment [25,26] research also reveals variability in individual response to treatment [27,28]. Treatment-related variables including greater treatment intensity [27,29-32], longer treatment duration [30-32], and greater total intervention time [33,34] have been linked to superior treatment outcomes. Furthermore, many patient-related variables have been associated with greater treatment gains. These include younger age [29,32,34-38], lower autism spectrum disorder symptom severity [35,36,38,39], and greater intellectual functioning [27,36,38-45].

Research evaluating the impact of challenging behaviors on treatment response in individuals with autism spectrum disorder is limited. Eikeseth and colleagues [46] investigated whether challenging behaviors, among other intake measures, were associated with treatment outcomes for adaptive behavior and autism spectrum disorder symptoms in children with autism spectrum disorder; however, challenging behaviors were not found to be a predictor of treatment outcome. Conversely, Remington and colleagues [39] found that higher rates of challenging behaviors at intake were associated with superior response to treatment and suggested that their counterintuitive findings may possibly be attributed to the sensitivity of the measure used to assess challenging behaviors. Given the prevalence of challenging behaviors among individuals with autism spectrum disorder, additional research is needed to investigate the impact of these behaviors on treatment response.

To account for the heterogeneity observed across individuals with autism spectrum disorder, researchers have investigated types of autism spectrum disorder [47]. Preliminary research has found behavioral types of autism spectrum disorder to have differences in gene expression [48-50], developmental trajectory [51-54], and treatment response [55]. In a recent study, Stevens and colleagues [55] used an unsupervised machine learning model to identify behavioral types of autism spectrum disorder and evaluate differences in treatment response across types. Participants included 2400 children with autism spectrum disorder. Data from a comprehensive assessment of skill deficits and treatment progress data were analyzed. A total of 16 autism spectrum disorder groups were identified using a Gaussian mixture model. Using a linear regression model, relationships between treatment hours and skill mastery were found to be strong within groups, accounting for 64% to 75% of variance. These findings are a preliminary step toward advancing targeted treatments and improving outcomes for individuals with autism spectrum disorder based on type membership.

Autism spectrum disorder types may also be identified based upon profiles of challenging behavior. Stevens and colleagues [56] conducted an analysis of challenging behaviors in a large sample of children with autism spectrum disorder (n=2116). Using *k*-means clustering, 8 diverse profiles, in which a single dominant challenging behavior was apparent, were identified. Furthermore, gender differences were observed when cluster analyses were performed separately for male and female participants. While all of the male clusters were found to exhibit a single dominant challenging behavior, 2 of the female clusters indicated equal engagement in 2 dominant challenging behaviors. These findings suggest that gender may play a role in the presentation of challenging behaviors in individuals with

autism spectrum disorder. Further investigations into autism spectrum disorder types based on challenging behaviors are warranted.

The study of challenging behaviors across types of autism spectrum disorder may help explain some of the variation observed in treatment outcomes across individuals with autism spectrum disorder and may advance efforts to develop targeted treatments to maximize outcomes. Preliminary evidence indicates there are autism spectrum disorder types based on challenging behaviors; however, little is known about how challenging behaviors impact treatment response. The purpose of this study was to identify types of autism spectrum disorder based on engagement in different challenging behaviors and evaluate differences in treatment response between groups and across gender.

Methods

Data Set

Deidentified retrospective treatment data for a large sample of children with autism spectrum disorder were used in this study. Data on the frequency of challenging behaviors and treatment progress were obtained from the Skills system software (Skills Global LLC [57]). Skills includes a thorough assessment of skill deficits with demonstrated reliability [58] and validity [59], a comprehensive curriculum to build individualized treatment plans, and tracking capabilities for challenging behaviors and ongoing treatment progress. In addition to Skills data, operational data on treatment hours were used in this study.

Participants included children with autism spectrum disorder who were receiving applied behavior analysis treatment from a community-based provider. A total of 2116 clinical records were reviewed based on the following inclusion criteria: (1) were between the ages 18 months and 12 years old; (2) had a diagnosis of autism spectrum disorder, autistic disorder, pervasive developmental disorder–not otherwise specified, or Asperger syndrome by an independent licensed clinician (eg, psychologist, pediatrician, etc); (3) received at least 20 hours of treatment per month; and (4) had at least 1 month of continuous services; (5) demonstrated repeated instances of challenging behavior as documented in their treatment history; and (6) had available treatment response data over the course of treatment. Parameters with respect to age were set based on the age range predominately represented in the data set to avoid potential outliers that may have affected the cluster analysis. Likewise, parameters regarding treatment intensity and duration were established so that each participant had adequate treatment response data to include in the analysis. After applying these criteria, a sample of 854 participants were included. Of the participants, 706 were male and 148 were female. The average age of participants was 7.59 (SD 2.17) years old, ranging from 2.74 years to 12 years. Participants resided in the states of Arizona, California, Colorado, Illinois, Louisiana, New York, Texas, and Virginia. The data used for this study were collected during a 36-month period (January 1, 2014 through December 31, 2016).

Measures

Data on engagement in challenging behaviors were used to identify potential clusters. While the classification of challenging behaviors is subjective in nature, there is agreement among the literature regarding operational definitions for common topographies of challenging behaviors exhibited by individuals with autism spectrum disorder [4]. While this may not be exhaustive, data were examined for the following topographies of challenging behaviors: aggression (eg, hitting, kicking), disruption (eg, interrupting, yelling), elopement (eg, wandering, bolting), noncompliance (eg, defiant behavior, refusing), obsessive behavior (eg, repeatedly talking about the same topic, preservation), self-injurious behavior (eg, head banging, hand biting), stereotypy (eg, hand flapping, toe walking, vocal stereotypy), and tantrums (eg, crying, falling). Skills is implemented as a relational database, which allows behavior interventionists to record observations in real time during a therapy session using an iPad and the corresponding Skills app. In the case of challenging behaviors, when such a behavior is observed, the therapist marks the type of behavior and provides a textual description of its context. This information is then timestamped and then stored in the underlying relational database. Aggregation of challenging behavior data for each patient can then be easily achieved using a simple database query (SQL format). An extra validation step was taken to compare identified challenging behaviors to the textual description provided by the behavior interventionist to ensure no challenging behavior observations were misidentified.

Data on mastered learning objectives were used to evaluate treatment response. Mastery criteria for learning objectives were determined by the patient's clinician and individualized to the patient. Typically, mastery was defined as 80% accuracy or greater for a minimum of 2 treatment sessions across 2 days.

Treatment

Participants received individualized applied behavior analysis–based treatment. Treatment comprehensively targeted deficits across developmental domains, including language, social, adaptive, cognitive, executive function, academic, play, and motor skills. Services were provided in the participant's home, clinic, school, community, or a variety of settings. Treatment was provided according to the Center for Autism and Related Disorders model [60].

Participants' treatment programs addressed skill acquisition and targeted the reduction of challenging behaviors. Interventions for challenging behaviors varied based on the target behavior's topography and function (determined using functional assessment). Possible interventions implemented by a participant's clinician included: antecedent-based interventions (ie, manipulations to the environment to reduce the target behavior) such as noncontingent reinforcement, demand fading, task modification, and choice; replacement behavior interventions including functional communication training, differential reinforcement of alternative behavior, and differential reinforcement of incompatible behavior; and consequence-based interventions (ie, manipulations to the events following the target behavior to reduce the likelihood of its reoccurrence) such as extinction, differential reinforcement of

other behavior, differential reinforcement of low rate behavior, and response interruption and redirection.

Data Analysis

Clustering

This analysis expanded on the work of Stevens and colleagues [56] to explore differences in treatment response across identified challenging behavior clusters in individuals with autism spectrum disorder. Patients were clustered based on relative frequency of 8 challenging behaviors (aggression, disruption, elopement, noncompliance, obsessive behavior, self-injurious behavior, stereotypy, and tantrums) using a *k*-means machine learning algorithm. This was achieved by creating an 8-dimensional feature vector for each patient. Relative frequency was calculated by finding the proportion of all challenging behaviors for each of the 8 categories for each patient. Duration and severity of the challenging behaviors were not taken into consideration for this value. Each vector element corresponded to the relative frequency of a specific challenging behavior observed for that patient. The 8-dimensional vectors were fed directly to the clustering algorithm without the use of feature selection because the dimensionality of the data was relatively small, and it was important to preserve each of the challenging behaviors in the final cluster model. Once clusters were identified using the *k*-means algorithm, multiple linear regression was performed to evaluate interactions between cluster assignment, treatment response, and gender.

The goal of clustering is to find latent groups, or clusters, in data. Patients within the same cluster will have more similar challenging behaviors profiles than patients in different clusters [61]. The *k*-means methods was selected for clustering because it is computationally efficient, easily implemented, and is a widely used prototype-based clustering algorithm, wherein each cluster is represented by a prototype. This prototype can either be the centroid of data points with similar continuous features or the medoid in the case of categorical features. This data set involved continuous features; therefore, each cluster had a centroid.

The *k*-means algorithm was implemented with 5 steps. (1) The best value of *k* (ie, the number of clusters) was identified by incrementally testing values between 2 and 20. (2) For each of these values, the algorithm picked *k* sample points from the data at random, which are the initial centroids (c_1, c_2, ..., c_k). (3) Each 8-dimensional data point d_i was assigned to the nearest centroid c_k using Euclidean distance to measure the distance between the point and the centroid. (4) The algorithm recalculated the centroids by taking the mean value (for each behavior) from all the data points currently in the cluster. (5) The algorithm repeated steps 3 and 4 until the cluster assignments did not change or a maximum number of iterations was reached.

To find the distance between the data points and the centroids in the data set, squared Euclidean distance was used. Similarity between data points is defined as the opposite of distance. A commonly used metric for finding the distance between data points *x* and *y* in *m*-dimensional space is the squared Euclidean distance.

Once similarities are measured, clustering becomes an optimization problem. An iterative approach was used to minimize the within-cluster sum of squared errors or cluster inertia. Once these errors were calculated, a graph of the errors were examined using the elbow method to find the best value for *k*. The elbow method involves examining the plot (ie, the arm) to determine the point at which diminishing returns are observed (ie, the elbow). As *k* increases, the sum of squared errors gets smaller. When *k* is equal to the number of points in the data set, the sum of squared errors is 0 and every point is its own cluster. Choosing *k* to correspond to the elbow in the graph thus provides an effective measure by which to prevent overfitting. The chosen *k* indicates the optimal number of clusters that are both cohesive and separate.

Linear Regression

A multiple linear regression model was used to evaluate the relationships between the target variable (skill mastery) and explanatory variables (treatment intensity, cluster assignment, and gender).

In univariate linear regression, the relationship between a single explanatory variable *x* and a response or target variable *y* is modeled. The equation used for linear models with only 1 predictor variable is defined as $y_i=\beta_0+\beta_1x_i+\varepsilon_i$, where the weight β_0 represents the *y*-axis intercept and β_1 is the coefficient of the explanatory variable. In simple linear regression, the goal is to find the weights of the equation to explain the relationship between the explanatory variable and the target variable. From this, the responses of new data points that were not part of the observed data may be predicted and coefficients of the model may be interpreted. The simple linear regression equation may be generalized to produce an equation for multiple linear regression that involves multiple explanatory variables.

Linear regression works by taking the explanatory variables and the response variable, and fitting a straight hyperplane to the data that minimizes the distance between an observed point and the fitted model. The explanatory variables were treatment intensity, cluster assignment, and gender, and the response variable was skill mastery.

An efficient way to quantitatively measure a model's performance is the mean squared error, which measures the average squared error between the model's prediction and the actual values. Mean squared error may be used to compare different regression models with the same outcome.

R^2 (another measure of model fit) is bounded between 0 and 1, with 1 indicating a perfect relationship between *x* and *y* and mean squared error is equal to 0. *R* allows for the specification of interaction terms in regression formulas. An interaction occurs when the product of 2 predictor variables is also a significant predictor [62]. In this study, there were 3 explanatory variables (treatment intensity, cluster assignment, and gender), and all interactions were included in the model.

Tukey Posthoc

Results from the regression model indicated that there was a significant difference between treatment hours, cluster assignment, and the interaction between cluster assignment and

gender. Posthoc analysis was conducted to determine which pairs of clusters significantly differed. The Tukey honestly significant difference method was used to correct for multiple comparisons.

The Tukey posthoc test assesses all the pairwise comparisons using the Tukey honestly significant difference formula

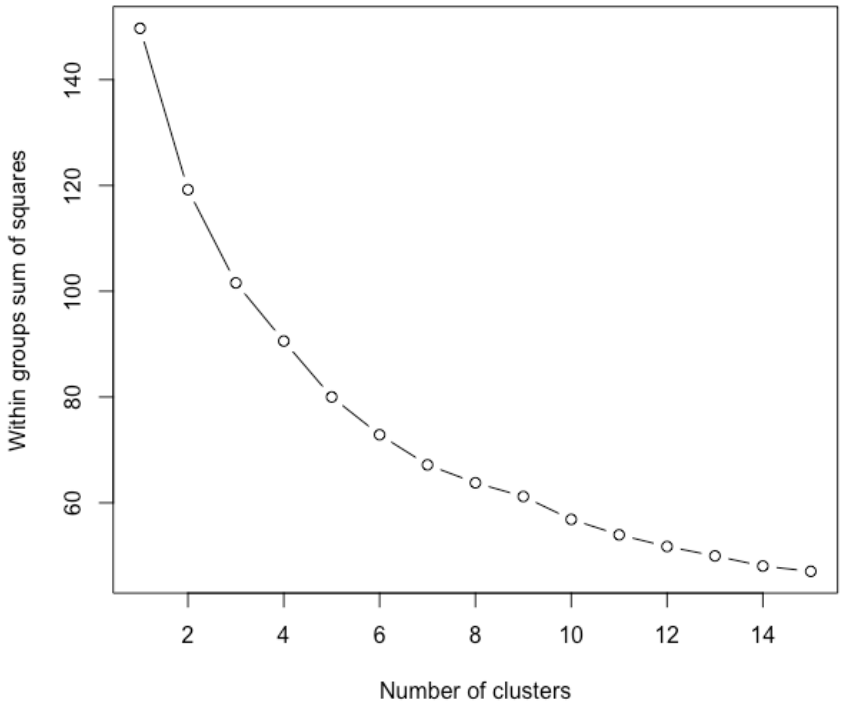
where $M_i - M_j$ is the difference between the pairs of means, MS_w is the mean square within, and n is the number of clusters.

Results

Clustering

Figure 1 shows within-group sum of squared errors for patients. The optimal value of k (the number of distinct challenging behavior profiles) was found to be 7, confirmed both by the elbow method and by silhouette score (the highest indicates most cohesive and separate). Each cluster corresponds to a phenotype and can be quantitatively represented with its centroid (the mean relative frequency for each of the 8 challenging behaviors for patients in that cluster). The dimensionality of each centroid is identical to the input feature space, which is preserved during the clustering process.

Figure 1. Within-cluster sum of squared errors for all patients, both male and female. The elbow method indicates that the best value of k is 7, meaning there are 7 clusters.



Seven phenotypes of autism spectrum disorder, most of which demonstrated 1 dominant challenging behavior, were identified based on average frequency (centroid) of 8 challenging behaviors (ie, aggression, disruption, elopement, noncompliance, obsessive behavior, self-injurious behavior, stereotypy, and tantrums) calculated for each cluster (Table 1). It is important to reiterate that the machine learning process is unsupervised. The phenotypes are identified by the algorithm without the need for human labels, which are required for supervised learning (classification).

Table 1. Breakdown of identified clusters.

Cluster	Dominant challenging behavior	Boys, n	Girls, n	All, n
1	Tantrums	60	14	74
2	Self-injurious behavior	79	8	87
3	Elopement	78	18	96
4	Stereotypy (low rate)	170	37	207
5	Noncompliance	87	26	113
6	Aggression	113	26	139
7	Stereotypy (high rate)	119	19	138

The radar graphs shown in Figure 2 and Figure 3 provide visual representations of each phenotype's engagement in the 8 challenging behaviors. The radar charts in Figure 3 were scaled from 0 to the average frequency of the dominant challenging behavior. For example, Cluster 1 was scaled from 0 to 0.6, to which tantrums extend. Cluster 4 was scaled from 0 to 0.4, to which stereotypy extends. It is worth noting that Cluster 4 and Cluster 7 both have stereotypy as their dominant challenging behavior, but their frequencies were different. Cluster 4 was found to engage in stereotypy at a lower rate than Cluster 7.

Figure 2. Radar graphs depicting engagement in challenging behaviors across clusters.

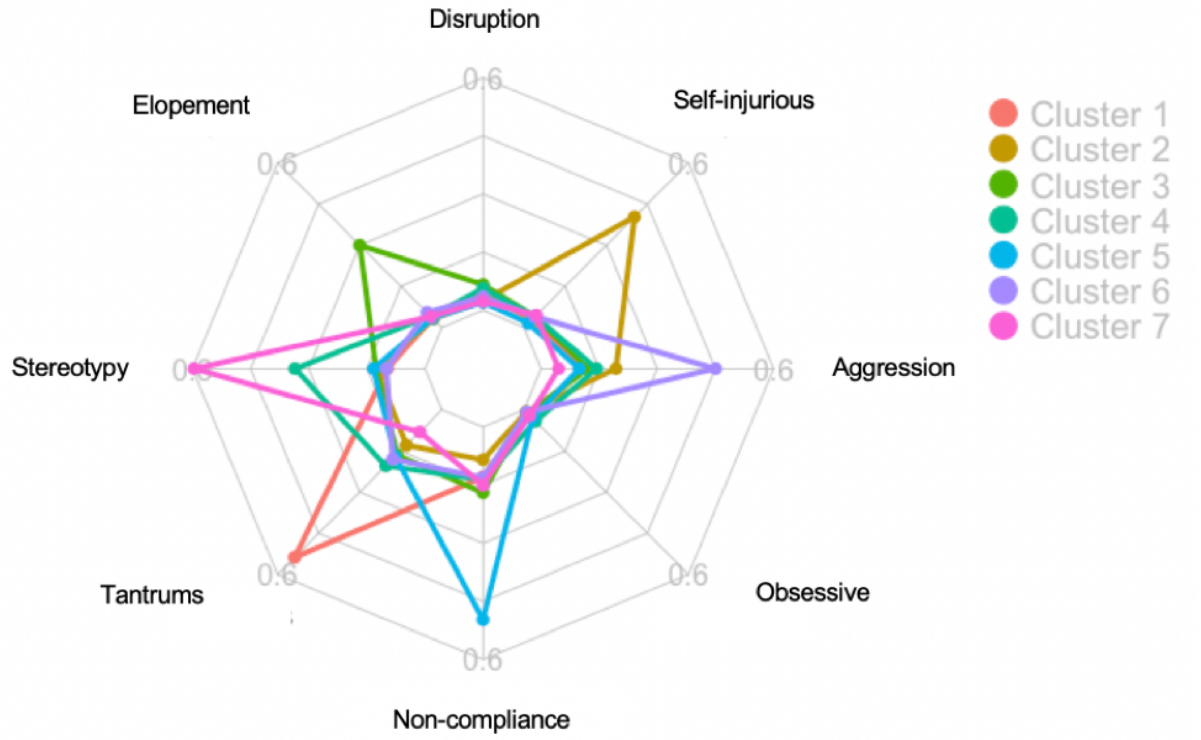


Figure 3. Radar graphs showing the dominant challenging behavior for each cluster. Note that the maximum varies between the clusters, particularly Cluster 4 and Cluster 7, in which patients demonstrate the same dominant challenging behavior.

Linear Regression

The R^2 value was found to be 0.67. The value for R^2 is the fraction of the variance of exemplar mastery that is explained by the model. Thus, the model explained 67% of the variance of exemplar mastery. The model was significantly predictive of mastery ($F_{27,826}$=61.05, P<.001).

Figure 4 shows the regression lines for male and female patients in each of the different clusters. The mean squared error for each cluster is shown in Table 2.

Figure 4. The line of best fit for each gender and cluster.

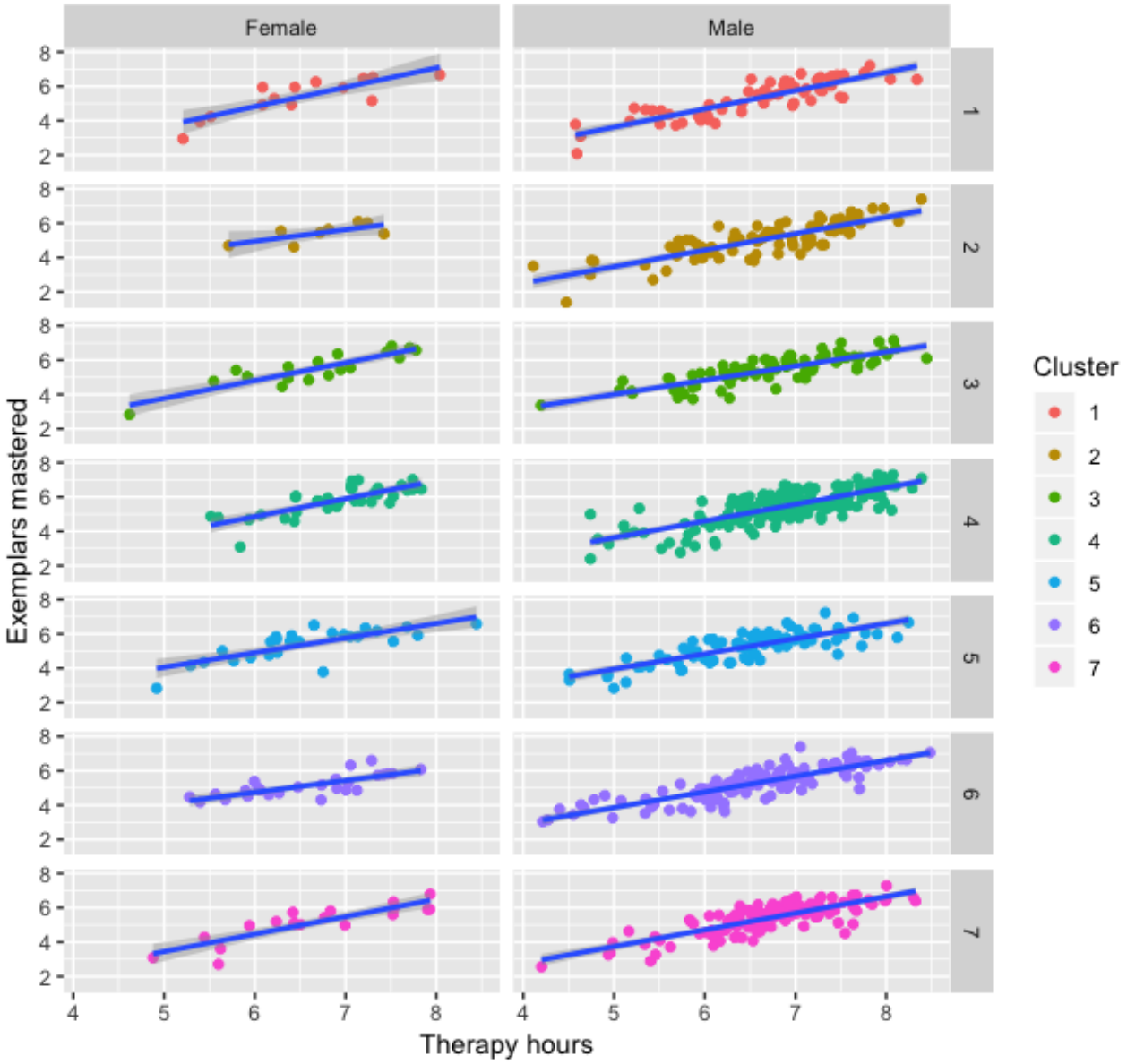
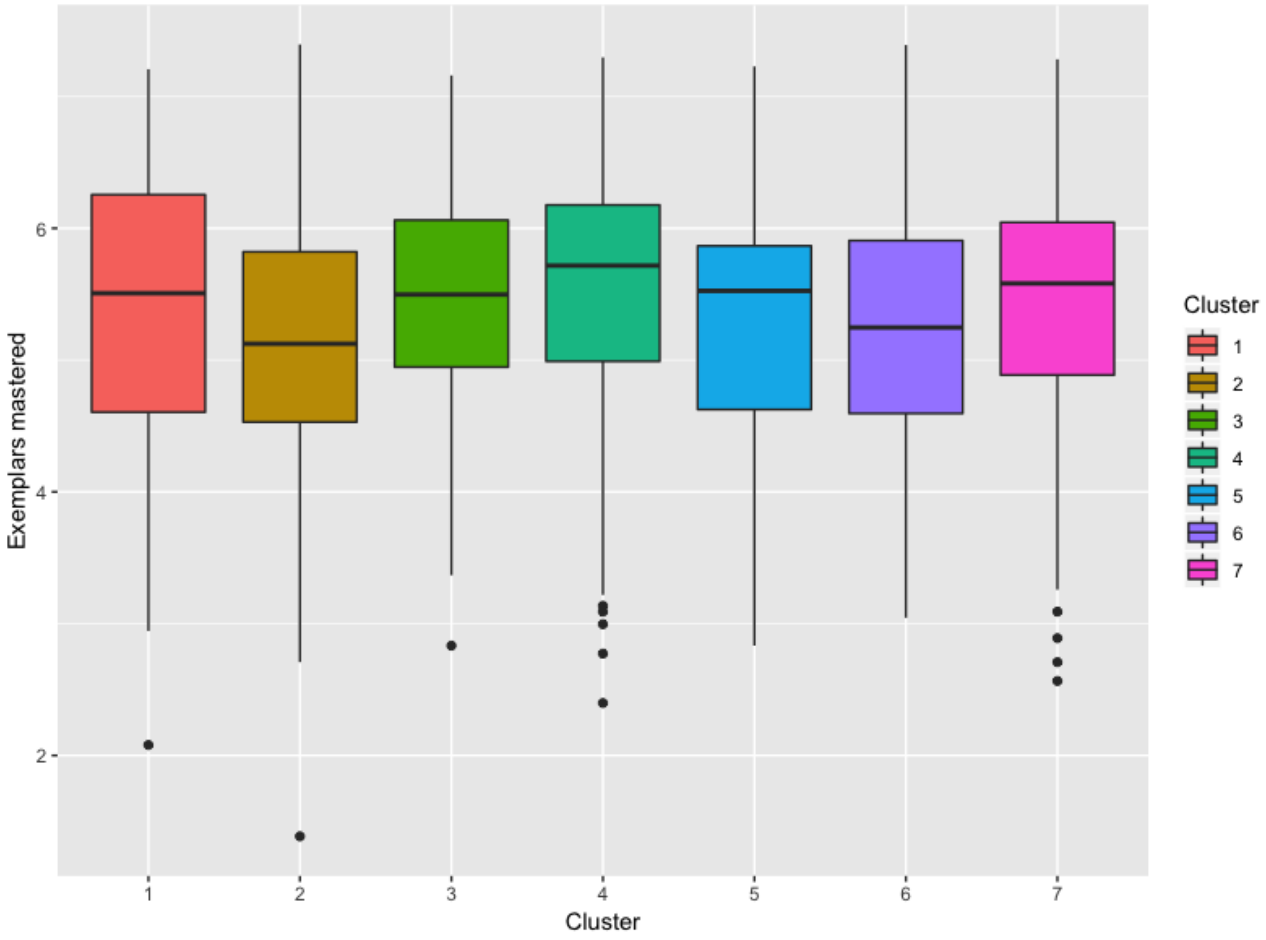


Table 2. Mean squared error comparison across clusters.

Cluster	Dominant challenging behavior	Mean squared error
1	Tantrums	0.30
2	Self-injurious behavior	0.34
3	Elopement	0.23
4	Stereotypy (low rate)	0.37
5	Noncompliance	0.30
6	Aggression	0.24
7	Stereotypy (high rate)	0.29

Box plots (Figure 5) for each of the 7 clusters depicts differences across clusters with respect to exemplar mastery and show the range of the exemplars mastered for each cluster, where the whiskers represent the minimum and maximum values (or 1.5 × the interquartile range, if outliers were present).

Figure 5. Box plots for each cluster. The box plots show the range of the exemplars mastered for each cluster, where the whiskers represent the minimum and maximum values. The line across each box is the median. The top of the box represents the third quartile. The bottom of the box represents the first quartile. Any points on the graph represent outliers in the clusters.



Increased treatment hours were associated with a significant increase in mastery (P<.001), and there were significant differences in mastery between clusters (P=.002); however, the interaction between treatment hours and cluster assignment was not significant (P=.28). Gender was nonsignificant (P=.051); however, the interaction between gender and cluster assignment did have a significant relationship with exemplar mastery (P=.018) (Table 3).

Table 3. Explanatory variables.

Variable	*P* value
Therapy hours	<.001
Gender	.051
Cluster	.002
Therapy hours × gender	.67
Therapy hours × cluster	.28
Gender and cluster	.02
Therapy hours, gender, and cluster	.63

Table 4 shows the averages for treatment hours and exemplars mastered for male, female, and combined clusters. Cluster 4 had the highest and Cluster 3 had the second-highest average number of exemplars mastered. Cluster 2 had the lowest average number of exemplars mastered.

Table 4. Average treatment hours and exemplars mastered across male, female, and combined gender clusters.

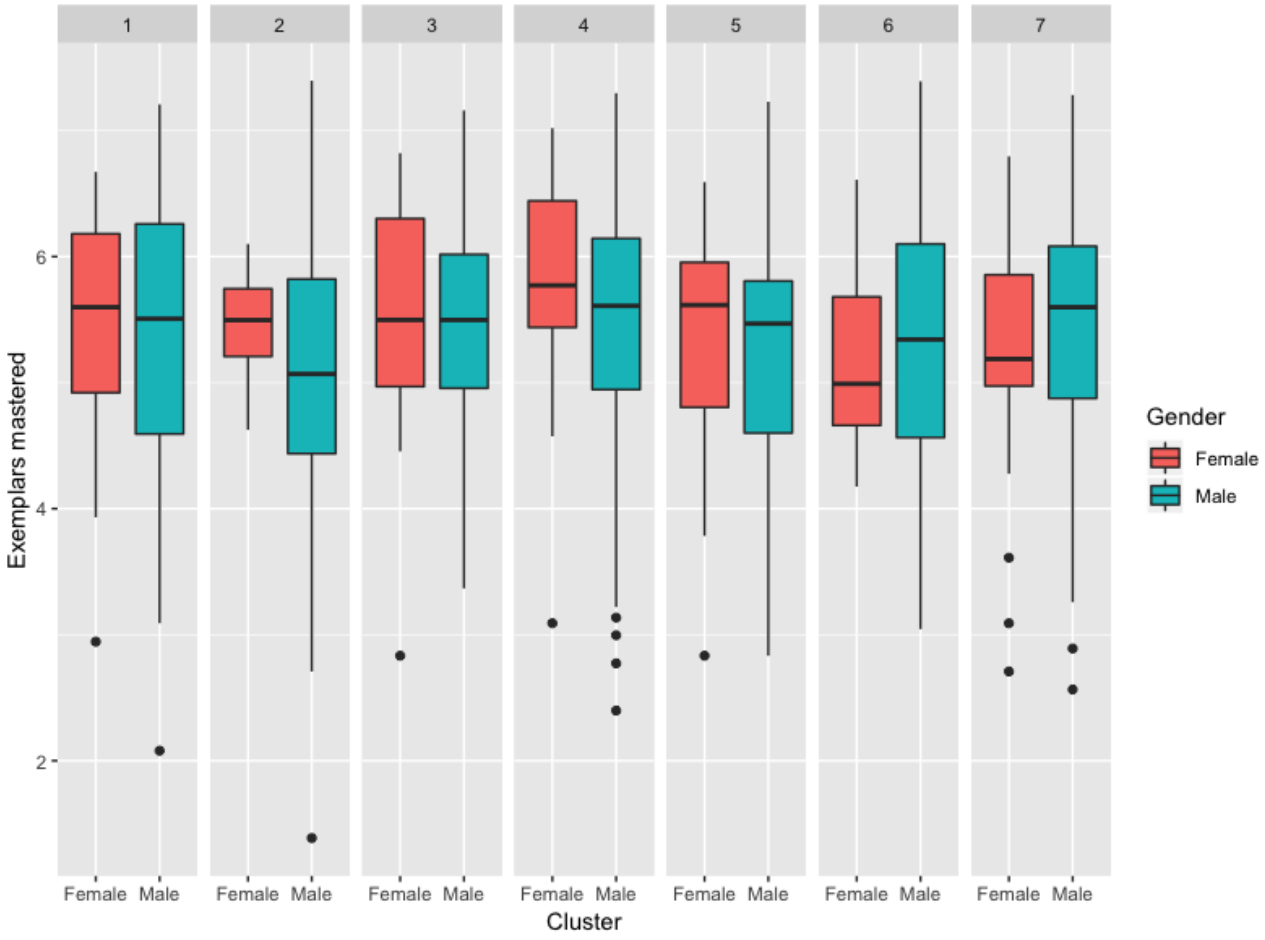
Cluster and dominant challenging behavior	Treatment hours	Exemplars mastered
Male		
Tantrums	6.61	5.33
Self-injurious behavior	6.64	5.05
Elopement	6.71	5.41
Stereotypy (low rate)	6.92	5.50
Noncompliance	6.47	5.26
Aggression	6.54	5.27
Stereotypy (high rate)	6.72	5.41
Female		
Tantrums	6.49	5.37
Self-injurious behavior	6.72	5.43
Elopement	6.66	5.50
Stereotypy (low rate)	6.90	5.80
Noncompliance	6.57	5.39
Aggression	6.58	5.15
Stereotypy (high rate)	6.66	5.13
Combined		
Tantrums	6.58	5.34
Self-injurious behavior	6.65	5.08
Elopement	6.70	5.43
Stereotypy (low rate)	6.91	5.55
Noncompliance	6.50	5.29
Aggression	6.55	5.24
Stereotypy (high rate)	6.71	5.38

Tukey Posthoc

Significant differences were found between Cluster 4 (low frequency stereotypy and moderate frequencies of other challenging behaviors) and Cluster 2 (self-injurious behavior) (P=.003) and between Cluster 4 and Cluster 6 (aggression) (P=.047). Overall, Cluster 4 had the highest rate of mastery while Cluster 2 had the lowest (Table 4); there was a significant difference between the clusters.

The interaction between gender and cluster assignment is depicted in Figure 6. Girls (P=.005) and boys (P=.03) in Cluster 4 mastered significantly more exemplars than the boys in Cluster 2. There was no significant difference between the girls in Cluster 2 and the girls and boys in Cluster 4. There were also no significant differences within clusters between genders (P=.003).

Figure 6. Box plots for each cluster and gender. The whiskers represent the minimum and maximum values. The line across each box is the median. The top of the box represents the third quartile. The bottom of the box represents the first quartile. Any points on the graph represent outliers in the clusters.



Discussion

The purpose of this study was to identify types of autism spectrum disorder based on engagement in 8 challenging behaviors (ie, aggression, disruption, elopement, noncompliance, obsessive behavior, self-injurious behavior, stereotypy, and tantrums) as well as examine group and gender differences in treatment response; *k*-means clustering analyses performed on male, female, and blended samples revealed 7 unique clusters. These findings differ from those of Stevens and colleagues [56], in which 8 male and female clusters were identified based on engagement in challenging behaviors. Similar to those found by Stevens and colleagues [56], the clusters in our study were found to have a single dominant challenging behavior. Only 2 of the measured challenging behaviors (ie, disruption and obsessive behaviors) did not appear as a dominant challenging behavior across the identified clusters. Furthermore, relatively low rates of disruption and obsessive behaviors were also observed across all the clusters. Cluster 1 had tantrums as its dominant challenging behavior, Cluster 2 had self-injurious behavior as its dominant challenging behavior, Cluster 3 had elopement as its dominant challenging behavior, Cluster 4 had stereotypy (low rate compared to cluster 7) as its dominant challenging behavior, Cluster 5 had noncompliance as its dominant challenging behavior, Cluster 6 had aggression as its dominant challenging behavior, and Cluster 7 had stereotypy (at a higher rate than Cluster 4) as its dominant challenging behavior. Neither obsessive behavior nor disruption appeared as a dominant behavior in any of the clusters.

To explore the relationship between skill mastery, treatment hours, cluster assignment, and gender, multiple linear regression was performed. Interactions between all the explanatory variables were also evaluated. In line with previous findings [30,31], the relationship between skill acquisition and treatment hours was found to be significant in our study (P<.001).

In addition to treatment hours, cluster assignment was found to be significantly related to skill mastery (P=.002). Results from the Tukey posthoc test revealed that Cluster 4, characterized by the dominant behavior stereotypy with moderate frequencies of other challenging behaviors, was found to have significantly stronger levels of skill mastery than both Cluster 2, characterized by self-injurious behavior, and Cluster 6, characterized by aggression (P=.003). These findings suggest that treatment response varies across individuals with autism spectrum disorder that engage in different topographies of challenging behaviors. In particular, participants who engaged in self-injurious behavior and aggression were found to have poorer response to treatment compared to those with low levels of stereotypy. It is likely that prioritizing the treatment of self-injurious behavior and aggression using appropriate behavior interventions based on

the identified function of the behavior [63] will result in better long-term treatment outcomes for these individuals.

The only interaction between explanatory variables that was found to be significant in this study was cluster assignment and gender (P=.018). No significant gender differences were found, with respect to skill acquisition, within the same cluster. That is, boys and girls in the same cluster were found to have similar rates of skill acquisition (Table 4). Significant gender differences were found across clusters, however. Specifically, both girls and boys in Cluster 4 (stereotypy) displayed stronger rates of skill mastery than boys in Cluster 2 (self-injurious behavior); however, no significant differences were found between boys and girls in Cluster 4 and girls in Cluster 2. In previous research, gender was found to be a risk factor for the occurrence of challenging behaviors in individuals with autism spectrum disorder [8-12]. While the role of gender is unclear, this finding provides further support for the significant differences in treatment response across clusters, particularly for Cluster 4 and Cluster 2.

This study has several limitations that are important to consider. As a retrospective study, the analysis was limited to the existing data in the data set. Data on race and ethnicity were not available in the data set; therefore, representation across those demographics and any potential disparities in this sample are unknown. Furthermore, variables such as autism spectrum disorder symptom severity and IQ were not measured. Both symptom severity and IQ have been found to be related to engagement in challenging behaviors [2,3,7-9] as well as related to treatment response [27,35,36,38-45]. In particular, aggression and self-injurious behavior, the behaviors associated with slower skill acquisition in our study, have been linked to low IQ scores [8]. It would be worth exploring these variables in future research. In addition, the method used to aggregate the data for clustering results in a relatively small feature space of only 8 dimensions. These dimensions correspond to broad categories of challenging behaviors but do not capture other aspects related to those behaviors such as function. A future study could improve on this work by starting with a higher-dimension behavioral feature space, including functions of behavior, and then utilizing contemporary feature selection algorithms to derive the most meaningful subset of features to be fed to the unsupervised learning algorithm. In this case, use of a clustering algorithm that is more sophisticated may be warranted; the *k*-means algorithm takes a simple approach to clustering that relies upon regularly shaped clusters throughout the feature space. Finally, we note that additional studies to validate the clusters identified here would be valuable. In particular, the use of an additional cohort of participants which could be assigned to clusters and then have their assignments verified by clinicians using a broader set of medical records would be important to verify that the clusters identified here are generalizable beyond the study population.

This study is among the first to investigate types of autism spectrum disorder based on engagement in challenging behaviors and the impact of challenging behaviors on treatment response. Findings suggest that challenging behaviors do impact treatment response with specific topographies (ie, self-injurious behavior, aggression) being particularly detrimental. In future investigations, it would be worthwhile to map the function of the behavior (eg, attention, escape, tangible, automatic), in addition to the topography, and explore its impact on treatment response. Future research should also explore targeted interventions to improve skill acquisition based on cluster assignment, particularly for the clusters characterized by self-injurious behavior and aggression. Until such investigations are conducted, treatment providers should be aware that these behaviors seem to have a particularly negative impact on skill acquisition and interventions addressing these behaviors may be worth prioritizing in treatment. To further improve outcomes across individuals with autism spectrum disorder, attention must be given to segmentation within the disorder. Investigations, such as these, show the promise of unsupervised machine learning models in identifying types of autism spectrum disorder so that targeted treatments based on type membership may be explored. We recommend that clinicians who are interested in further exploring latent structural features of the autism spectrum, including challenging behaviors, proactively collect data to the greatest extent that is practical and unobtrusive. Such data, especially in aggregate, will be essential for gaining additional insights into autism spectrum disorder types with the ultimate goal of personalizing and optimizing treatment plans.

Acknowledgments

CPP was supported by the National Science Foundation (grant GRFP1849569).

Authors' Contributions

EL, ES, and DD conceived the research design. JGH, CPP, and EL conducted the analysis. ES, MN, and JGH performed the literature review. All authors contributed to drafting and revising the manuscript.

Conflicts of Interest

None declared.

References

1. American Psychiatric Association DSM-5 Task Force. Diagnostic and Statistical Manual of Mental Disorders Fifth Edition. Washington, DC: American Psychiatric Publishing Inc; 2013.

2. Jang J, Dixon DR, Tarbox J, Granpeesheh D. Symptom severity and challenging behavior in children with ASD. Res Autism Spectr Disord 2011 Jul;5(3):1028-1032. [doi: 10.1016/j.rasd.2010.11.008]
3. Matson JL, Wilkins J, Macken J. The relationship of challenging behaviors to severity and symptoms of autism spectrum disorders. J Ment Health Res Intellect Disabil 2008 Dec 29;2(1):29-44. [doi: 10.1080/19315860802611415]
4. Matson JL, Nebel-Schwalm M. Assessing challenging behaviors in children with autism spectrum disorders: a review. Res Dev Disabil 2007 Nov;28(6):567-579. [doi: 10.1016/j.ridd.2006.08.001] [Medline: 16973329]
5. Lecavalier L, Leone S, Wiltz J. The impact of behaviour problems on caregiver stress in young people with autism spectrum disorders. J Intellect Disabil Res 2006 Mar;50(Pt 3):172-183 [FREE Full text] [doi: 10.1111/j.1365-2788.2005.00732.x] [Medline: 16430729]
6. Sikora D, Moran E, Orlich F, Hall T, Kovacs E, Delahaye J, et al. The relationship between family functioning and behavior problems in children with autism spectrum disorders. Res Autism Spectr Disord 2013 Feb;7(2):307-315 [FREE Full text] [doi: 10.1016/j.rasd.2012.09.006]
7. Bishop SL, Richler J, Lord C. Association between restricted and repetitive behaviors and nonverbal IQ in children with autism spectrum disorders. Child Neuropsychol 2006 Aug;12(4-5):247-267. [doi: 10.1080/09297040600630288] [Medline: 16911971]
8. McTiernan A, Leader G, Healy O, Mannion A. Analysis of risk factors and early predictors of challenging behavior for children with autism spectrum disorder. Res Autism Spectr Disord 2011 Jul;5(3):1215-1222. [doi: 10.1016/j.rasd.2011.01.009]
9. Murphy O, Healy O, Leader G. Risk factors for challenging behaviors among 157 children with autism spectrum disorder in Ireland. Res Autism Spectr Disord 2009 Apr;3(2):474-482 [FREE Full text] [doi: 10.1016/j.rasd.2008.09.008]
10. Baghdadli A, Pascal C, Grisi S, Aussilloux C. Risk factors for self-injurious behaviours among 222 young children with autistic disorders. J Intellect Disabil Res 2003 Nov;47(Pt 8):622-627. [doi: 10.1046/j.1365-2788.2003.00507.x] [Medline: 14641810]
11. Hartley S, Sikora D, McCoy R. Prevalence and risk factors of maladaptive behaviour in young children with autistic disorder. J Intellect Disabil Res 2008 Oct;52(10):819-829 [FREE Full text] [doi: 10.1111/j.1365-2788.2008.01065.x] [Medline: 18444989]
12. Kozlowski AM, Matson JL, Rieske RD. Gender effects on challenging behaviors in children with autism spectrum disorders. Res Autism Spectr Disord 2012 Apr;6(2):958-964. [doi: 10.1016/j.rasd.2011.12.011]
13. Cooper J, Heron T, Heward W. Applied Behavior Analysis, Second Edition. Upper Saddle River, NJ: Pearson; 2007.
14. Granpeesheh D, Tarbox J, Dixon DR. Applied behavior analytic interventions for children with autism: a description and review of treatment research. Ann Clin Psychiatry 2009;21(3):162-173. [Medline: 19758537]
15. Hong E, Dixon DR, Stevens E, Burns CO, Linstead E. Topography and function of challenging behaviors in individuals with autism spectrum disorder. Adv Neurodev Disord 2018 Apr 30;2(2):206-215. [doi: 10.1007/s41252-018-0063-7]
16. Machalicek W, Raulston T, Knowles C, Ruppert T, Carnett A, Alresheed F. Challenging behavior. In: Matson J, editor. Comorbid Conditions Among Children with Autism Spectrum Disorders. New York: Springer; Oct 2015.
17. Bearss K, Johnson C, Smith T, Lecavalier L, Swiezy N, Aman M, et al. Effect of parent training vs parent education on behavioral problems in children with autism spectrum disorder: a randomized clinical trial. JAMA 2015 Apr 21;313(15):1524-1533 [FREE Full text] [doi: 10.1001/jama.2015.3150] [Medline: 25898050]
18. Aman M, Mcdougle C, Scahill L, Handen B, Arnold L, Johnson C, et al. Medication and parent training in children with pervasive developmental disorders and serious behavior problems: results from a randomized clinical trial. J Am Acad Child Adolesc Psychiatry 2009 Dec;48(12):1143-1154 [FREE Full text] [doi: 10.1097/chi.0b013e3181bfd669]
19. Lindgren S, Wacker D, Suess A, Schieltz K, Pelzel K, Kopelman T, et al. Telehealth and autism: treating challenging behavior at lower cost. Pediatrics 2016 Feb;137 Suppl 2:S167-S175 [FREE Full text] [doi: 10.1542/peds.2015-2851O] [Medline: 26908472]
20. Campbell JM. Efficacy of behavioral interventions for reducing problem behavior in persons with autism: a quantitative synthesis of single-subject research. Res Dev Disabil 2003;24(2):120-138. [doi: 10.1016/s0891-4222(03)00014-3] [Medline: 12623082]
21. Didden R, Korzilius H, van Oorsouw W, Sturmey P. Behavioral treatment of challenging behaviors in individuals with mild mental retardation: meta-analysis of single-subject research. Am J Ment Retard 2006 Jul;111(4):290-298. [doi: 10.1352/0895-8017(2006)111[290:BTOCBI]2.0.CO;2] [Medline: 16792430]
22. Heyvaert M, Saenen L, Campbell JM, Maes B, Onghena P. Efficacy of behavioral interventions for reducing problem behavior in persons with autism: an updated quantitative synthesis of single-subject research. Res Dev Disabil 2014 Oct;35(10):2463-2476. [doi: 10.1016/j.ridd.2014.06.017] [Medline: 24992447]
23. Rogers SJ, Vismara LA. Evidence-based comprehensive treatments for early autism. J Clin Child Adolesc Psychol 2008 Jan 03;37(1):8-38 [FREE Full text] [doi: 10.1080/15374410701817808] [Medline: 18444052]
24. Smith T, Iadarola S. Evidence base update for autism spectrum disorder. J Clin Child Adolesc Psychol 2015 Oct 02;44(6):897-922. [doi: 10.1080/15374416.2015.1077448] [Medline: 26430947]
25. Eldevik S, Hastings RP, Hughes JC, Jahr E, Eikeseth S, Cross S. Meta-analysis of early intensive behavioral intervention for children with autism. J Clin Child Adolesc Psychol 2009 May 19;38(3):439-450. [doi: 10.1080/15374410902851739] [Medline: 19437303]

26. Reichow B. Overview of meta-analyses on early intensive behavioral intervention for young children with autism spectrum disorders. J Autism Dev Disord 2012 Apr 15;42(4):512-520. [doi: 10.1007/s10803-011-1218-9] [Medline: 21404083]
27. Eldevik S, Hastings R, Hughes JC, Jahr E, Eikeseth S, Cross S. Using participant data to extend the evidence base for intensive behavioral intervention for children with autism. Am J Intellect Dev Disabil 2010 Sep;115(5):381-405. [doi: 10.1352/1944-7558-115.5.381] [Medline: 20687823]
28. Howlin P, Magiati I, Charman T. Systematic review of early intensive behavioral interventions for children with autism. Am J Intellect Dev Disabil 2009;114(1):23-41. [doi: 10.1352/2009.114:23;nd41]
29. Granpeesheh D, Dixon DR, Tarbox J, Kaplan AM, Wilke AE. The effects of age and treatment intensity on behavioral intervention outcomes for children with autism spectrum disorders. Res Autism Spectr Disord 2009 Oct;3(4):1014-1022. [doi: 10.1016/j.rasd.2009.06.007]
30. Linstead E, Dixon DR, French R, Granpeesheh D, Adams H, German R, et al. Intensity and learning outcomes in the treatment of children with autism spectrum disorder. Behav Modif 2017 Mar 21;41(2):229-252. [doi: 10.1177/0145445516667059] [Medline: 27651097]
31. Linstead E, Dixon DR, Hong E, Burns CO, French R, Novack MN, et al. An evaluation of the effects of intensity and duration on outcomes across treatment domains for children with autism spectrum disorder. Transl Psychiatry 2017 Sep 19;7(9):e1234-e1234 [FREE Full text] [doi: 10.1038/tp.2017.207] [Medline: 28925999]
32. Makrygianni M, Reed P. A meta-analytic review of the effectiveness of behavioural early intervention programs for children with Autistic Spectrum Disorders. Res Autism Spectr Disord 2010 Oct;4(4):577-593 [FREE Full text] [doi: 10.1016/j.rasd.2010.01.014]
33. Virués-Ortega J. Applied behavior analytic intervention for autism in early childhood: meta-analysis, meta-regression and dose-response meta-analysis of multiple outcomes. Clin Psychol Rev 2010 Jun;30(4):387-399 [FREE Full text] [doi: 10.1016/j.cpr.2010.01.008] [Medline: 20223569]
34. Virués-Ortega J, Rodríguez V, Yu C. Prediction of treatment outcomes and longitudinal analysis in children with autism undergoing intensive behavioral intervention. Int J Clin Health Psychol 2013 May;13(2):91-100 [FREE Full text] [doi: 10.1016/s1697-2600(13)70012-7]
35. Ben Itzchak E, Zachor D. Who benefits from early intervention in autism spectrum disorders? Res Autism Spectr Disord 2011 Jan;5(1):345-350 [FREE Full text] [doi: 10.1016/j.rasd.2010.04.018]
36. Eldevik S, Hastings R, Jahr E, Hughes JC. Outcomes of behavioral intervention for children with autism in mainstream pre-school settings. J Autism Dev Disord 2012 Feb;42(2):210-220 [FREE Full text] [doi: 10.1007/s10803-011-1234-9] [Medline: 21472360]
37. Flanagan H, Perry A, Freeman N. Effectiveness of large-scale community-based Intensive Behavioral Intervention: a waitlist comparison study exploring outcomes and predictors. Res Autism Spectr Disord 2012 Apr;6(2):673-682 [FREE Full text] [doi: 10.1016/j.rasd.2011.09.011]
38. Perry A, Cummings A, Geier J, Freeman N, Hughes S, Managhan T, et al. Predictors of outcome for children receiving intensive behavioral intervention in a large, community-based program. Res Autism Spectr Disord 2011 Jan;5(1):592-603 [FREE Full text] [doi: 10.1016/j.rasd.2010.07.003]
39. Remington B, Hastings R, Kovshoff H, degli Espinosa F, Jahr E, Brown T, et al. Early intensive behavioral intervention: outcomes for children with autism and their parents after two years. Am J Ment Retard 2007 Nov;112(6):418-438. [doi: 10.1352/0895-8017(2007)112[418:EIBIOF]2.0.CO;2] [Medline: 17963434]
40. Ben-Itzchak E, Zachor D. The effects of intellectual functioning and autism severity on outcome of early behavioral intervention for children with autism. Res Dev Disabil 2007;28(3):287-303 [FREE Full text] [doi: 10.1016/j.ridd.2006.03.002] [Medline: 16730944]
41. Eikeseth S, Smith T, Jahr E, Eldevik S. Intensive behavioral treatment at school for 4- to 7-year-old children with autism. a 1-year comparison controlled study. Behav Modif 2002 Jan;26(1):49-68 [FREE Full text] [doi: 10.1177/0145445502026001004] [Medline: 11799654]
42. Eikeseth S, Smith T, Jahr E, Eldevik S. Outcome for children with autism who began intensive behavioral treatment between ages 4 and 7: a comparison controlled study. Behav Modif 2007 May;31(3):264-278 [FREE Full text] [doi: 10.1177/0145445506291396] [Medline: 17438342]
43. Hayward D, Eikeseth S, Gale C, Morgan S. Assessing progress during treatment for young children with autism receiving intensive behavioural interventions. Autism 2009 Nov;13(6):613-633 [FREE Full text] [doi: 10.1177/1362361309340029] [Medline: 19933766]
44. Magiati I, Charman T, Howlin P. A two-year prospective follow-up study of community-based early intensive behavioural intervention and specialist nursery provision for children with autism spectrum disorders. J Child Psychol Psychiatry 2007 Aug;48(8):803-812 [FREE Full text] [doi: 10.1111/j.1469-7610.2007.01756.x] [Medline: 17683452]
45. Magiati I, Moss J, Charman T, Howlin P. Patterns of change in children with autism spectrum disorders who received community based comprehensive interventions in their pre-school years: A seven year follow-up study. Res Autism Spectr Disord 2011 Jul;5(3):1016-1027 [FREE Full text] [doi: 10.1016/j.rasd.2010.11.007]

46. Eikeseth S, Klintwall L, Jahr E, Karlsson P. Outcome for children with autism receiving early and intensive behavioral intervention in mainstream preschool and kindergarten settings. Res Autism Spectr Disord 2012 Apr;6(2):829-835 [FREE Full text] [doi: 10.1016/j.rasd.2011.09.002]
47. Beglinger L, Smith T. A review of subtyping in autism and proposed dimensional classification model. J Autism Dev Disord 2001 Aug;31(4):411-422. [doi: 10.1023/a:1010616719877] [Medline: 11569587]
48. Bruining H, de Sonneville L, Swaab H, de Jonge M, Kas M, van Engeland H, et al. Dissecting the clinical heterogeneity of autism spectrum disorders through defined genotypes. PLoS One 2010 May 28;5(5):e10887-e10887 [FREE Full text] [doi: 10.1371/journal.pone.0010887] [Medline: 20526357]
49. Hu V, Lai Y. Developing a predictive gene classifier for autism spectrum disorders based upon differential gene expression profiles of phenotypic subgroups. N Am J Med Sci (Boston) 2013;6(3):1-18 [FREE Full text] [doi: 10.7156/najms.2013.0603107] [Medline: 24363828]
50. Veatch O, Veenstra-Vanderweele J, Potter M, Pericak-Vance MA, Haines JL. Genetically meaningful phenotypic subgroups in autism spectrum disorders. Genes Brain Behav 2014 Mar;13(3):276-285 [FREE Full text] [doi: 10.1111/gbb.12117] [Medline: 24373520]
51. Fein D, Stevens M, Dunn M, Waterhouse L, Allen D, Rapin I, et al. Subtypes of pervasive developmental disorder: clinical characteristics. Child Neuropsychol 2010 Aug 09;5(1):1-23 [FREE Full text] [doi: 10.1076/chin.5.1.1.7075]
52. Lord C, Bishop S, Anderson D. Developmental trajectories as autism phenotypes. Am J Med Genet C Semin Med Genet 2015 Jun;169(2):198-208 [FREE Full text] [doi: 10.1002/ajmg.c.31440] [Medline: 25959391]
53. Pickles A, Anderson D, Lord C. Heterogeneity and plasticity in the development of language: a 17-year follow-up of children referred early for possible autism. J Child Psychol Psychiatry 2014 Dec;55(12):1354-1362 [FREE Full text] [doi: 10.1111/jcpp.12269] [Medline: 24889883]
54. Stevens M, Fein D, Dunn M, Allen D, Waterhouse LH, Feinstein C, et al. Subgroups of children with autism by cluster analysis: a longitudinal examination. J Am Acad Child Adolesc Psychiatry 2000 Mar;39(3):346-352 [FREE Full text] [doi: 10.1097/00004583-200003000-00017] [Medline: 10714055]
55. Stevens E, Dixon D, Novack MN, Granpeesheh D, Smith T, Linstead E. Identification and analysis of behavioral phenotypes in autism spectrum disorder via unsupervised machine learning. Int J Med Inform 2019 Sep;129:29-36 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.05.006] [Medline: 31445269]
56. Stevens E, Atchison A, Stevens L, Hong E, Granpeesheh D, Dixon D, et al. A cluster analysis of challenging behaviors in autism spectrum disorder. 2017 Dec Presented at: 16th IEEE International Conference on Machine Learning and Applications; Devember 18-21; Cancun, Mexico p. 661-666 URL: https://doi.org/10.1109/ICMLA.2017.00-85 [doi: 10.1109/ICMLA.2017.00-85]
57. Skills software. Skills Global LLC. URL: https://www.skillsforautism.com/ [accessed 2019-02-01]
58. Dixon D, Tarbox J, Najdowski A, Wilke A, Granpeesheh D. A comprehensive evaluation of language for early behavioral intervention programs: the reliability of the SKILLS language index. Res Autism Spectr Disord 2011 Jan;5(1):506-511 [FREE Full text] [doi: 10.1016/j.rasd.2010.06.016]
59. Persicke A, Bishop M, Coffman C, Najdowski A, Tarbox J, Chi K, et al. Evaluation of the concurrent validity of a skills assessment for autism treatment. Res Autism Spectr Disord 2014 Mar;8(3):281-285 [FREE Full text] [doi: 10.1016/j.rasd.2013.12.011]
60. Granpeesheh D, Tarbox J, Najdowski A, Kornack J. Evidence-Based Treatment for Children with Autism: The CARD Model. New York: Elsevier; 2015.
61. Raschka S. Python Machine Learning. Birmingham, United Kingdom: Packt Publishing; 2015.
62. Teetor P. R Cookbook: Proven Recipes for Data Analysis, Statistics, and Graphics. Sebastopol, CA: O'Reilly Media; 2011.
63. Napolitano D, Knapp V, Speares E, McAdam D, Brown H. The role of functional assessment in treatment planning. In: Matson JL, editor. Functional Assessment for Challenging Behaviors. New York: Springer; 2012:195-211.

Edited by G Eysenbach; submitted 06.02.21; peer-reviewed by S Iadarola, Q Zhan; comments to author 01.03.21; revised version received 23.04.21; accepted 29.04.21; published 02.06.21.

Please cite as:
Gardner-Hoag J, Novack M, Parlett-Pelleriti C, Stevens E, Dixon D, Linstead E
Unsupervised Machine Learning for Identifying Challenging Behavior Profiles to Explore Cluster-Based Treatment Efficacy in Children With Autism Spectrum Disorder: Retrospective Data Analysis Study
JMIR Med Inform 2021;9(6):e27793
URL: https://medinform.jmir.org/2021/6/e27793
doi:10.2196/27793
PMID:34076577

Original Paper

Informing Developmental Milestone Achievement for Children With Autism: Machine Learning Approach

Munirul M Haque[1*], PhD; Masud Rabbani[2*], BSc; Dipranjan Das Dipal[2*], MSc; Md Ishrak Islam Zarif[2*], BSc; Anik Iqbal[2*], PhD; Amy Schwichtenberg[3*], PhD; Naveen Bansal[4*], PhD; Tanjir Rashid Soron[5*], MD; Syed Ishtiaque Ahmed[6*], PhD; Sheikh Iqbal Ahamed[2*], PhD

[1]R.B. Annis School of Engineering, University of Indianapolis, Indianapolis, IN, United States

[2]Ubicomp Lab, Department of Computer Science, Marquette University, Milwaukee, WI, United States

[3]College of Health and Human Sciences, Purdue University, West Lafayette, IN, United States

[4]Department of Mathematical and Statistical Sciences, Marquette University, Milwaukee, WI, United States

[5]Telepsychiatry Research and Innovation Network Ltd, Dhaka, Bangladesh

[6]Department of Computer Science, University of Toronto, Toronto, ON, Canada

[*]all authors contributed equally

Corresponding Author:
Masud Rabbani, BSc
Ubicomp Lab, Department of Computer Science
Marquette University
1422 W Kilbourn Ave
102
Milwaukee, WI, 53233-1784
United States
Phone: 1 4143267769
Email: masud.rabbani@marquette.edu

Abstract

Background: Care for children with autism spectrum disorder (ASD) can be challenging for families and medical care systems. This is especially true in low- and- middle-income countries such as Bangladesh. To improve family–practitioner communication and developmental monitoring of children with ASD, mCARE (Mobile-Based Care for Children with Autism Spectrum Disorder Using Remote Experience Sampling Method) was developed. Within this study, mCARE was used to track child milestone achievement and family sociodemographic assets to inform mCARE feasibility/scalability and family asset–informed practitioner recommendations.

Objective: The objectives of this paper are threefold. First, it documents how mCARE can be used to monitor child milestone achievement. Second, it demonstrates how advanced machine learning models can inform our understanding of milestone achievement in children with ASD. Third, it describes family/child sociodemographic factors that are associated with earlier milestone achievement in children with ASD (across 5 machine learning models).

Methods: Using mCARE-collected data, this study assessed milestone achievement in 300 children with ASD from Bangladesh. In this study, we used 4 supervised machine learning algorithms (decision tree, logistic regression, K-nearest neighbor [KNN], and artificial neural network [ANN]) and 1 unsupervised machine learning algorithm (K-means clustering) to build models of milestone achievement based on family/child sociodemographic details. For analyses, the sample was randomly divided in half to train the machine learning models and then their accuracy was estimated based on the other half of the sample. Each model was specified for the following milestones: *Brushes teeth, Asks to use the toilet, Urinates in the toilet or potty, and Buttons large buttons*.

Results: This study aimed to find a suitable machine learning algorithm for milestone prediction/achievement for children with ASD using family/child sociodemographic characteristics. For *Brushes teeth*, the 3 supervised machine learning models met or exceeded an accuracy of 95% with logistic regression, KNN, and ANN as the most robust sociodemographic predictors. For *Asks to use toilet,* 84.00% accuracy was achieved with the KNN and ANN models. For these models, the family sociodemographic predictors of “family expenditure” and “parents’ age” accounted for most of the model variability. The last 2 parameters, *Urinates in toilet or potty* and *Buttons large buttons*, had an accuracy of 91.00% and 76.00%, respectively, in ANN. Overall, the ANN

had a higher accuracy (above ~80% on average) among the other algorithms for all the parameters. Across the models and milestones, “family expenditure,” “family size/type,” “living places,” and “parent’s age and occupation” were the most influential family/child sociodemographic factors.

Conclusions: mCARE was successfully deployed in a low- and middle-income country (ie, Bangladesh), providing parents and care practitioners a mechanism to share detailed information on child milestones achievement. Using advanced modeling techniques this study demonstrates how family/child sociodemographic elements can inform child milestone achievement. Specifically, families with fewer sociodemographic resources reported later milestone attainment. Developmental science theories highlight how family/systems can directly influence child development and this study provides a clear link between family resources and child developmental progress. Clinical implications for this work could include supporting the larger family system to improve child milestone achievement.

(JMIR Med Inform 2021;9(6):e29242) doi:10.2196/29242

KEYWORDS

autism spectrum disorders; machine learning; digital health; mobile health; mhealth; predictive modeling; milestone parameters; Autism and Developmental Disabilities Monitoring (ADDM); early intervention

Introduction

Background

Autism spectrum disorder (ASD) is a global problem [1] and a heterogeneous neurodevelopmental disorder [2]. In 1943, Kanner [3] first described this disorder in children’s behavior [3]. In this neurodevelopmental disorder, children have social communication issues, repetitive behaviors, restrictive interests, and professional impairments throughout their lifespan [4,5]. In developed countries, 1%-1.5% of children have ASD [4], whereas in the United States, 1 out of 54 children have ASD [6,7]. Although it is the fastest growing developmental disorder, the number of individuals affected globally remains largely unknown [8]. In low- and middle-income countries, this rate is estimated to vary between 0.15% and 0.8%, whereas in a developing country such as Bangladesh this rate is reported to be 3% [9-11]. ASD symptoms gradually show up before 1 year of age, with nearly 80% of problems being identified by 2 years of age [12,13]. In particular, boys are affected 3 to 4 times more than girls with ASD [14]. Unfortunately, nearly 46% of children with ASD do not receive the proper treatment following diagnosis [8].

Medically, early identification and diagnosis of ASD will improve positive functional outcomes in later life for these children [15-18]. As a result, in 2000, the American Academy of Neurology and Child Neurology recommended to screen every child for ASD [14,19-21]. In other words, a reliable ASD diagnosis should be performed in children before 24 months of age [19], as this substantially improves the opportunities for recovery and also reduces the burden on caregivers (diagnostic odyssey) [16]. The major barriers to making improvements in ASD diagnosis and treatment are lack of proper knowledge about ASD, lack of motivation and patience of parents or caregivers, and delayed identification and diagnosis of ASD. Early identification and diagnosis help the care practitioners to make evidence-based decisions during intervention, which has both positive and long-term outcomes on the improvement of patients with ASD [5,22]. Physical therapy or exercise is much more important than medicine in the development of many patients with ASD, and in such cases early intervention can play an important role [23-26].

Besides the early identification and diagnosis, parents’ or caregivers’ demography, social or environmental demography, race, and ethnicity can play a vital role in the developmental process of children with ASD [15,27-31]. Concerning parents’ demography, educational level, occupation, family income and expenditures, number of siblings, and living area remain very important factors in the development of children with ASD [27-30]. Environmental factors such as the socioeconomic condition, neighborhood, and society’s attitudes toward children with ASD are very significant [12,13]. Although genes increasing the risk for ASD in children are mostly prenatal [32], demography of parents remains very important [33], as it can affect the improvement of patients with ASD. In this study, we will use the parents’, environmental, and social demography as a parameter to develop a machine learning model for predicting the improvement level of milestone parameters in children with ASD.

Based on the demography, machine learning models can predict the milestone parameters in children with ASD during their early intervention period. In this study, we have used 10 important demographic information in 4 supervised machine learning models to predict the improvement level of “daily living skills.” In the “Decision Tree” [34] machine learning algorithm, we have used the “Classification Trees” category to build the predictive model. To build a statistical model for our binary dependent variable, we deployed “Logistic Regression” with the sigmoid function [35,36] as the logistic function. We then deployed our preprocessed data sets in the K-nearest neighbor (KNN) algorithm using the “Euclidean distance” [37] to find the nearest neighbor. In the end, we used an artificial neural network (ANN) to build our last predictive model. In ANN, we have used “relu” as the hidden layer’s activation function, and “sigmoid” as the output layer’s activation function.

Prior Work

In our previous work (Mobile-Based Care for Children with Autism Spectrum Disorder Using Remote Experience Sampling Method [mCARE]), we developed a mobile-based system to regularly monitor children with ASD with the help of caregivers in Bangladesh. In mCARE, we deployed a remote experience sampling method to monitor the milestone and behavioral parameters. These longitudinal data can be used in the

intervention process, where the care practitioners can make evidence-based decisions based on the data. This tool was very effective in the development process of children with ASD; using this tool, the caregiver and care practitioner can observe the improvement level over a certain period on a graphical view. This tool not only assists the care practitioners but also motivates the caregivers. Besides, this tool has some renowned applications and studies to assist with the ASD diagnosis process in different phases [2,15,19,38,39]. While most studies have been performed for the early identification or recognition of ASD [16,19,22,40-44], little work has been done so far on the prediction of improvement level of ASD parameters or the timeframe for a certain level of improvement, or on the factors that need to be improved. In this study, we developed a relationship between the parents' demography and the improvement in ASD milestone parameters by deploying a real data set of the mCARE system.

Goal of This Study

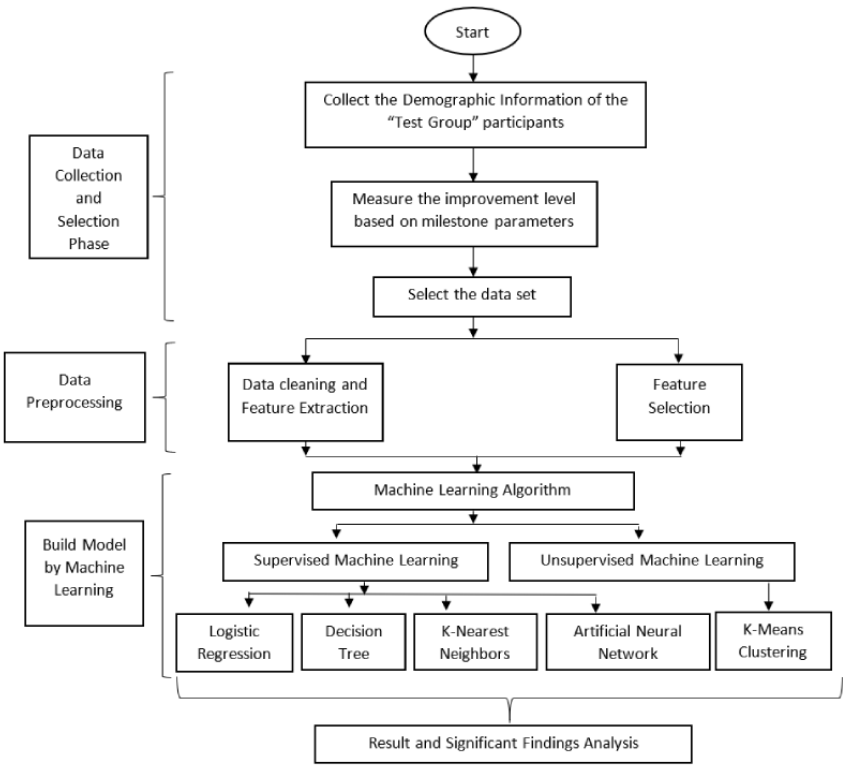
Demographic data such as family income, living place, facilities, parents' age, education and occupation, family types, and number of siblings affect parental stress and psychology [45]. This parental stress and psychological stress definitely impact the mental development of children with ASD, especially "daily living skills" [46-48]. For this reason, cognitive behavioral therapy is very effective on the daily living skill development of children with ASD [49]. In this study, our main goal was to predict the improvement in an ASD milestone parameter (ie, daily living skills) using a machine learning algorithm based on demographic data of caregivers. To achieve our goal, first, we measure the improvement level of the milestone in children with ASD from the mCARE tools. Second, we will deploy an mCARE data set in 5 supervised and 1 unsupervised machine learning algorithm to build the best milestone parameters improvement prediction model. Finally, we will describe the importance of the caregiver-specific demography in predicting the improvement level of certain milestone parameters in children with ASD.

Methods

mCARE

mCARE is a mobile-based app for monitoring the milestone and behavioral parameters of children with ASD regularly and remotely. This project was awarded by the National Institutes of Health (NIH) [50] and has been implemented in Bangladesh for 2 years. For this study, we used data from the mCARE study, which was approved by the Institutional Review Board of the Marquette University on July 9, 2020 (protocol number HR-1803022959). The mCARE study recruited 316 participants, of which we recruited 300 for this study. We deployed the remote experience sampling method to collect data on children with ASD, which was achieved by their caregiver using a smartphone app or an SMS text message. This mobile-based app has significance in the mental health intervention process, where by using the mCARE: Data Management Portal (mCARE: DMP), a caregiver can observe the longitudinal behavioral or milestone data graphically for a certain period. This feature helped the caregiver to make evidence-based decisions in the intervention process. In this study, we will first measure the improvement level of the "test group" participants based on milestone parameters. Using the test group data set, we will build the machine learning–based prediction model for a specific milestone parameter. We will use the test group patients' demography for constructing the prediction model. Figure 1 summarizes the research design in a simple flowchart.

Figure 1. Outline of research design.



Data Collection and Selection Phase

Following approval from the Marquette University Institutional Review Board (Protocol number HR-1803022959), the mCARE project recruited a total of 300 children with ASD (aged 2-9) from Bangladesh. We incorporated diversity in terms of age, sex, ASD severity, and family socioeconomic recourses. We divided the whole sample population into 2 groups: (1) the test group and (2) the control group. Patients in the test group were intervened and monitored regularly, whereas those in the control group were monitored over a certain period. Data from the control group and the test group were compared. This study took place in 4 major institutes of Bangladesh located in 2 geographical locations (Dhaka and Chittagong). We collaborated with 2 government organizations for ASD treatment and research, namely, The National Institute of Mental Health (NIMH) [51] and The Institute of Pediatric Neuro-disorder & Autism (IPNA) [52], to recruit 100 caregivers of children with ASD from each. The participants from each organization were divided into 2 groups: mCARE-APP (n=50) and mCARE-SMS (n=50). Each group was further divided equally into the test (n=25) and control (n=25) groups. Typically, in Bangladesh, families with low and high socioeconomic status receive treatment from public and private organizations, respectively. Therefore, to include participants from all socioeconomic classes, we included 2 private organizations, namely, Nishpap [53] and Autism Welfare Foundation (AWF) [54]. A total of 50 participants chosen from each of these schools were divided into the test group (n=25) and the control group (n=25) only for the mCARE-APP study group. The patient distribution among the 4 centers and the participant demography are presented in Tables 1 and 2, respectively.

Table 1. Patient distribution among the 4 centers.

Serial	Center name	Patients distribution	
		Test group (n=150)	Control group (n=150)
1	The National Institute of Mental Health (NIMH)	50	50
2	The Institute of Pediatric Neuro-disorder & Autism (IPNA)	50	50
3	Autism Welfare Foundation (AWF)	25	25
4	Nishpap Autism Foundation	25	25

Table 2. Demographic information of participants in the test group (n=150).

Demographics	mCARE: test group, n (%)
Age (years)	
2-6	37 (24.7)
6-9	113 (75.3)
Sex	
Male	124 (82.7)
Female	26 (17.3)
Education of children	
Never went to school	34 (22.7)
Went to usual academic school but failed to continue study	22 (14.7)
Went to specialized school but failed to continue study	4 (2.7)
Currently he/she is going to usual academic school	12 (8.0)
Currently he/she is going to specialized academic school	78 (52.0)
Father's education	
Primary	29 (19.3)
Secondary	23 (15.3)
Undergraduate	23 (15.3)
Graduate	29 (19.3)
Postgraduate	46 (30.7)
Mother's education	
Primary	19 (12.7)
Secondary	37 (24.7)
Undergraduate	25 (16.7)
Graduate	32 (21.3)
Postgraduate	37 (24.7)
Student	0.0 (0.0)
Unemployed	4 (2.7)
Father's occupation	
Service	70 (46.7)
Business	45 (30.0)
Cultivation	1 (0.7)
Other	7 (4.7)
Unemployed	23 (15.3)
Mother's occupation	
Student	0.0 (0.0)
Unemployed	0.0 (0.0)
Housewife	124 (82.7)
Service	17 (11.3)
Business	4 (2.7)
Cultivation	0 (0.0)
Maid	1 (0.7)
Other	1 (0.7)
Not applied	3 (2.0)

Demographics	mCARE: test group, n (%)
Average family spending per month (in thousand Taka)[a]	
<15 K	19 (12.7)
15-30 K	44 (29.3)
30-50 K	31 (20.7)
>50 K	56 (37.3)
Family type	
Nuclear	113 (75.3)
Extended	37 (24.7)
Geographic location	
Urban	120 (80.0)
Semiurban	15 (10.0)
Rural	15 (10.0)
Slum	0.0 (0.0)

[a]US $1=84.77 Taka (as of March 18, 2021).

Demographic Information of the Participants in the "Test Group"

We collected demographic information about participants in the test group (n=150). In Table 2, we present in detail the demographic information of participants that took part in the mCARE study.

Measuring the Improvement Level Based on Milestone Parameters

In the mCARE project, there were 4 types of milestone for every test group patient. These were "daily living skills," "communication," "motor skills," and "socialization." Further, for every patient, based on his/her condition, the recruited care practitioner set different types of parameter from every milestone category. Table 3 lists the 4 types of parameters from each milestone group along with the participant numbers (n). Here the participant number (n) is different for different milestone parameters, as every participant did not have the same milestone parameter initially set by the care practitioner. At the beginning of this project, the care practitioners obtained the baseline information for every milestone parameter by screening the participant. Then, in the project timeline (2 years), the caregiver continuously updated the milestone parameter using the mCARE: APP or mCARE: SMS tool based on the child's condition. At the end of the project, one can generate the participant's end improvement level for different levels of their milestone parameters. By comparing the baseline milestone data with the end participant's improvement data, we can calculate the improvement level (in percentage) for every milestone parameter (described in Table 3). In this table, besides the improvement level, we calculated the 95% CI for the validation of our results. As our sample size was 150, we used the Z value (1.96 for 95% CI) [55] for calculating the 95% CI using the following formula:

where  is the mean, Z is 1.96 (chosen from the Z-value table [55]), S is the SD, and n is the average sample number.

Table 3. Improvement level of the test group (mCARE) on their milestone parameters.

Milestone type and parameter with total participants (n)	Improvement level (%)	95% CI[a]		
		Average sample (n)	Lower-upper bound	Average improvement
Daily living skills		117	77.88-86.12	82
Asks to use toilet (n=106)	61 (57.5)			
Brushes teeth (n=140)	113 (80.7)			
Buttons large buttons in front, in correct buttonholes (n=109)	70 (64.2)			
Urinates in toilet or potty (n=113)	84 (74.3)			
Communication		90	33.48-42.01	37.75
Listens to a story for at least 15 minutes (n=101)	35 (34.7)			
Points to at least five body parts when asked (n=117)	62 (52.9)			
Says month and day of birthday when asked (n=116)	42 (36.2)			
Says own phone number when asked (n=23)	12 (52.1)			
Motor skills		123	80.94-86.06	83.5
Draws circle freehand while looking at an example (n=136)	100 (73.5)			
Glues or pastes 2 or more pieces together (n=130)	87 (66.9)			
Jumps with both feet off the floor (n=104)	65 (62.5)			
Runs smoothly without falling (n=119)	82 (68.9)			
Socialization		96	32.58-45.42	39
Ends conversation appropriately (eg, “good bye” or “khoda hafez”) (n=14)	4 (28.5)			
Keeps comfortable distance between self and others in social situations (n=130)	76 (58.4)			
Talks with others about shared interests (eg, sports, TV shows, cartoons) (n=126)	50 (39.6)			
Uses words to express emotions (eg, “I am happy” or “I am scared”) (n=110)	24 (21.8)			

Data Set Selection

In the mCARE study, among the 4 categories (Table 3) of milestone parameters, the “daily living skills” showed the highest improvement level. In this study, we selected this category for building the prediction model based on the participant’s demography. In this milestone type, there are 4 different parameters: *Asks to use toilet*, *Brushes teeth*, *Buttons large buttons in front, in correct buttonholes*, and *Urinates in toilet or potty*. We took the demographic information for every participant who had these milestone parameters and created 4 data sets. In each data set, there were 18 features regarding the participant’s demographic (Multimedia Appendix 1) and 1 value for the “end improvement level” for each participant (this is the label value that will be used in supervised machine learning). We titled each data set by the name of the milestone parameter; for example, *Asks to use toilet*, which has 106 instances; *Brushes teeth*, which has 140 instances; *Buttons large buttons in front, in correct buttonholes*, which has 109 instances, and *Urinates in toilet or potty*, which has 113 instances. In the following sections, we describe the different machine learning models based on these 4 data sets.

Data Preprocessing

Before building the prediction model, we have preprocessed our data set into 3 steps. In the following section, we will describe these steps.

Data Cleaning and Feature Extraction

In the data cleaning step, we observed some missing data, especially with regard to age and salary, in our data sets. We handled this by replacing the empty cell with the mean value for that particular data set. In our data sets, out of 19 columns, only 6 had a numerical value, whereas others had a string input. Therefore, we created dummy variables for every column and converted the string input into a numerical input to handle this. For example, we categorized the column “gender” into 2 subcolumns, namely, “male” and “female.” The corresponding binary codes were set as “1” if the original input is male; otherwise “0.” By using a similar approach we set the female column. We could thus convert our whole data set into a numeric

type by this feature extraction, but the problem is it increased the feature number to 48 from 19. Besides the feature extraction, we used the MinMaxScaler [56,57] to convert all of our features from the 0 to 1 range, as it increases the performance of the machine learning algorithm [58].

Feature Selection

To get the most important features, we first created an extended data set from the "daily living skills" parameter with 18 features. We have used 3 different feature selection methods (univariate selection [59,60], feature importance [59,61-63], and correlation matrix with heatmap [59,64]) with our domain knowledge to select the 10 most important features from the extended data sets. From univariate selection [60] and feature importance [61-63], we obtained 10 important features with their score from each approach (Multimedia Appendices 2 and 3, respectively). We also prepared an important correlation matrix (Multimedia Appendix 4) with heatmap [64] for the features. After computing the most important features with their scores from the 3 feature selection methods, we selected the 10 most important features using these results and our domain knowledge. These features were "family expenditure," "mother age," "father age," "going to specialized school," "number of siblings," "housewife-mother," "father in service," "living in urban," "nuclear family," and "mother education level (undergraduate)." After that, we again split the extended data set into 4 data sets (ie, Brushes teeth; Buttons large buttons in front, in correct buttonholes; Urinates in toilet or potty; and Asks to use toilet) using only these 10 features and with the "end improvement level." These feature-selected data sets are very important in machine learning algorithm to boost up model performance.

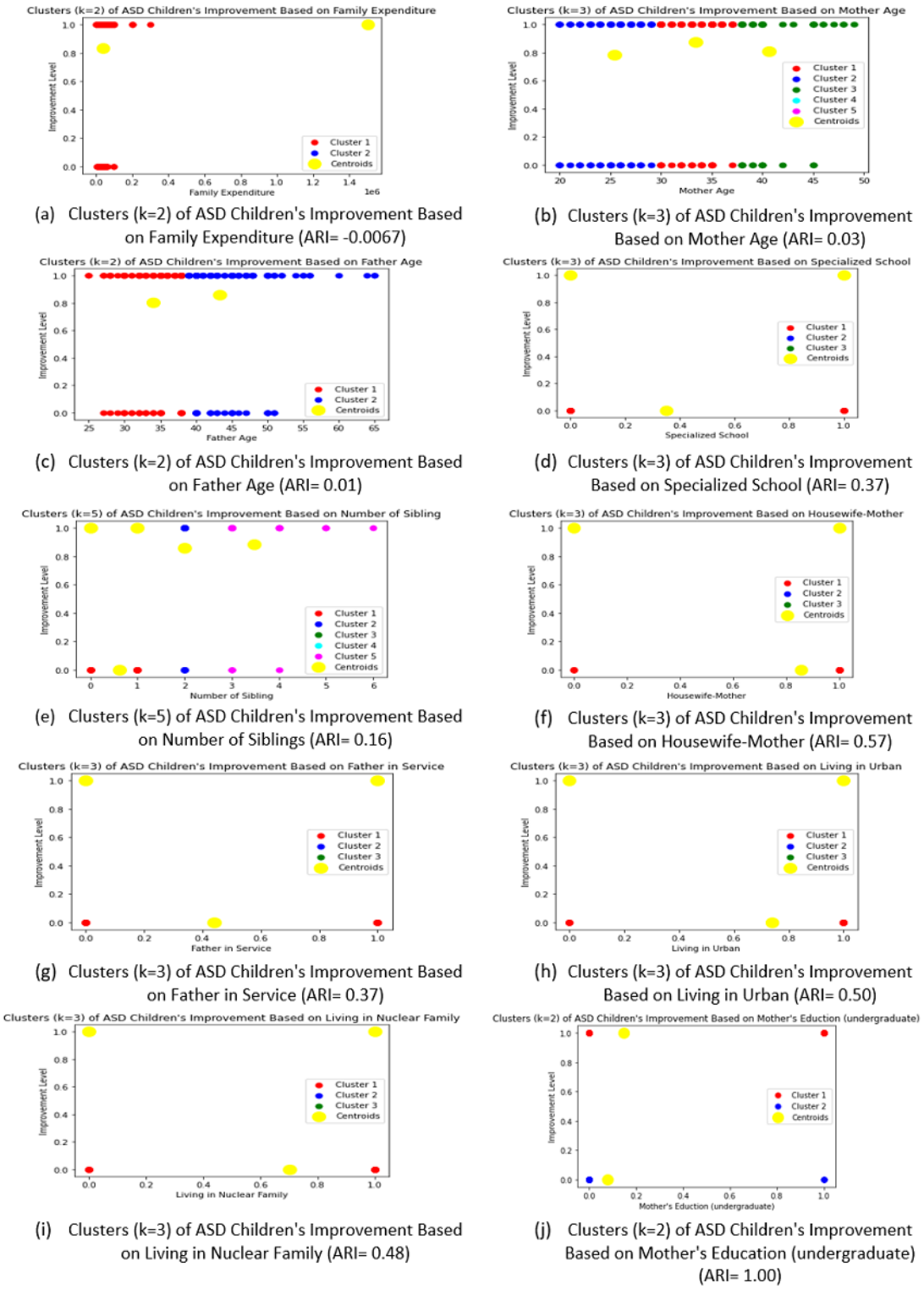
Exploring the Relationship and Associations Underlying the Data Set by Unsupervised Machine Learning: K-Means Clustering

To understand the relation of the 10 selected features (described in the "Feature Selection" section) with the improvement level of "daily living skills" of children with ASD, we implemented K-means clustering [65] to create clusters. As our improvement level is "0" and "1," we have to describe the children's improvement clusters by the "cluster centroid." Figure 2 shows the 10 clusters for the 10 selected features in "daily living skills." We have selected the cluster number (k) by using the "elbow method" [66]. All elbow graphs are shown in Multimedia Appendix 5. We also validated the cluster number by "Adjusted Random Index" [67].

From the cluster in Figure 2A, we can see that the improvement of children with ASD from high-income families is better than those from low-income families. Age of parents is an important factor in the development of children with ASD, as middle-aged mothers (from Figure 2B) and old-aged fathers (Figure 2C) can take better care of their children's development. We also obtained similar types of clusters from Figure 2F and 2G, where occupation of parents plays a vital role in the development of their children with ASD. The number of siblings, living in the urban area, and family size (nuclear) are also important factors in our data set. From the clusters in Figure 2E, 2H, and 2I, we can see that small families with less siblings in the urban area can help improve the children in their "daily living skills." Education levels of children with ASD, especially in specialized school, and their parent's education, especially mother's higher education, can also be helpful for their "daily living skills" development (Figure 2D and 2J).

From the explanation of the clusters in Figure 2, we can find the association between our selected feature and the development of children with ASD. Further, using these data, we can validate our main findings, which is described in detail in the "Principal Results" section.

Figure 2. Cluster for the Selected Features of "Daily Living Skills" using K-Means Algorithm. ARI: Adjusted Random Index; ASD: Autism spectrum disorder.



(a) Clusters (k=2) of ASD Children's Improvement Based on Family Expenditure (ARI= -0.0067)

(b) Clusters (k=3) of ASD Children's Improvement Based on Mother Age (ARI= 0.03)

(c) Clusters (k=2) of ASD Children's Improvement Based on Father Age (ARI= 0.01)

(d) Clusters (k=3) of ASD Children's Improvement Based on Specialized School (ARI= 0.37)

(e) Clusters (k=5) of ASD Children's Improvement Based on Number of Siblings (ARI= 0.16)

(f) Clusters (k=3) of ASD Children's Improvement Based on Housewife-Mother (ARI= 0.57)

(g) Clusters (k=3) of ASD Children's Improvement Based on Father in Service (ARI= 0.37)

(h) Clusters (k=3) of ASD Children's Improvement Based on Living in Urban (ARI= 0.50)

(i) Clusters (k=3) of ASD Children's Improvement Based on Living in Nuclear Family (ARI= 0.48)

(j) Clusters (k=2) of ASD Children's Improvement Based on Mother's Education (undergraduate) (ARI= 1.00)

Building the Model by Machine Learning

We have used 4 supervised machine learning algorithms (decision tree [68], logistic regression [36,69,70], KNN [71,72], and ANN [73-75]) to build the prediction model and compared the results to find out the best machine learning algorithm that can be used for the prediction from this kind of problem and data sets. We used 4 data sets (described in the "Select the Data Set" section) for each algorithm. We used 80% of data for training purposes and 20% for testing purposes from every data set for all the algorithms. We validated our models by k-fold cross-validation (where k=5) [76,77] and took the score's average as the model's accuracy. We describe the models based on different machine learning algorithms in the following sections.

Supervised Machine Learning

Decision Tree

For implementation of the decision tree classification algorithm, we used the tree.DecisionTreeClassifier [78] from the sklearn library [79] of Python [80] to build models for 4 distinguished data sets. The highest accuracy (87.85%; average of fivefold cross-validation score) was obtained for the *Brushes teeth* data set among the 4 models. These models were implemented in Python's Jupyter Notebook [81].

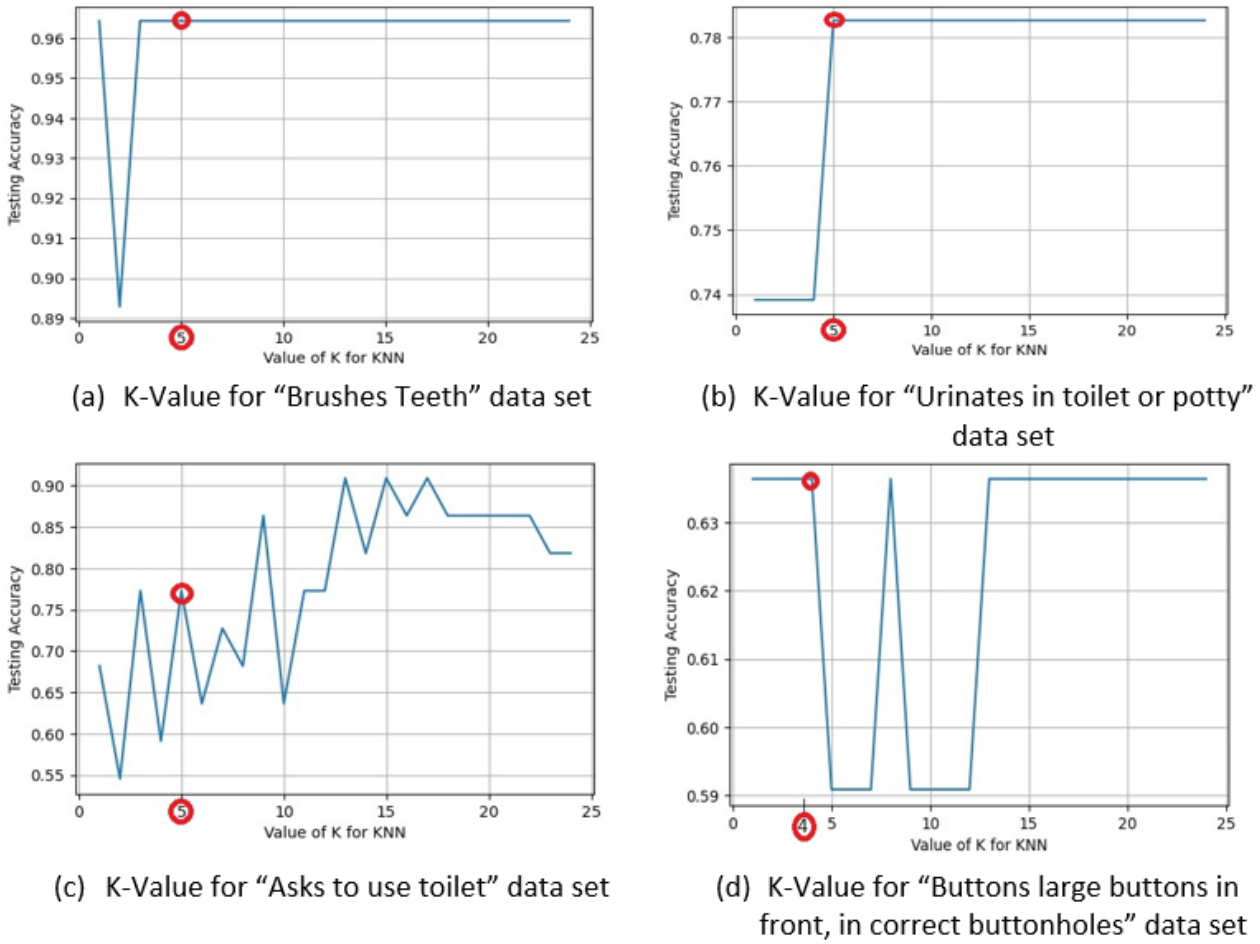
Logistic Regression

For implementation of the classification model, we used the LogisticRegression class [82] from the sklearn library [79] of Python [80] to build 4 predictive models from the "daily living skills" milestone parameter. We calculated the accuracy of the model based on the average fivefold cross-validation score, with accuracies for *Brushes teeth*, *Asks to use toilet*, *Urinates in toilet or potty*, and *Buttons large buttons in front, in correct buttonholes* being 95.00%, 77.35%, 84.98%, and 71.55%, respectively.

K-Nearest Neighbor

We implemented this model in Python Jupyter Notebook using the KNeighborsClassifier [83] from the sklearn library [79] of Python [80]. In this algorithm, the K-value selection is the key to measure the model's performance. For this reason, to build the relationship between the K-value and testing accuracy, we created a plot for a range of K-values against the accuracy for every data set (Figure 3). From the graphical representation, we can easily pick the right K-value for a standard accuracy data set. For example, from Figure 3A, we have chosen K=5 for the *Brushes teeth* data set and applied it in the KNeighborsClassifier [83], which created 95.00% (average fivefold cross-validation score) of the model. For other data sets, similarly, we used the K-value from the graphical representation of Figure 3 and obtained satisfactory accuracy (details of outcomes are described in the "Results" section).

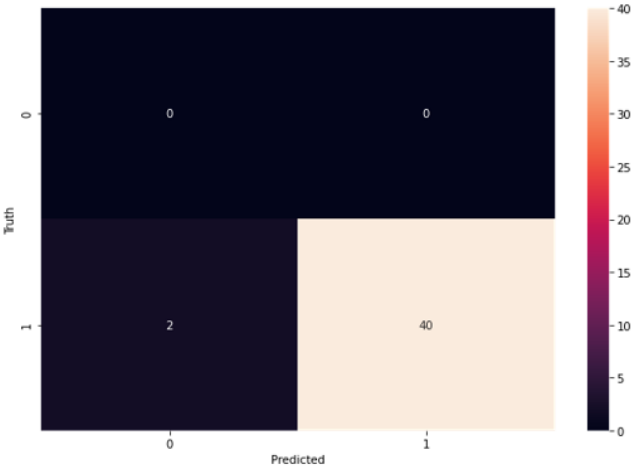
Figure 3. Graphical representation for calculating the best K- value against the test accuracy for the datasets. KNN: K-Nearest Neighbor.



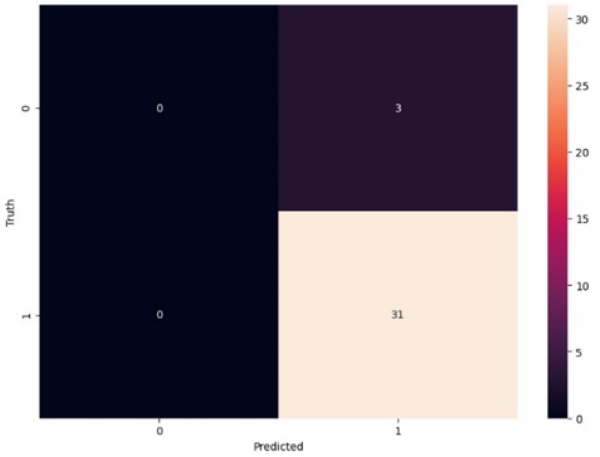
Artificial Neural Network

We have used the keras.Sequential [84] model from the TensorFlow [85] library to build the models. Figure 4 shows the confusion matrix for the 4 data sets using ANN. Table 4 shows the ANN model's overall classification report for all data sets.

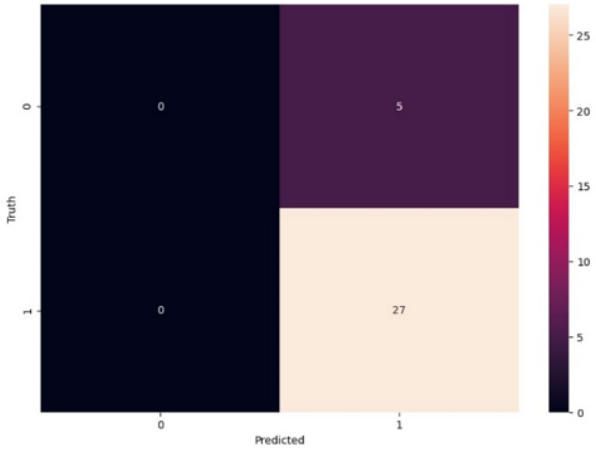
Figure 4. Confusion Matrix for all the Datasets.



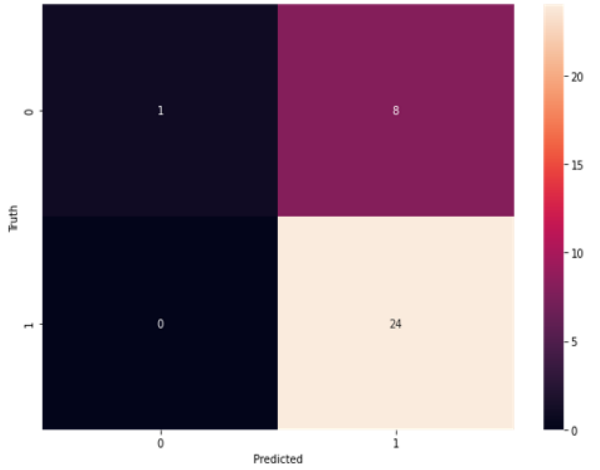
(a) Confusion Matrix for “Brushes Teeth” data set



(b) Confusion Matrix for “Urinates in toilet or potty” data set



(c) Confusion Matrix for “Asks to use toilet” data set



(d) Confusion Matrix for “Buttons large buttons in front, in correct buttonholes” data set

Table 4. The artificial neural network model's overall classification report for all data sets.

Data set and classification report	Precision	Recall	F1 score	Support
Brushes teeth				
0	0.00	0.00	0.00	0
1	1.00	0.95	0.98	42
Accuracy	N/A[a]	N/A	0.95	42
Macro average	0.50	0.48	0.49	42
Weighted average	1.00	0.95	0.98	42
Urinates in toilet or potty				
0	0.00	0.00	0.00	3
1	0.91	1.00	0.95	31
Accuracy	N/A	N/A	0.91	34
Macro average	0.46	0.50	0.48	34
Weighted average	0.83	0.91	0.87	34
Asks to use toilet				
0	0.00	0.00	0.00	5
1	0.84	1.00	0.92	27
Accuracy	N/A	N/A	0.84	32
Macro average	0.42	0.50	0.46	32
Weighted average	0.71	0.84	0.77	32
Buttons large buttons in front, in correct buttonholes				
0	1.00	0.11	0.20	9
1	0.75	1.00	0.86	24
Accuracy	N/A	N/A	0.76	33
Macro average	0.88	0.56	0.53	33
Weighted average	0.82	0.76	0.68	33

[a]N/A: not applicable.

Results

In this study, we have implemented 4 supervised machine languages to build predictive models for the "daily living skill" milestone parameter of children with ASD based on their demography. A summary of the results for different machine learning algorithms for predicting this milestone parameter is presented in Table 5.

We validated the model's result by a fivefold validation score. From Table 5, we can conclude that, based on the demography, *Daily living skills* and *Brushes teeth* data sets had the highest accuracy in all machine learning–based models. The "ANN" performed well among the machine learning algorithms studied. In conclusion, if we need to develop an automated system to predict the "daily living skill" milestone parameter development based on the demography, then from this study's outcome, we can recommend developing a system based on machine learning algorithm, especially ANN.

We validated the performance of our classifiers by receiver operating characteristic–area under the curve (ROC–AUC) [86] scores (Table 6), with score "1" considered the outstanding classifier. Rice and Harris [87] suggested that, in applied psychology and prediction model of future behavior, the ROC–AUC values of 0.70 or higher would be considered to have strong effects. The average ROC–AUC scores (from 4 parameters) of the decision tree, logistic regression, KNN, and ANN were 0.84, 0.86, 0.76, and 0.83, respectively (Table 6). The ROC curves of these classifiers are presented in Multimedia Appendices 6-9.

Table 5. Summary of the accuracy of all prediction models based on demography for "daily living skills."

Parameter types	Decision tree (fivefold cross-validation score)	Logistic regression (fivefold cross-validation score)	K-nearest neighbor fivefold cross-validation score)	Artificial neural network
Brushes teeth	87.85%	95.00%	95.00% (K=5)	95.00%
Asks to use toilet	71.64%	77.35%	84.00% (K=13)	84.00%
Urinates in toilet or potty	72.52%	84.98%	85.02% (K=5)	91.00%
Buttons large buttons in front, in correct buttonholes	73.46%	71.55%	66.88% (K=5)	76.00%

Table 6. Summary of receiver operating characteristic–area under the curve for all prediction models based on demography for "daily living skills."

Parameter types	Decision tree	Logistic regression	K-nearest neighbor	Artificial neural network
Brushes teeth	0.68	0.91	0.65	0.80
Asks to use toilet	0.95	0.77	0.77	0.76
Urinates in toilet or potty	0.78	0.89	0.86	0.91
Buttons large buttons in front, in correct buttonholes	0.94	0.86	0.75	0.84

Discussion

Principal Findings

This study reports on some major evidence-based findings regarding patients with ASD and their development in the milestone categories based on demography.

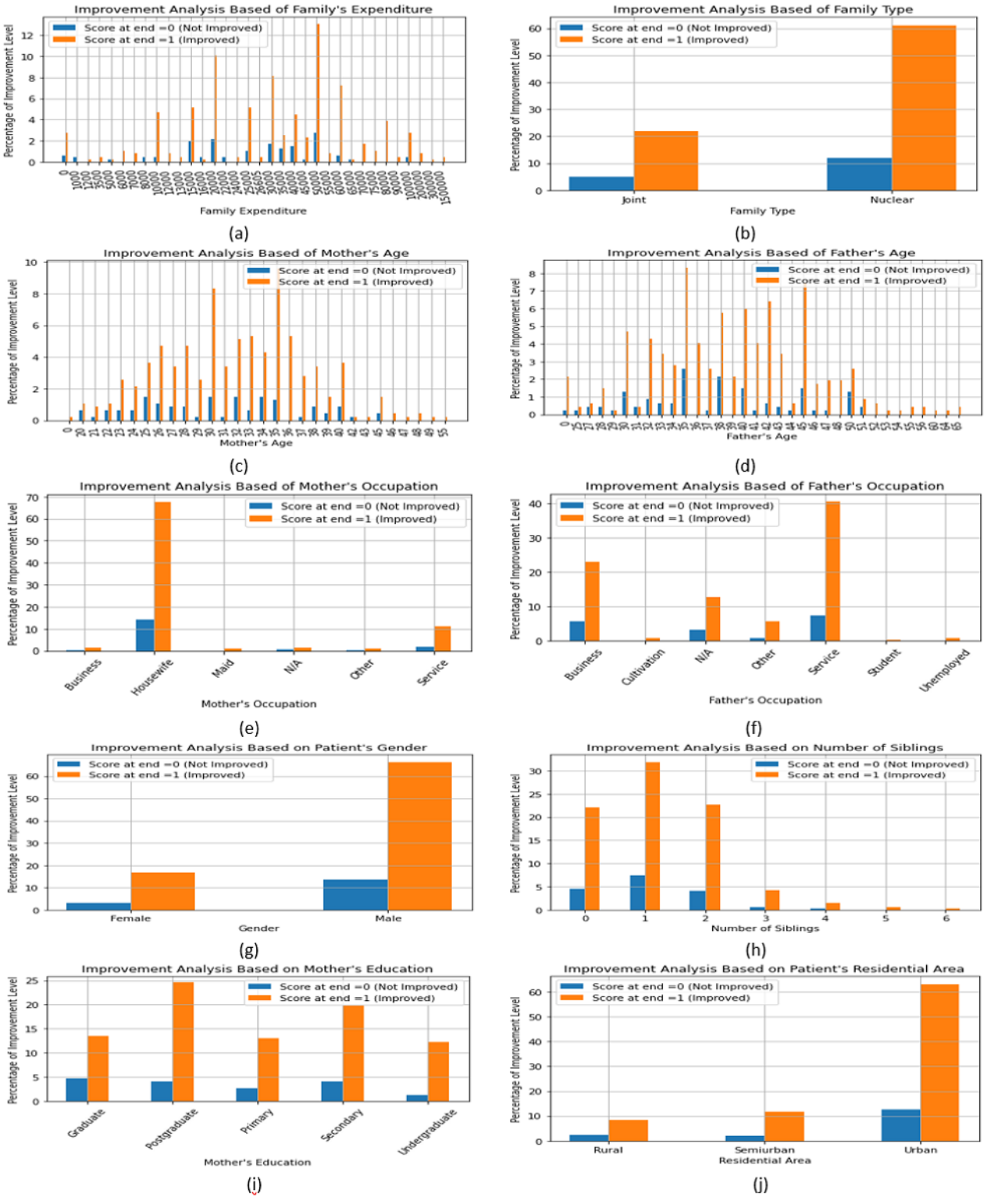
Finding 1

Among the 4 major milestone categories, “daily living skills” had the highest improvement level. Thus, it can be concluded that the caregiver and care practitioner give more importance to developing the daily living skills of children with ASD so that they can live independently without requiring any help from others.

Finding 2

The demography of children with ASD impacts the development of their milestone parameters. In Figure 5, we have summarized the demography that impacts the development of their “daily living skills” parameter. Here, “score_at_end”=1 is the final improvement point of the children with ASD. We see that family income or expenditure (Figure 5A) in the middle range helps children with ASD to develop. Besides, a nuclear family (Figure 5B) with a small number of siblings (Figure 5H) in the urban area (Figure 5J) shows the higher improvement rate of children with ASD. The age of parents is also an important factor in the development of children with ASD; generally, middle-aged (aged 25-45) parents can take better care of their children during the course of their development (Figure 5C and 5D). Occupation and education of parents are other good factors to consider; our results show that a mother who works in the house (Figure 5E) but has good education (Figure 5I) and an employed father (Figure 5F) can help achieve significant development in their child. Lastly, gender of patients remains another significant demography in our study, with male children’s development being far better than that of female children (Figure 5G).

Figure 5. The Summary of the Demography's importance behind the ASD Children's Milestone Parameter Development.



Finding 3

We implemented 4 supervised machine learning algorithms to predict the "daily living skills" improvement level of children with ASD based on their demography. Among the 4 algorithms, the ANN performs better than others, and it has, on average, an average accuracy of over 80% from the same data set we have used in other algorithms. Thus, we can conclude and recommend the ANN to develop a demography-based prediction tool in the intervention or treatment process of children with ASD.

Limitations

Although we achieved some satisfactory results and reported important findings in this study, our data set lacks in some aspects. The first limitation of the data set is its scattered property, which makes it challenging to find patterns for analysis, but still we achieved good accuracy from this data set.

Increasing the number of data can help resolve this problem. Although some studies had been done in this area, the real data set remains very rare. Therefore, we could not compare our study results and findings with other studies and data sets.

Comparison With Prior Work

Most mental health work is related to identification or recognition and symptom analysis of ASD [88]. In this study, we have implemented machine learning models to predict the improvement level of children with ASD based on their demography. A few studies have been performed in this area, and these are described in the following section.

Scheer et al [38] built a clinical model to predict proximal junctional kyphosis and proximal junctional failure. They used the baseline demographic, radiographic, and surgical factors for 510 patients to build the prediction model. The model's overall accuracy was 86.3%, which has a great significance in caregiving decision making, risk analysis, and risk prediction before surgery. To build this model, they used the decision tree machine learning algorithm with 5 different bootstrapped models. This model would have been more sophisticated had they used more than 1 machine learning model for the prediction.

Another machine learning–based work has been performed by Tariq et al [2] to detect developmental delay in patients with autism, wherein they used home videos of Bangladeshi children to train and validate the model. Their study's main objective was to determine the "risk scores" for autism. Using a 2-classification layer neural network, they achieved 85% accuracy for predicting developmental delay. This work has been very effective not only for predicting developmental delay but also for early detection of autism remotely. The authors trained the model with the US data set, but they achieved only low accuracy when applying the Bangladeshi data set. Thus, the model had no cultural divergence.

To evaluate the ADDM status of children, Maenner et al [39] have developed a machine learning–based model using the words and phrases in children's developmental evaluation. This model has been built with the random forest classifier by deploying the *2008 Georgia* data set containing data on 1162 children. With 86.5% accuracy, the machine learning–based algorithm significantly differentiated between the children that do and do not meet ASD surveillance criteria. As is the case with Scheer et al [38], this work would have been more in-depth had there been more than 1 machine learning algorithm for building the model.

Nowell et al [15] summarized in their review that patients' demographic has an influence on their ASD development. The main finding of their study was that "myriad demographic factors influence the diagnosis of ASD." Their study proves that the patient's demography, including race, socioeconomic status, ethnicity, and parental education, is the most important factor in ASD diagnosis. However, most of the studies reviewed were based on children in the United States.

A sufficient number of studies have been performed to detect ASD by both supervised [2,89-97] and unsupervised machine [98,99] learning methods. In our study, supervised machine learning has mainly been used for the detection of ASD through behavioral or neuroimaging data, whereas unsupervised machine learning was deployed for predicting ASD assessment. In supervised machine learning, logistic regression, KNN, neural network, convolution neural network, naive Bayes, support vector machine, and rule-based machine learning models have been used to detect ASD. Raj and Masood [89] deployed some supervised machine learning models with 3 nonclinical ASD data sets to predict and analyze the problem of ASD. Feature selection–based machine learning has been used to detect ASD with accuracy greater than 90% [90]. Tariq et al [2,91] used home videos of Bangladeshi children with ASD in supervised machine learning to detect their speech and language problems. Küpper et al [92] deployed the clinical behavioral feature in support vector machine to detect the ASD problems in adolescents. Besides these studies, rule-based [93] classification approaches such as decision trees, random forest, and linear discriminant analysis [94-97] have been used to detect ASD. By contrast, unsupervised machine learning has been used for predicting ASD assessment or analysis of ASD problem in children [98,99].

Comparison With Our Study

Most of the work on children with ASD concerned generalized development, but in this study, we developed prediction models for specific milestone parameters concerning development in children with ASD. Unlike other previous studies, we have validated the prediction result for a specific milestone parameter with more than 1 machine learning algorithm. Our study used the same cultural demographic data set (from Bangladesh) for both training and predicting the models, which helps to get an accurate result from the models.

Conclusions

This study implies 3 significant factors in the area of mental health development of children with ASD in low- and middle-income countries such as Bangladesh. First, we evaluated the improvement in milestone parameters in children with ASD from the mCARE project. The "daily living skills" and "motor skills" had significant improvement after deploying mCARE tools. We have developed 4 supervised machine learning models based on the demographic information of children with ASD to predict their "daily living skills" development. By comparing the accuracy of the algorithms, we can conclude that the ANN with 1 hidden layer can provide the appropriate prediction for the improvement in "daily living skills" of children with ASD. At the end of the study, from the supervised and unsupervised algorithms, we found some important demographic characteristics that can impact the improvement level in children with ASD. In conclusion, successful and accurate prediction tools deploying this study's findings will make a renovation in the area of mental health, especially in the development of children with ASD.

Acknowledgments

This study has been partially supported by an NIH grant (1R21MH116726-01). The authors are thankful to 4 specialist autism health care centers and institutions in Bangladesh: The Institute of Pediatric Neuro-disorder & Autism (IPNA) Bangladesh, The National Institute of Mental Health (NIMH), Autism Welfare Foundation (AWF), and Nishpap Autism Foundation and their respective departments for their continuous support throughout this study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data set features: 18 features about the participant's demographic for this study.
[DOCX File , 13 KB - medinform_v9i6e29242_app1.docx]

Multimedia Appendix 2
Best 10 features from "daily living skill" by univariate selection method.
[DOCX File , 13 KB - medinform_v9i6e29242_app2.docx]

Multimedia Appendix 3
Top 10 Features from "Daily Living Skills" by Feature Importance method.
[PNG File , 43 KB - medinform_v9i6e29242_app3.png]

Multimedia Appendix 4
Correlation Matrix with Heatmap for "Daily Living Skill" Dataset.
[PNG File , 4390 KB - medinform_v9i6e29242_app4.png]

Multimedia Appendix 5
Cluster Analysis by "Elbow Method" for the Selected Features of "Daily Living Skills".
[PNG File , 299 KB - medinform_v9i6e29242_app5.png]

Multimedia Appendix 6
ROC curve for four parameters of "Daily Living Skills" by "Decision Tree" model.
[PNG File , 205 KB - medinform_v9i6e29242_app6.png]

Multimedia Appendix 7
ROC curve for four parameters of "Daily Living Skills" by "Logistic Regression" model.
[PNG File , 204 KB - medinform_v9i6e29242_app7.png]

Multimedia Appendix 8
ROC curve for four parameters of "Daily Living Skills" by "K-Nearest Neighbor" model.
[PNG File , 200 KB - medinform_v9i6e29242_app8.png]

Multimedia Appendix 9
ROC curve for four parameters of "Daily Living Skills" by "Artificial Neural Network" model.
[PNG File , 211 KB - medinform_v9i6e29242_app9.png]

References

1. Wallace GL, Kenworthy L, Pugliese CE, Popal HS, White EI, Brodsky E, et al. Real-World Executive Functions in Adults with Autism Spectrum Disorder: Profiles of Impairment and Associations with Adaptive Functioning and Co-morbid Anxiety and Depression. J Autism Dev Disord 2016 Mar 16;46(3):1071-1083 [FREE Full text] [doi: 10.1007/s10803-015-2655-7] [Medline: 26572659]
2. Tariq Q, Fleming SL, Schwartz JN, Dunlap K, Corbin C, Washington P, et al. Detecting Developmental Delay and Autism Through Machine Learning Models Using Home Videos of Bangladeshi Children: Development and Validation Study. J Med Internet Res 2019 Apr 24;21(4):e13822 [FREE Full text] [doi: 10.2196/13822] [Medline: 31017583]
3. Kanner L. Autistic disturbances of affective contact. Acta Paedopsychiatr 1968;35(4):100-136. [Medline: 4880460]

4. Sealey L, Hughes B, Sriskanda A, Guest J, Gibson A, Johnson-Williams L, et al. Environmental factors in the development of autism spectrum disorders. Environ Int 2016 Mar;88:288-298. [doi: 10.1016/j.envint.2015.12.021] [Medline: 26826339]
5. Hisle-Gorman E, Susi A, Stokes T, Gorman G, Erdie-Lalena C, Nylund CM. Prenatal, perinatal, and neonatal risk factors of autism spectrum disorder. Pediatr Res 2018 Aug;84(2):190-198. [doi: 10.1038/pr.2018.23] [Medline: 29538366]
6. DiGuiseppi CG, Daniels JL, Fallin DM, Rosenberg SA, Schieve LA, Thomas KC, et al. Demographic profile of families and children in the Study to Explore Early Development (SEED): Case-control study of autism spectrum disorder. Disabil Health J 2016 Jul;9(3):544-551 [FREE Full text] [doi: 10.1016/j.dhjo.2016.01.005] [Medline: 26917104]
7. Speaks A. Autism Statistics and Facts. autism speaks. URL: https://www.autismspeaks.org/autism-statistics-asd [accessed 2021-03-18]
8. Happé FG, Mansour H, Barrett P, Brown T, Abbott P, Charlton RA. Demographic and Cognitive Profile of Individuals Seeking a Diagnosis of Autism Spectrum Disorder in Adulthood. J Autism Dev Disord 2016 Nov;46(11):3469-3480. [doi: 10.1007/s10803-016-2886-2] [Medline: 27549589]
9. Cromer J. Autism: fastest-growing developmental disability. Autism: U.S. ARMY; 2018. URL: https://www.army.mil/article/203386/autism_fastest_growing_developmental_disability [accessed 2021-03-17]
10. Baio J, Wiggins L, Christensen DL, Maenner MJ, Daniels J, Warren Z, et al. Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2014. MMWR Surveill Summ 2018 Apr 27;67(6):1-23 [FREE Full text] [doi: 10.15585/mmwr.ss6706a1] [Medline: 29701730]
11. Hossain MD, Ahmed HU, Jalal Uddin MM, Chowdhury WA, Iqbal MS, Kabir RI, et al. Autism Spectrum disorders (ASD) in South Asia: a systematic review. BMC Psychiatry 2017 Aug 01;17(1):281 [FREE Full text] [doi: 10.1186/s12888-017-1440-x] [Medline: 28826398]
12. Baghdadli A, Pascal C, Grisi S, Aussilloux C. Risk factors for self-injurious behaviours among 222 young children with autistic disorders. J Intellect Disabil Res 2003 Nov;47(Pt 8):622-627. [doi: 10.1046/j.1365-2788.2003.00507.x] [Medline: 14641810]
13. De Giacomo A, Fombonne E. Parental recognition of developmental abnormalities in autism. European Child & Adolescent Psychiatry 1998 Oct 12;7(3):131-136. [doi: 10.1007/s007870050058]
14. Fombonne E. The epidemiology of autism: a review. Psychol Med 1999 Jul;29(4):769-786. [doi: 10.1017/s0033291799008508] [Medline: 10473304]
15. Nowell KP, Brewton CM, Allain E, Mire SS. The Influence of Demographic Factors on the Identification of Autism Spectrum Disorder: A Review and Call for Research. Rev J Autism Dev Disord 2015 Jul 10;2(3):300-309. [doi: 10.1007/s40489-015-0053-x]
16. Zwaigenbaum L, Bryson S, Garon N. Early identification of autism spectrum disorders. Behav Brain Res 2013 Aug 15;251:133-146. [doi: 10.1016/j.bbr.2013.04.004] [Medline: 23588272]
17. Harris S, Handleman J. Age and IQ at intake as predictors of placement for young children with autism: a four- to six-year follow-up. J Autism Dev Disord 2000 Apr;30(2):137-142. [doi: 10.1023/a:1005459606120] [Medline: 10832778]
18. Turner LM, Stone WL, Pozdol SL, Coonrod EE. Follow-up of children with autism spectrum disorders from age 2 to age 9. Autism 2006 May;10(3):243-265. [doi: 10.1177/1362361306063296] [Medline: 16682397]
19. Webb S, Jones E. Early Identification of Autism: Early Characteristics, Onset of Symptoms, and Diagnostic Stability. Infants Young Child 2009;22(2):100-118 [FREE Full text] [doi: 10.1097/IYC.0b013e3181a02f7f] [Medline: 28090148]
20. Filipek PA, Accardo PJ, Ashwal S, Baranek GT, Cook EH, Dawson G, et al. Practice parameter: screening and diagnosis of autism: report of the Quality Standards Subcommittee of the American Academy of Neurology and the Child Neurology Society. Neurology 2000 Aug 22;55(4):468-479. [doi: 10.1212/wnl.55.4.468] [Medline: 10953176]
21. Lord C, Volkmar F. Genetics of childhood disorders: XLII. Autism, part 1: Diagnosis and assessment in autistic spectrum disorders. J Am Acad Child Adolesc Psychiatry 2002 Sep;41(9):1134-1136. [doi: 10.1097/00004583-200209000-00015] [Medline: 12218436]
22. Zwaigenbaum L, Bauman ML, Stone WL, Yirmiya N, Estes A, Hansen RL, et al. Early Identification of Autism Spectrum Disorder: Recommendations for Practice and Research. Pediatrics 2015 Oct;136 Suppl 1:S10-S40. [doi: 10.1542/peds.2014-3667C] [Medline: 26430168]
23. Bauman ML. Medical comorbidities in autism: challenges to diagnosis and treatment. Neurotherapeutics 2010 Jul;7(3):320-327 [FREE Full text] [doi: 10.1016/j.nurt.2010.06.001] [Medline: 20643385]
24. Sowa M, Meulenbroek R. Effects of physical exercise on Autism Spectrum Disorders: A meta-analysis. Research in Autism Spectrum Disorders 2012 Jan;6(1):46-57. [doi: 10.1016/j.rasd.2011.09.001]
25. Courchesne E, Carper R, Akshoomoff N. Evidence of brain overgrowth in the first year of life in autism. JAMA 2003 Jul 16;290(3):337-344. [doi: 10.1001/jama.290.3.337] [Medline: 12865374]
26. Ursano R, Bell C, Eth S, Friedman M, Norwood A, Pfefferbaum B, Work Group on ASDPTSD, Steering Committee on Practice Guidelines. Practice guideline for the treatment of patients with acute stress disorder and posttraumatic stress disorder. Am J Psychiatry 2004 Nov;161(11 Suppl):3-31. [Medline: 15617511]
27. Dyches TT, Wilder LK, Sudweeks RR, Obiakor FE, Algozzine B. Multicultural Issues in Autism. J Autism Dev Disord 2004 Apr;34(2):211-222. [doi: 10.1023/b:jadd.0000022611.80478.73]

28. Mandell DS, Listerud J, Levy SE, Pinto-Martin JA. Race differences in the age at diagnosis among medicaid-eligible children with autism. J Am Acad Child Adolesc Psychiatry 2002 Dec;41(12):1447-1453. [doi: 10.1097/00004583-200212000-00016] [Medline: 12447031]
29. Ravindran N, Myers BJ. Cultural Influences on Perceptions of Health, Illness, and Disability: A Review and Focus on Autism. J Child Fam Stud 2011 May 12;21(2):311-319. [doi: 10.1007/s10826-011-9477-9]
30. Thomas P, Zahorodny W, Peng B, Kim S, Jani N, Halperin W, et al. The association of autism diagnosis with socioeconomic status. Autism 2012 Mar;16(2):201-213. [doi: 10.1177/1362361311413397] [Medline: 21810908]
31. Sathyabama R. Clinical characteristics and demographic profile of children with Autism Spectrum Disorder (ASD) at child development clinic (CDC), Penang Hospital, Malaysia. Med J Malaysia 2019 Oct;74(5):372-376 [FREE Full text] [Medline: 31649211]
32. Courchesne E, Gazestani VH, Lewis NE. Prenatal Origins of ASD: The When, What, and How of ASD Development. Trends Neurosci 2020 May;43(5):326-342 [FREE Full text] [doi: 10.1016/j.tins.2020.03.005] [Medline: 32353336]
33. Siller M, Reyes N, Hotez E, Hutman T, Sigman M. Longitudinal change in the use of services in autism spectrum disorder: understanding the role of child characteristics, family demographics, and parent cognitions. Autism 2014 May;18(4):433-446. [doi: 10.1177/1362361313476766] [Medline: 24108191]
34. Decision Trees for Classification: A Machine Learning Algorithm. Resources X. 2017 Sep 7. URL: https://www.xoriant.com/blog/product-engineering/decision-trees-machine-learning-algorithm.html [accessed 2021-03-20]
35. Molnar C. Logistic Regression. Interpretable Machine Learning. URL: https://christophm.github.io/interpretable-ml-book/logistic.html [accessed 2021-03-20]
36. Logistic regression. Wikipedia. URL: https://en.wikipedia.org/wiki/Logistic_regression [accessed 2021-03-19]
37. Euclidean distance. Wikipedia. URL: https://en.wikipedia.org/wiki/Euclidean_distance [accessed 2021-03-20]
38. Scheer J, Osorio J, Smith J, Schwab F, Lafage V, Hart R, International Spine Study Group. Development of Validated Computer-based Preoperative Predictive Model for Proximal Junction Failure (PJF) or Clinically Significant PJK With 86% Accuracy Based on 510 ASD Patients With 2-year Follow-up. Spine (Phila Pa 1976) 2016 Nov 15;41(22):E1328-E1335. [doi: 10.1097/BRS.0000000000001598] [Medline: 27831987]
39. Maenner MJ, Yeargin-Allsopp M, Van Naarden Braun K, Christensen DL, Schieve LA. Development of a Machine Learning Algorithm for the Surveillance of Autism Spectrum Disorder. PLoS One 2016;11(12):e0168224 [FREE Full text] [doi: 10.1371/journal.pone.0168224] [Medline: 28002438]
40. Crais ER, Watson LR, Baranek GT, Reznick JS. Early identification of autism: how early can we go? Semin Speech Lang 2006 Aug;27(3):143-160. [doi: 10.1055/s-2006-948226] [Medline: 16941286]
41. Chakrabarti S. Early identification of autism. Indian Pediatr 2009 May;46(5):412-414 [FREE Full text] [Medline: 19179745]
42. Barbaro J, Halder S. Early Identification of Autism Spectrum Disorder: Current Challenges and Future Global Directions. Curr Dev Disord Rep 2016 Feb 20;3(1):67-74. [doi: 10.1007/s40474-016-0078-6]
43. Eaves LC, Ho HH. The very early identification of autism: outcome to age 4 1/2-5. J Autism Dev Disord 2004 Aug;34(4):367-378. [doi: 10.1023/b:jadd.0000037414.33270.a8] [Medline: 15449513]
44. Guthrie W, Swineford L, Nottke C, Wetherby A. Early diagnosis of autism spectrum disorder: stability and change in clinical diagnosis and symptom presentation. J Child Psychol Psychiatry 2013 May;54(5):582-590 [FREE Full text] [doi: 10.1111/jcpp.12008] [Medline: 23078094]
45. Hsiao Y. Autism Spectrum Disorders: Family Demographics, Parental Stress, and Family Quality of Life. Journal of Policy and Practice in Intellectual Disabilities 2018 Mar 09;15(1):70-79. [doi: 10.1111/jppi.12232]
46. Green SA, Carter AS. Predictors and course of daily living skills development in toddlers with autism spectrum disorders. J Autism Dev Disord 2014 Feb;44(2):256-263 [FREE Full text] [doi: 10.1007/s10803-011-1275-0] [Medline: 21598046]
47. Estes A, Munson J, Dawson G, Koehler E, Zhou X, Abbott R. Parenting stress and psychological functioning among mothers of preschool children with autism and developmental delay. Autism 2009 Jul;13(4):375-387 [FREE Full text] [doi: 10.1177/1362361309105658] [Medline: 19535467]
48. Estes A, Olson E, Sullivan K, Greenson J, Winter J, Dawson G, et al. Parenting-related stress and psychological distress in mothers of toddlers with autism spectrum disorders. Brain Dev 2013 Feb;35(2):133-138 [FREE Full text] [doi: 10.1016/j.braindev.2012.10.004] [Medline: 23146332]
49. Drahota A, Wood JJ, Sze KM, Van Dyke M. Effects of cognitive behavioral therapy on daily living skills in children with high-functioning autism and concurrent anxiety disorders. J Autism Dev Disord 2011 Mar;41(3):257-265 [FREE Full text] [doi: 10.1007/s10803-010-1037-4] [Medline: 20508979]
50. U.S. Department of Health and Human Services U. National Institutes of Health (NIH). NIH. URL: https://www.nih.gov/ [accessed 2021-03-18]
51. Ministry of Health and Family Welfare (MoHFW). National Institute of Mental Health and Hospital. Facility Registry-Government of People's Republic of Bangladesh Ministry of Health and Family Welfare. URL: http://facilityregistry.dghs.gov.bd/org_profile.php?org_code=10000010 [accessed 2021-03-18]
52. Institute For Peadiatric Neurodisorder And Autism in BSMMU. (IPNA) IoPNA. 2019. URL: http://ipnabsmmu.edu.bd/ [accessed 2021-03-18]

53. Global T. Nishpap Autism Foundation. Onsite Training with Nishpap Autism Foundation in Chattogram. 2018. URL: https://www.therapglobal.net/globalblog/onsite-training-with-nishpap-autism-foundation-in-chattogram-bangladesh/ [accessed 2021-03-18]
54. Awfbd. Working for the Brighter Future of Person with Autism. Autism Welfare Foundation. URL: https://awfbd.org/ [accessed 2021-03-18]
55. MathsIsFun. Confidence Intervals. Math Fun Advanced. URL: https://www.mathsisfun.com/data/confidence-interval.html [accessed 2021-03-18]
56. Roy B. All about Feature Scaling. Towards Data Science. URL: https://towardsdatascience.com/all-about-feature-scaling-bcc0ad75cb35 [accessed 2021-03-19]
57. Brownlee J. How to Use StandardScaler and MinMaxScaler Transforms in Python. Machine Learning Mastery. URL: https://machinelearningmastery.com/standardscaler-and-minmaxscaler-transforms-in-python/ [accessed 2021-03-19]
58. Chong J. What Is Feature Scaling & Why Is it Important in Machine Learning?. URL: https://towardsdatascience.com/what-is-feature-scaling-why-is-it-important-in-machine-learning-2854ae877048 [accessed 2021-03-18]
59. Shaikh R. Feature Selection Techniques in Machine Learning with Python. 2018. URL: https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e [accessed 2021-03-18]
60. Brewer JK, Hills JR. Univariate selection: The effects of size of correlation, degree of skew, and degree of restriction. Psychometrika 1969 Sep;34(3):347-361. [doi: 10.1007/bf02289363]
61. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics 2010 May 15;26(10):1340-1347. [doi: 10.1093/bioinformatics/btq134] [Medline: 20385727]
62. Zien A, Krämer N, Sonnenburg S, Rätsch G. The Feature Importance Ranking Measure. 2009 Sep 06 Presented at: Joint European Conference on Machine Learning and Knowledge Discovery in Databases; 2009; Berlin, Germany p. 694-709. [doi: 10.1007/978-3-642-04174-7_45]
63. Hooker S, Erhan D, Kindermans P, Kim B. Evaluating feature importance estimates. 2019 Nov 05 Presented at: 33rd Conference on Neural Information Processing Systems (NeurIPS 2019); 2019; Vancouver, BC, Canada.
64. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 2016 Sep 15;32(18):2847-2849. [doi: 10.1093/bioinformatics/btw313] [Medline: 27207943]
65. Likas A, Vlassis N, J. Verbeek J. The global k-means clustering algorithm. Pattern Recognition 2003 Feb;36(2):451-461. [doi: 10.1016/s0031-3203(02)00060-2]
66. Elbow method (clustering). Wikipedia. URL: https://en.wikipedia.org/wiki/Elbow_method_(clustering) [accessed 2021-03-21]
67. Yeung KY, Ruzzo WL. Principal component analysis for clustering gene expression data. Bioinformatics 2001 Sep;17(9):763-774. [doi: 10.1093/bioinformatics/17.9.763] [Medline: 11590094]
68. Decision tree. Wikipedia. URL: https://en.wikipedia.org/wiki/Decision_tree# [accessed 2021-03-19]
69. Wright R. Logistic regression. American Psychological Association 1995:2017-2244.
70. Kleinbaum D, Dietz K, Gail M, Klein M, Klein M. Logistic Regression. New York, NY: Springer; 2010.
71. Kramer O. K-Nearest Neighbors. Berlin, Germany: Springer; 2013:13-23.
72. k-Nearest Neighbors Algorithm. Wikipedia. URL: https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm [accessed 2021-03-19]
73. Wang S. Artificial Neural Network. Boston, MA: Springer; 2003:81-100.
74. Artificial Neural Network. Wikipedia. URL: https://en.wikipedia.org/wiki/Artificial_neural_network [accessed 2021-03-19]
75. Snijders T, Bosker R. Fundamentals of Artificial Neural Networks. London, UK: MIT press; 1999.
76. Cross-Validation (statistics). Wikipedia. URL: https://en.wikipedia.org/wiki/Cross-validation_(statistics) [accessed 2021-03-19]
77. Rodriguez J, Perez A, Lozano J. Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation. IEEE Trans. Pattern Anal. Mach. Intell 2010 Mar;32(3):569-575. [doi: 10.1109/tpami.2009.187]
78. A Decision Tree Classifier: Scikit Learn. URL: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html [accessed 2021-03-20]
79. Varoquaux G, Buitinck L, Louppe G, Grisel O, Pedregosa F, Mueller A. Scikit-learn. GetMobile: Mobile Comp. and Comm 2015 Jun;19(1):29-33. [doi: 10.1145/2786984.2786995]
80. Python Foundation. python. URL: https://www.python.org/ [accessed 2021-03-20]
81. Jupyter. URL: https://jupyter.org/ [accessed 2021-03-20]
82. LogisticRegression. URL: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html [accessed 2021-03-20]
83. KNeighborsClassifier. URL: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html [accessed 2021-03-20]
84. Keras K. The Sequential Model. 2020. URL: https://keras.io/guides/sequential_model/ [accessed 2021-03-20]
85. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C. TensorFlow: Large-scale machine learning on heterogeneous systems. arXiv 2016 Mar 14 [FREE Full text]
86. Jin Huang, Ling C. Using AUC and accuracy in evaluating learning algorithms. IEEE Trans. Knowl. Data Eng 2005 Mar;17(3):299-310. [doi: 10.1109/tkde.2005.50]

87. Rice ME, Harris GT. Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. Law Hum Behav 2005 Oct;29(5):615-620. [doi: 10.1007/s10979-005-6832-7] [Medline: 16254746]
88. Elsevier. Most Downloaded Research in Autism Spectrum Disorders Articles. 2021. URL: https://www.journals.elsevier.com/research-in-autism-spectrum-disorders/most-downloaded-articles [accessed 2021-03-18]
89. Raj S, Masood S. Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques. Procedia Computer Science 2020;167:994-1004. [doi: 10.1016/j.procs.2020.03.399]
90. Kosmicki JA, Sochat V, Duda M, Wall DP. Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. Transl Psychiatry 2015 Mar 24;5(2):e514-e514 [FREE Full text] [doi: 10.1038/tp.2015.7] [Medline: 25710120]
91. Tariq Q, Daniels J, Schwartz JN, Washington P, Kalantarian H, Wall DP. Mobile detection of autism through machine learning on home video: A development and prospective validation study. PLoS Med 2018 Nov 27;15(11):e1002705 [FREE Full text] [doi: 10.1371/journal.pmed.1002705] [Medline: 30481180]
92. Küpper C, Stroth S, Wolff N, Hauck F, Kliewer N, Schad-Hansjosten T, et al. Identifying predictive features of autism spectrum disorders in a clinical sample of adolescents and adults using machine learning. Sci Rep 2020 Mar 18;10(1):4805. [doi: 10.1038/s41598-020-61607-w] [Medline: 32188882]
93. Thabtah F, Peebles D. A new machine learning model based on induction of rules for autism detection. Health Informatics J 2020 Mar 29;26(1):264-286 [FREE Full text] [doi: 10.1177/1460458218824711] [Medline: 30693818]
94. Duda M, Ma R, Haber N, Wall DP. Use of machine learning for behavioral distinction of autism and ADHD. Transl Psychiatry 2016 Mar 09;6(2):e732 [FREE Full text] [doi: 10.1038/tp.2015.221] [Medline: 26859815]
95. Hyde KK, Novack MN, LaHaye N, Parlett-Pelleriti C, Anden R, Dixon DR, et al. Applications of Supervised Machine Learning in Autism Spectrum Disorder Research: a Review. Rev J Autism Dev Disord 2019 Feb 19;6(2):128-146. [doi: 10.1007/s40489-019-00158-x]
96. Thabtah F. Machine learning in autistic spectrum disorder behavioral research: A review and ways forward. Inform Health Soc Care 2019 Oct 13;44(3):278-297. [doi: 10.1080/17538157.2017.1399132] [Medline: 29436887]
97. Eslami T, Mirjalili V, Fong A, Laird AR, Saeed F. ASD-DiagNet: A Hybrid Learning Approach for Detection of Autism Spectrum Disorder Using fMRI Data. Front Neuroinform 2019 Nov 27;13:70 [FREE Full text] [doi: 10.3389/fninf.2019.00070] [Medline: 31827430]
98. Pratap A, Kanimozhiselvi C. Predictive assessment of autism using unsupervised machine learning models. IJAIP 2014;6(2):113. [doi: 10.1504/ijaip.2014.062174]
99. Stevens E, Dixon DR, Novack MN, Granpeesheh D, Smith T, Linstead E. Identification and analysis of behavioral phenotypes in autism spectrum disorder via unsupervised machine learning. Int J Med Inform 2019 Sep;129:29-36 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.05.006] [Medline: 31445269]

Abbreviations

ADDM: Autism and Developmental Disabilities Monitoring
ANN: artificial neural network
ASD: autism spectrum disorder
AUC: area under the curve
AWF: Autism Welfare Foundation
IPNA: Institute of Pediatric Neuro-disorder & Autism
KNN: K-nearest neighbor
NIH: National Institutes of Health
NIMH: National Institute of Mental Health
ROC: receiver operating characteristic

Edited by G Eysenbach; submitted 30.03.21; peer-reviewed by A Das, M Elbattah, V Jain; comments to author 22.04.21; revised version received 10.05.21; accepted 12.05.21; published 08.06.21.

Please cite as:
Haque MM, Rabbani M, Dipal DD, Zarif MII, Iqbal A, Schwichtenberg A, Bansal N, Soron TR, Ahmed SI, Ahamed SI
Informing Developmental Milestone Achievement for Children With Autism: Machine Learning Approach
JMIR Med Inform 2021;9(6):e29242
URL: https://medinform.jmir.org/2021/6/e29242
doi:10.2196/29242
PMID:33984830

Original Paper

Implementing Vertical Federated Learning Using Autoencoders: Practical Application, Generalizability, and Utility Study

Dongchul Cha[1,2], MD; MinDong Sung[1], MD; Yu-Rang Park[1], PhD

[1]Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, Republic of Korea

[2]Department of Otorhinolaryngology, Yonsei University College of Medicine, Seoul, Republic of Korea

Corresponding Author:
Yu-Rang Park, PhD
Department of Biomedical Systems Informatics
Yonsei University College of Medicine
50-1 Yonsei-ro
Sinchon-dong, Seodaemun-gu
Seoul, 03722
Republic of Korea
Phone: 82 2 2228 2363
Email: yurangpark@yuhs.ac

Abstract

Background: Machine learning (ML) is now widely deployed in our everyday lives. Building robust ML models requires a massive amount of data for training. Traditional ML algorithms require training data centralization, which raises privacy and data governance issues. Federated learning (FL) is an approach to overcome this issue. We focused on applying FL on vertically partitioned data, in which an individual's record is scattered among different sites.

Objective: The aim of this study was to perform FL on vertically partitioned data to achieve performance comparable to that of centralized models without exposing the raw data.

Methods: We used three different datasets (Adult income, Schwannoma, and eICU datasets) and vertically divided each dataset into different pieces. Following the vertical division of data, overcomplete autoencoder-based model training was performed for each site. Following training, each site's data were transformed into latent data, which were aggregated for training. A tabular neural network model with categorical embedding was used for training. A centrally based model was used as a baseline model, which was compared to that of FL in terms of accuracy and area under the receiver operating characteristic curve (AUROC).

Results: The autoencoder-based network successfully transformed the original data into latent representations with no domain knowledge applied. These altered data were different from the original data in terms of the feature space and data distributions, indicating appropriate data security. The loss of performance was minimal when using an overcomplete autoencoder; accuracy loss was 1.2%, 8.89%, and 1.23%, and AUROC loss was 1.1%, 0%, and 1.12% in the Adult income, Schwannoma, and eICU dataset, respectively.

Conclusions: We proposed an autoencoder-based ML model for vertically incomplete data. Since our model is based on unsupervised learning, no domain-specific knowledge is required in individual sites. Under the circumstances where direct data sharing is not available, our approach may be a practical solution enabling both data protection and building a robust model.

(JMIR Med Inform 2021;9(6):e26598) doi:10.2196/26598

KEYWORDS

federated learning; vertically incomplete data; privacy; machine learning; coding; data; performance; model; security; training; dataset; unsupervised learning; data sharing; protection

Introduction

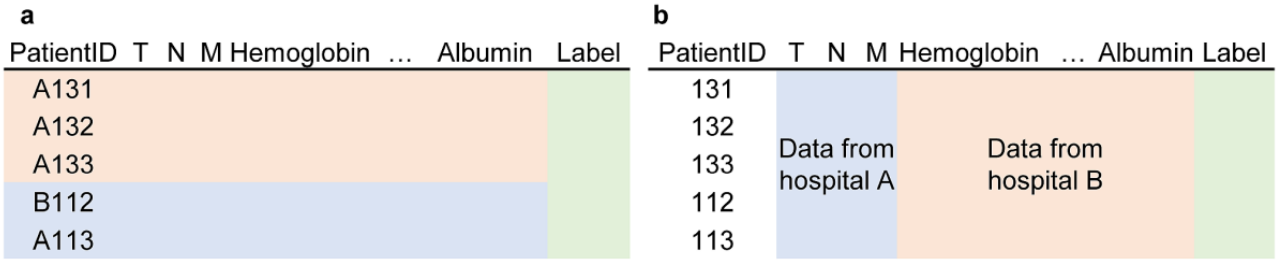
Machine learning (ML) is widely deployed in our daily lives, including, but not limited to, personalized digital media, product recommendations, and health care services. Building high-quality ML models requires a huge amount of data for training [1]. Conventional ML algorithms typically require the training data to reside where the models are trained. Recently, there has been an increasing level of concern about data privacy [2]. The EU General Data Protection Regulation and the US

Health Insurance Portability and Accountability Act are examples of regulations to secure sensitive information when gathering such information centrally. Moreover, as more data are needed for a robust ML model, raw data are a crucial asset. Sharing raw data raises data governance issues, making data owners hesitant about sharing their data.

An alternative approach to overcome such concerns is federated learning (FL). FL is a learning process in which the individual data owners train a model collaboratively without exposing the original data to others [2]. For the protection of data privacy, k-anonymity [3], l-diversity [4], and t-closeness [5] are well-established methods. Differential privacy [6] is another semantic method to add noise to data. Using such methods enables the aggregation of perturbed data with fewer concerns of exposing the original data. However, stronger protection of privacy requires stronger perturbations of the original data, which reduces the utility; in other words, this results in low-quality ML models. An alternative approach is homomorphic encryption [7], which offers training with encrypted data. However, training such a model is relatively slow, possibly making it impractical to be used in real-world applications [8].

FL could be divided into horizontal and vertical frameworks [2]. In horizontal FL, the data have the same feature space but are distributed among different organizations. In other words, all rows share the same columns but could originate from different sites. In contrast, vertical FL takes vertically partitioned data for training. For each row in the database, columns (features) originate from several different sites. Consider a database of colorectal cancer patients consisting of tumor-node-metastasis staging and laboratory results gathered from different hospitals, and we want to build an ML model to predict survival. In the horizontal FL setting, different organizations train ML models in their individual databases but share the same feature space (Figure 1a). However, in the vertical FL setting, individual tests are spread among different hospitals (eg, tumor stage in hospital A and laboratory tests in hospital B), and ML training is performed without aggregation of raw patient data (Figure 1b). Study results based on horizontal FL [9,10] show comparable performance to that of ML models trained centrally. For vertical FL, there is the possibility of logistic regression [11], linear regression [12], boosting model [13], a model capable of linear and logistic regressions, and neural network models [14].

Figure 1. Classification of federated learning. Assume a colorectal cancer patient dataset, and only the target label is gathered centrally. (a) In horizontally partitioned data, patients share the same feature space, but features are collected at different sites. (b) In vertically partitioned data, patient features are present in different sites.



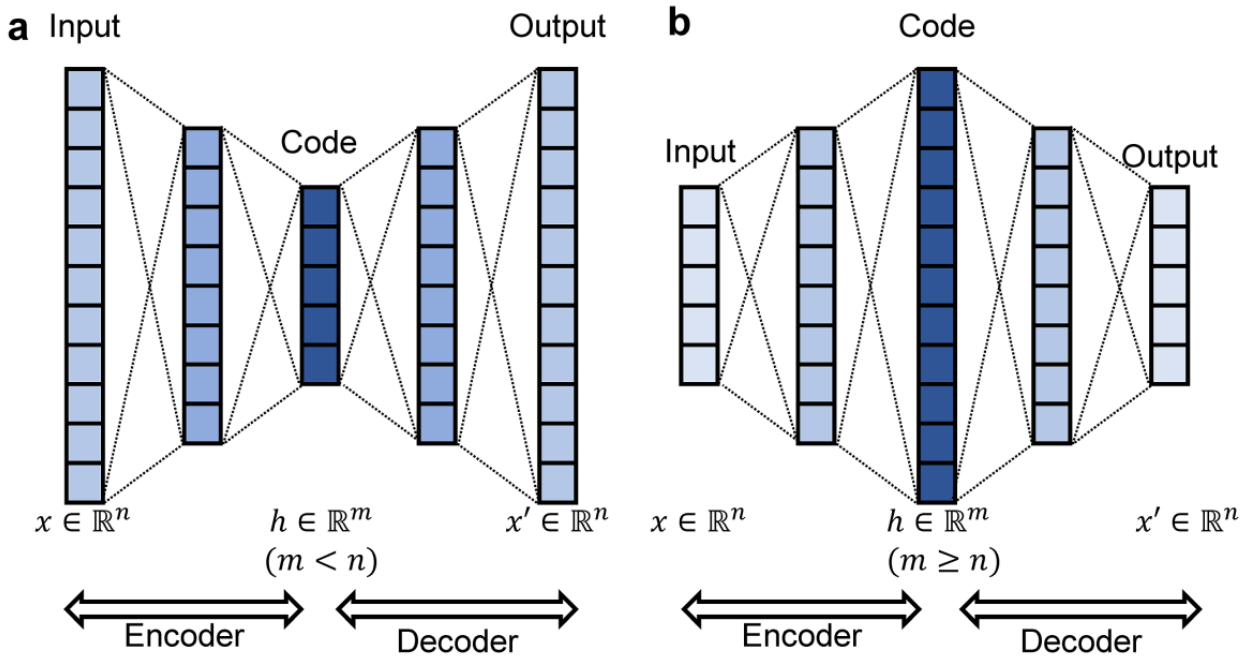
We here present a simple, practical, robust, and novel vertical FL method based on autoencoder neural networks [15], more specifically, an overcomplete autoencoder, in which hidden layers have a higher dimension than input layers. We tested our method in three datasets, including two medical datasets, to demonstrate generalizability and utility.

Methods

Overcomplete Autoencoder for the Latent Representation of Original Data

An autoencoder is a feed-forward neural network with the same inputs and outputs that are trained in an unsupervised manner. The network is fully connected and consists of an encoder and a decoder. The encoder transforms the input into a latent representation, and the decoder maps the latent representation back to the original input. During training, the machine learns both the encoder's and decoder's weights by minimizing the reconstruction loss. There are three main layers of an autoencoder: an input layer, hidden (including code) layer, and output layer. By adding a hidden layer with constraints such as fewer dimensions than the given input (Figure 2a, $h \in \mathbb{R}^m$ [$m<n$]) the machine tries to learn essential features in the given input. Since a conventional autoencoder reduces dimension, there is an inevitable loss of information. In an overcomplete autoencoder, hidden layers are larger than or equal to the input layer (Figure 2b, $h \in \mathbb{R}^m$ [$m \geq n$]). By having more feature space in the code layer, information loss could be minimized, especially when datasets have a small number of features. Additionally, latent representation differs from the original input data, enabling both security and performance.

Figure 2. The autoencoder network, which is an unsupervised machine learning algorithm. Input and output are the same; thus, they have identical feature space. (a) The conventional autoencoder has a latent space dimension smaller than the input space (m<n). (b) The overcomplete autoencoder has equal or higher dimensions in the latent space (m≥n).



Datasets and Vertical Division of Data

Adult Income Dataset

The adult income dataset [16] has two labels: whether or not a person earns over 50,000 per year, with eight categorical and six continuous variables as input variables. The dataset included 37,155 individuals with a salary ≤50,000 and 11,687 individuals with a salary >50,000 per year. We randomly sampled from the 11,687 individuals with a salary under 50,000 to balance the dataset (random undersampling), so that the total dataset comprised 23,374 individuals, and set the prediction chance level to 50%. We vertically divided this dataset into three pieces, assuming three different organizations possessing partial data over individuals (Table 1).

Table 1. Dataset composition and training parameters with division to simulate vertically partitioned data.

Dataset	Division	Dataset size (number of rows)	Feature dimension	Autoencoder layers	Aggregated dimension
Adult income	3 sites	23,374	5, 5, 4	64-128-64	384×23,374
Schwannoma	3 sites	50	7, 3, 5	64-128-64	384×50
eICU	7 sites	15,762	3, 4, 9, 3, 3, 4, 6	64-128-64	896×15,762

Vestibular Schwannoma Dataset

The vestibular schwannoma dataset [17] is an anonymized, private, medical dataset to predict hearing disabilities following surgery. We included this dataset to demonstrate its feasibility in a relatively low number of training samples with sparse data. The dataset included 50 patients, one categorical variable, 14 continuous variables as input, and binary classification labels as output. Since the dataset had 22 and 28 binary target labels, no additional undersampling was performed. The data were vertically split into three sites (Table 1).

The eICU Collaborative Research Database

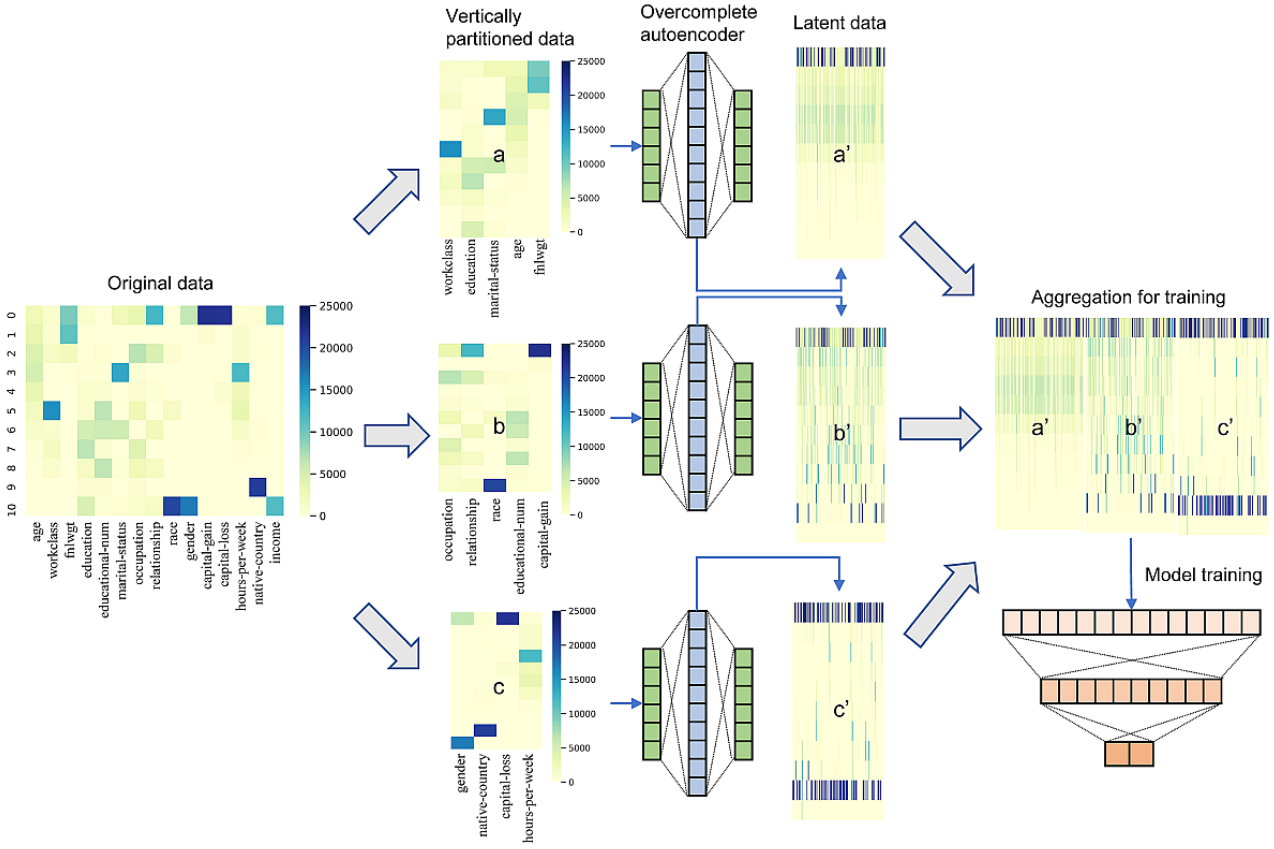
The eICU collaborative research database [18] is a database containing variables used in deriving Acute Physiologic Assessment and Chronic Health Evaluation (APACHE) [19] scores to predict a given patient's mortality (binary classification). The initial database contained 148,532 intensive care unit (ICU) stays with APACHE version IVa. We only included ICU stays with more than 15 (62.5%) nonnull values, excluding 712 ICU stays. We also excluded 15,968 rows without labels. Therefore, a total of 131,852 rows (ICU stays) were used, 7881 of which were labeled as expired. We randomly picked 7881 alive rows to rule out the class imbalance problem, making the baseline dataset contain 15,762 rows, and vertically divided the dataset into 7 sites (Table 1).

Training Workflow and Parameters

All three datasets were vertically divided. In all three datasets, we assumed that a third-party relay server performs data alignment between different servers. For example, the third row in server A is also the third row in servers B and C. To test the generalizability of our approach, we divided the dataset into various numbers (Table 1). Following the vertical division of

data, overcomplete autoencoder-based model training was performed for each site (Figure 3 a, b, c). Following training, each site's latent data (Figure 3 a', b,' c,' representations in the code layer) were aggregated for training. We used PyTorch [20] with the Fastai [21] library for this task. Each site was vertically divided to simulate vertically partitioned data among different sites. Accuracy and area under the receiver operating characteristic curve (AUROC) were used as the evaluation metrics for classification tasks.

Figure 3. The workflow of vertical federated learning using overcomplete autoencoder. The UCI adult income dataset is illustrated as an example. The dataset consists of 14 features and 1 target label (income). Original data are vertically divided into several datasets, three in this case, to assume data distribution among different sites. The heatmaps show each feature with prevalence. Each site (a, b, c) trains an autoencoder and transmits latent data, which are differently distributed, as seen in the heatmaps (a', b,' c'). The latent data are aggregated for training to a server, and the server performs model training. The accuracy of models created using the original data versus aggregated latent data is compared.



Autoencoder models were trained using an initial learning rate of 0.01 and a learning rate decay of 0.99. There is a concern that the ML algorithm might learn an identity function, which may not correctly perturb (or encode) the data. However, a previous study [22] using stochastic gradient descent (SGD) when training resulted in a useful data representation. In addition to using SGD, we also used a weight decay of 0.1 to prevent the autoencoder models from learning the identity function and overfitting.

A tabular neural network model with categorical embedding was used when training. A centrally based model was used as a baseline model. For each vertically split data, both models were trained: an ML model based on each vertically split dataset (Figure 3 a, b, c) and an ML model based on latent representations of each split dataset (Figure 3 a', b', c'). Finally, the central-based model was compared to the latent data aggregated model for benchmarking our vertical federated neural network model.

Code Availability

Since autoencoders are widely implemented in various environments, we do not offer the source code publicly. However, codes will be available upon request to the corresponding author for noncommercial, educational purposes.

Results

Transformation of the Data to Latent Representations

The autoencoder-based network successfully transformed the original data into latent representations with no domain knowledge applied. These altered data were different from the original data in terms of both the feature space and data distributions (Figure 3 a', b', c'), indicating appropriate data security.

Classification Performance

Following latent data aggregation, we tested the built model against centralized models and individually trained models using the vertically incomplete original data and latent data (Table 2; see Multimedia Appendix 1 for a detailed division of the feature

space). The performance of the autoencoder increased as the number of the code layers increased (see Multimedia Appendix 2 for detailed results). Of note, since we used categorical embeddings when putting categorical variables to the autoencoder and tabular neural network model, latent representations of the original data were continuous variables. The adult income and eICU datasets, which have a relatively large number of rows, did not suffer from a fluctuation of accuracy and AUROC. Although the schwannoma dataset, which only has 50 rows, showed a fluctuation of accuracy and AUROC among different sites, the overall accuracy and AUROC penalties were still acceptable. The eICU dataset had abundant feature space and was vertically divided into seven sites. There was still a minimal loss of accuracy and AUROC, implying good utility while preserving data privacy (Table 2).

Table 2. Classification results of the three datasets.

Site	Adult income dataset		Schwannoma dataset		eICU dataset	
	Accuracy	AUROC[a]	Accuracy	AUROC	Accuracy	AUROC
Central						
Before VFL[b]	0.83	0.91	0.90	0.84	0.81	0.89
After VFL[c]	0.82	0.90	0.82	0.84	0.80	0.88
Difference[d]	–1.20	–1.10	–8.89	0	–1.23	–1.12
A						
Before VFL	0.81	0.89	0.82	0.81	0.70	0.72
After VFL	0.77	0.83	0.78	0.86	0.70	0.72
Difference	–4.94	–6.74	–4.88	+6.17	0	0
B						
Before VFL	0.81	0.90	0.76	0.82	0.73	0.80
After VFL	0.77	0.83	0.78	0.83	0.72	0.79
Difference	–4.94	–7.78	+2.63	+1.22	–1.37	–1.25
C						
Before VFL	0.67	0.73	0.48	0.60	0.55	0.57
After VFL	0.76	0.83	0.62	0.71	0.56	0.57
Difference	+13.43	+13.70	+29.17	+18.33	1.82	0

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

[b]VFL: vertical federated learning.

[c]Corresponding to the latent representation of original data (central, A, B, or C) in the code layer.

[d]The difference is compared between AUROCs in classification tasks.

Discussion

Principal Results

We have successfully transformed original data into latent representations and trained ML models with perturbed data, resulting in minimal loss of accuracy while preserving data privacy. In an autoencoder network, ML models learn data representation in an unsupervised manner. Therefore, no domain knowledge is required to train the model. Since the code layer has more layers than the input layer, resulting in high dimensionality, this method requires more computing power compared to that required for traditional autoencoders. However, loss of information is minimal, even though the data are severely perturbed (Figure 3 a', b', c'). Although slight, there was still a loss of accuracy and AUROC in the trained ML model (Table 2). We suspect this was due to redundant information generated by the network, which acts as noise when training an ML model. The model's design is somewhat similar to local differential privacy [23] in that each site performs training of ML models independently before sending the perturbed data to a central server. The main difference is that differential privacy has an equal number of feature space dimensions as in the original dataset, whereas our approach alters the feature space to a predefined number of hidden layers.

To check its generalizability, we tested three different datasets with various vertically split datasets. Training the ML model worked well in all datasets, even with a relatively small number of rows. Moreover, some datasets were vigorously divided, but the accuracy remained comparable to that of the centralized ML model. In real-life practice, our model may enable building an ML model without the direct exchange of sensitive information among different data owners. For example, a patient may undergo some routine complete blood count test in one hospital, obtain imaging studies in another, and perform electrolyte tests in the other hospital. When building a classifier model, three sites (hospitals) may train our proposed model individually and

share the latent feature space to train the model without directly exposing the patient's data.

Comparison With Prior Work and Limitations

Earlier works on federated ML using vertically partitioned data focused on the logistic regression [11], linear regression [12], and boosting [13] models. Hardy et al [12] also utilized additively homomorphic encryption. In their study, both nonprivate and federated settings showed the same accuracy, AUROC, and F1 score. However, the training time was in the order of hours per epoch in high-performance cloud-based machines, which may not be practical. Cheng et al [13] proposed SecureBoost, which exhibited a performance comparable to that of nonprivacy-preserving gradient boosting machine models. They theoretically proved that if both ML models have identical initialization and parameters, the SecureBoost algorithm is lossless; that is, the model shows comparable accuracy to the nonfederated boosting model. Mohassel et al [14] suggested a system capable of linear regression, logistic regression, and neural networks. They used a secure multiparty computation [24] framework with two noncolluding servers (secure two-party computation) to train ML models in a privacy-preserving fashion. The results were promising, but the authors suggested that the neural network model is not yet practical due to the high number of interactions and communications costs.

In this study, we assumed that each client performs autoencoder-based data alteration; therefore, file transmission happens only once when building an ML model. Continuous network connections are not necessary. In addition, training an overcomplete autoencoder is not computationally expensive, which makes our proposed model practical. Similar to other privacy-preserving methods, our model ensures no data leakage beyond data owners. Moreover, we have demonstrated that our approach enables more than two participants to aggregate the latent data, allowing more features per person as the number of participating institutions increases.

Our study has limitations. First, even though the data are differently shaped, data owners still need to transmit the coded data to a central location, which may have room for reverse engineering. However, unless the original feature space is revealed to the recipient, reverse engineering may be difficult. Moreover, the latent space is much bigger than the original feature space, making data transmission redundant. Given sufficient network capacity, this should not be a critical issue. Second, more rigorous results are genuinely needed using cross-validation. Last but not least is the explainability of the model. Since the model transforms feature space into latent space, each feature's meaning in the aggregated data is somewhat different; it cannot be directly associated with the original feature space. Indirectly, site-wise comparison of accuracy using only part of available data could be used to measure feature importance, but future studies should be performed to overcome this limitation.

Conclusions

We proposed an overcomplete autoencoder–based ML model for vertically incomplete data. Since our model is based on unsupervised learning, no domain-specific knowledge is required in individual sites. Under the circumstances where direct data sharing is not available, our approach may be a practical solution enabling both data protection and building a robust model.

Acknowledgments

This study received support from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number HI19C1015). This study was also supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (NRF2019M3E5D4064682).

Authors' Contributions

DC and YRP designed the study. DC proposed the algorithm and wrote the machine learning code with MDS. DC and MDS wrote the first draft of the manuscript, and all other authors reviewed, modified, and approved the final manuscript. DC and YRP are guarantors of the study. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Feature distribution of individual sites in three datasets.
[DOCX File , 15 KB - medinform_v9i6e26598_app1.docx]

Multimedia Appendix 2
Classification results of the three datasets, and comparison with conventional undercomplete autoencoders.
[DOCX File , 26 KB - medinform_v9i6e26598_app2.docx]

References

1. Halevy A, Norvig P, Pereira F. The unreasonable effectiveness of data. IEEE Intell Syst 2009 Mar;24(2):8-12. [doi: 10.1109/mis.2009.36]
2. Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning. ACM Trans Intell Syst Technol 2019 Feb 28;10(2):1-19. [doi: 10.1145/3298981]
3. Sweeney L. k-anonymity: A model for protecting privacy. Int J Unc Fuzz Knowl Based Syst 2012 May 02;10(05):557-570. [doi: 10.1142/s0218488502001648]
4. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. L-diversity: privacy beyond k-anonymity. In: ACM Trans Knowl Discov Data. 2007 Mar Presented at: 22nd International Conference on Data Engineering (ICDE'06); April 3-7, 2006; Atlanta, GA p. 3. [doi: 10.1145/1217299.1217302]
5. Li N, Li T, Venkatasubramanian S. t-closeness: Privacy beyond k-anonymity and l-diversity. 2007 Apr 15 Presented at: IEEE 23rd International Conference on Data Engineering; April 15, 2007; Istanbul, Turkey p. 106-115. [doi: 10.1109/icde.2007.367856]
6. Dwork C, Roth A. The algorithmic foundations of differential privacy. FNT Theor Comput Sci 2014;9(3-4):211-407. [doi: 10.1561/0400000042]
7. Rivest R, Adleman L, Dertouzos M. On data banks and privacy homomorphisms. Found Secure Comput 1978;4(11):169-180.
8. Naehrig M, Lauter K, Vaikuntanathan V. Can homomorphic encryption be practical? 2011 Oct 17 Presented at: CCSW '11: Proceedings of the 3rd ACM workshop on Cloud computing security workshop; October 17, 2011; Chicago, IL p. 113-124. [doi: 10.1145/2046660.2046682]
9. Sheller M, Reina G, Edwards B, Martin J, Bakas S. Multi-institutional deep learning modeling without sharing patient data: a feasibility study on brain tumor segmentation. Brainlesion 2019;11383:92-104 [FREE Full text] [doi: 10.1007/978-3-030-11723-8_9] [Medline: 31231720]
10. Li W, Milletarì F, Xu D, Rieke N, Hancox J, Zhu W. Privacy-preserving federated brain tumour segmentation. 2019 Oct 10 Presented at: International Workshop on Machine Learning in Medical Imaging; October 13, 2019; Shenzhen, China p. 133-141. [doi: 10.1007/978-3-030-32692-0_16]
11. Li Y, Jiang X, Wang S, Xiong H, Ohno-Machado L. VERTIcal Grid lOgistic regression (VERTIGO). J Am Med Inform Assoc 2016 May;23(3):570-579 [FREE Full text] [doi: 10.1093/jamia/ocv146] [Medline: 26554428]
12. Hardy S, Henecka W, Ivey-Law H, Nock R, Patrini G, Smith G. Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption. arXiv preprint. 2017 Nov 29. URL: https://arxiv.org/abs/1711.10677 [accessed 2021-05-28]
13. Cheng K, Fan T, Jin Y, Liu Y, Chen T, Yang Q. Secureboost: A lossless federated learning framework. arXiv preprint. 2019 Jan 25. URL: https://arxiv.org/abs/1901.08755 [accessed 2021-05-28]
14. Mohassel P, Zhang Y. SecureML: A system for scalable privacy-preserving machine learning. 2017 May 22 Presented at: 2017 IEEE Symposium on Security and Privacy (SP); May 22, 2017; San Jose, CA p. 22-26. [doi: 10.1109/sp.2017.12]
15. Kramer MA. Nonlinear principal component analysis using autoassociative neural networks. AIChE J 1991 Feb;37(2):233-243. [doi: 10.1002/aic.690370209]
16. Asuncion A, Newman D. Adult Data Set. UCI machine learning repository. 2007. URL: https://archive.ics.uci.edu/ml/datasets/Adult [accessed 2021-05-28]
17. Cha D, Shin SH, Kim SH, Choi JY, Moon IS. Machine learning approach for prediction of hearing preservation in vestibular schwannoma surgery. Sci Rep 2020 Apr 28;10(1):7136. [doi: 10.1038/s41598-020-64175-1] [Medline: 32346085]
18. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data 2018 Sep 11;5:180178. [doi: 10.1038/sdata.2018.178] [Medline: 30204154]
19. Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. Crit Care Med 2006 May;34(5):1297-1310. [doi: 10.1097/01.CCM.0000215112.84523.F0] [Medline: 16540951]
20. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G. Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems. 2019 Presented at: 33rd Conference on Neural Information Processing Systems (NeurIPS 2019); December 8-14, 2019; Vancouver, BC p. 8026-8037. [doi: 10.7551/mitpress/11474.003.0014]
21. Howard J, Gugger S. Fastai: A layered API for deep learning. Information 2020 Feb 16;11(2):108. [doi: 10.3390/info11020108]
22. Bengio Y. Learning deep architectures for AI. FNT Machine Learn 2009;2(1):1-127. [doi: 10.1561/2200000006]
23. Zhao Y, Zhao J, Yang M, Wang T, Wang N, Lyu L, et al. Local differential privacy-based federated learning for internet of things. IEEE Internet Things J 2021 Jun 1;8(11):8836-8853. [doi: 10.1109/jiot.2020.3037194]
24. Zhao C, Zhao S, Zhao M, Chen Z, Gao C, Li H, et al. Secure multi-party computation: theory, practice and applications. Inf Sci 2019 Feb;476(7):357-372. [doi: 10.1016/j.ins.2018.10.024]

Abbreviations

APACHE: Acute Physiologic Assessment and Chronic Health Evaluation
AUROC: area under the receiver operating characteristics curve
FL: federated learning
ICU: intensive care unit
ML: machine learning
SGD: stochastic gradient descent

Edited by C Lovis; submitted 17.12.20; peer-reviewed by KW Kim, M Elbattah; comments to author 13.01.21; revised version received 31.01.21; accepted 03.05.21; published 09.06.21.

Please cite as:
Cha D, Sung M, Park YR
Implementing Vertical Federated Learning Using Autoencoders: Practical Application, Generalizability, and Utility Study
JMIR Med Inform 2021;9(6):e26598
URL: https://medinform.jmir.org/2021/6/e26598
doi:10.2196/26598
PMID:34106083

Original Paper

Enhancing Obstructive Sleep Apnea Diagnosis With Screening Through Disease Phenotypes: Algorithm Development and Validation

Daniela Ferreira-Santos[1,2], BSc, MSc; Pedro Pereira Rodrigues[1,2], BSc, MSc, PhD

[1]MEDCIDS-FMUP – Community Medicine, Information and Decision Sciences, Faculty of Medicine of the University of Porto, Porto, Portugal
[2]CINTESIS – Center for Health Technology and Services Research, Porto, Portugal

Corresponding Author:
Daniela Ferreira-Santos, BSc, MSc
MEDCIDS-FMUP – Community Medicine, Information and Decision Sciences
Faculty of Medicine of the University of Porto
Rua Dr. Plácido da Costa, s/n
Porto, 4200-450
Portugal
Phone: 351 22 551 3622
Email: danielasantos@med.up.pt

Abstract

Background: The American Academy of Sleep Medicine guidelines suggest that clinical prediction algorithms can be used in patients with obstructive sleep apnea (OSA) without replacing polysomnography, which is the gold standard.

Objective: This study aims to develop a clinical decision support system for OSA diagnosis according to its standard definition (apnea-hypopnea index plus symptoms), identifying individuals with high pretest probability based on risk and diagnostic factors.

Methods: A total of 47 predictive variables were extracted from a cohort of patients who underwent polysomnography. A total of 14 variables that were univariately significant were then used to compute the distance between patients with OSA, defining a hierarchical clustering structure from which patient phenotypes were derived and described. Affinity from individuals at risk of OSA phenotypes was later computed, and cluster membership was used as an additional predictor in a Bayesian network classifier (model B).

Results: A total of 318 patients at risk were included, of whom 207 (65.1%) individuals were diagnosed with OSA (111, 53.6% with mild; 50, 24.2% with moderate; and 46, 22.2% with severe). On the basis of predictive variables, 3 phenotypes were defined (74/207, 35.7% low; 104/207, 50.2% medium; and 29/207, 14.1% high), with an increasing prevalence of symptoms and comorbidities, the latter describing older and obese patients, and a substantial increase in some comorbidities, suggesting their beneficial use as combined predictors (median apnea-hypopnea indices of 10, 14, and 31, respectively). Cross-validation results demonstrated that the inclusion of OSA phenotypes as an adjusting predictor in a Bayesian classifier improved screening specificity (26%, 95% CI 24-29, to 38%, 95% CI 35-40) while maintaining a high sensitivity (93%, 95% CI 91-95), with model B doubling the diagnostic model effectiveness (diagnostic odds ratio of 8.14).

Conclusions: Defined OSA phenotypes are a sensitive tool that enhances our understanding of the disease and allows the derivation of a predictive algorithm that can clearly outperform symptom-based guideline recommendations as a rule-out approach for screening.

(JMIR Med Inform 2021;9(6):e25124) doi:10.2196/25124

KEYWORDS

obstructive sleep apnea; screening; risk factors; phenotypes; Bayesian network classifiers

Introduction

Background

Obstructive sleep apnea (OSA) is a common sleep-related breathing disorder characterized by clinical symptoms (eg, daytime sleepiness) and at least five events per hour of narrowing (apnea or hypopnea) of the upper airway that impairs normal ventilation during sleep [1]. An apnea consists of a cessation of airflow higher than 90% of the baseline, a hypopnea is a reduction in airflow along with a decreased saturation of 3% from pre-event baseline and/or associated with an arousal, and the apnea-hypopnea index (AHI) is the number of such events per hour of sleep. OSA prevalence has been underestimated, with studies varying significantly, both in the population being studied and in OSA definition. A study using a simpler hypopnea definition (4% desaturation) estimated a prevalence of 14% in men and 5% in women [2]. In 2 other studies, the prevalence was substantially higher but was estimated for specific populations, such as patients being evaluated for bariatric surgery [3] or patients who have had a transient ischemic attack or stroke [4], reaching values of 70% and 72%, respectively. The latest study by Benjafield et al [5] estimated that 936 million adults have OSA; in Portugal, it represents 17%, and approximately 74% have moderate to severe OSA. Overall, this disease is largely unrecognized and undiagnosed, representing a significant burden to the health care system [6], especially for patients who remain untreated or at an increased risk of developing cardiovascular disease, metabolic dysregulation, or diabetes [1,7-11]. The failure to clinically recognize OSA leads to significant morbidity and mortality, making it essential to anticipate its recognition, diagnosis, and treatment [1]. OSA diagnosis, for which a comprehensive sleep evaluation (sleep history and physical examination) plus polysomnography (PSG) is the gold standard [1], can effectively decrease health care utilization and costs, whereas timely treatment can improve quality of life, lower the rates of motor vehicle crashes, and reduce the risk of chronic health consequences [12].

In 2017, a new clinical practice guideline for diagnostic testing for adults with OSA was issued by the American Academy of Sleep Medicine (AASM) [1], updating 2 previous AASM guidelines from 2005 [8] and 2007 [13]. Of the 9 PICO (patient, population or problem, intervention, comparison, and outcome) questions raised in this new guideline, the task force reported insufficient evidence to directly address the first one: "In adult patients with suspected OSA, do clinical prediction algorithms accurately identify patients with a high pretest probability for OSA compared to history and physical exam?," as no studies comparing the efficacy of clinical prediction algorithms with clinical history and physical examination were identified. Therefore, they compared the efficacy of clinical prediction algorithms with PSG, crafting recommendation 1: "We recommend that clinical tools, questionnaires and prediction algorithms not to be used to diagnose OSA in adults, in the absence of PSG," affirming that clinical prediction algorithms can, however, be used in patients with suspected OSA, as long as not to establish the need for PSG or to become a substitute for PSG. Rather, these tools can be more helpful, in specialties other than sleep-oriented ones, to identify patients with an increased risk for OSA.

Objective

In this study, we aim to establish a new clinical prediction algorithm to allow OSA screening (high pretest probability for OSA) based on demographics, physical examination, clinical history, and comorbidities, using standard OSA definition (AHI ≥5 plus symptoms), extending traditional approaches that assess only preestablished symptoms, such as snoring, witnessed apneas, and excessive daytime sleepiness.

Methods

Overview

Using retrospective data from a cohort of patients who underwent PSG, after proper referral by a physician, significant predictive variables were selected and used to compute distances among patients with OSA, which supported a clustering algorithm to derive patient phenotypes from resulting clusters, with missing data being analyzed and imputed as needed. To assess the consistency of our phenotypes, each healthy individual was also tested against the clustering structure, and the resulting phenotyping was analyzed. Then, to assess the benefit of this phenotyping strategy, cluster membership was used as an additional predictive variable and included in a Bayesian network classifier, with validity compared with an equivalent classifier without phenotype information, following the 2015 STARD (Standards for Reporting Diagnostic Accuracy Studies) guideline.

Patients

Data from patients referred to undergo PSG at Vila Nova de Gaia and Espinho Hospital Center Sleep Laboratory were retrospectively collected. Patients who underwent PSG between January and May 2015 were included if they were aged >18 years and were suspected of having OSA. Nonetheless, exclusion criteria included patients already diagnosed (performing positive airway pressure therapies), patients suspected of having another sleep disease, patients with severe lungs or neurological conditions, and pregnant women. In case of multiple examinations of the same patient, the one with the best sleep efficiency was selected. This study was approved by the Ethics Commission of Vila Nova de Gaia and Espinho Hospital Center, in accordance with the Declaration of Helsinki.

Predictive Variables

An author-performed literature review on PubMed (April 19, 2015) supported the definition of the relevant variables to be collected from medical and/or sleep laboratory records in which the presence or absence of each information was assessed by a physician, resulting in a total of 47 predictive variables, all in accordance with the current and previous OSA guidelines. The search contained "risk factors," "sleep apnea, obstructive," and "diagnosis" as MeSH terms, obtaining 1397 articles, of which 47 were used for variable definition (full review description and references used in this phase are not shown for space purposes but can be provided on request). Selected variables included basic demographic data (gender and age), physical examination

(BMI, neck and abdominal circumferences, modified Mallampati classification, and craniofacial and upper airway abnormalities), clinical history (daytime sleepiness, snoring, witnessed apneas, gasping and/or choking, sleep fragmentation, nonrepairing sleep, behavior changes, decreased concentration, morning headaches, decreased libido, sleeping body position, sleep efficiency, participation in vehicle crashes, truck driver activity, driving sleepiness, nocturia, alcohol consumption, smoking, coffee intake, use of sedatives before sleep, family history or genetic evidence, and Epworth Sleepiness Scale), and comorbidities and cointerventions (stroke, myocardial infarction, pulmonary infarction, arterial or pulmonary hypertension, congestive heart failure, arrhythmias, respiratory changes, diabetes, dyslipidemia, renal failure, hypothyroidism, gastroesophageal reflux, anxiety and/or depression, insomnia, glaucoma, pacemaker or implantable cardioverter-defibrillator, and bariatric surgery).

Data Set Description

Clinical data from each patient (47 predictive variables plus the outcome) were extracted from the central clinical data registry (all records were fulfilled by a physician) along with sleep laboratory data and adequately anonymized to ensure patient privacy. Original files included structured demographic data, structured PSG reports, and unstructured textual annotations from the medical records, with many abbreviations and short-form text. The outcome measure was obtained from the AHI, categorized as mild (AHI between 5 and 14), moderate (AHI between 15 and 29), and severe (AHI >30). Given the categorical characteristic of our modeling strategies, all continuous variables were discretized, and the following common cutoffs were extracted from the literature: (1) age (20-44 years, 45-64 years, and 65-90 years), (2) BMI (<25 kg/m^2 as normal weight, 25-30 kg/m^2 as overweight, and ≥30 kg/m^2 as obese), (3) female neck circumference (≤37 cm as normal and >38 cm as increased), (4) male neck circumference (≤41 cm as normal and >42 cm as increased), (5) female abdominal circumference (≤80 cm as normal and >81 cm as increased), (6) male abdominal circumference (≤94 cm as normal and >95 cm as increased), (7) Epworth Sleepiness Scale (0-10 as normal and 11-24 as excessive daytime sleepiness), and (8) AHI (0-4 as normal, 5-14 as mild, 15-29 as moderate, and ≥30 as severe).

Missing Data Imputation

Although we had all the electronic clinical records from the included patients, after screening all unstructured text reports, some predictive variables were not fully present or described, as physicians normally do not mention the absence of a disease or it could only be noted in paper records (missing data proportions ranged from 0% for gender to 97% for bariatric surgery). In our previous study [14], we studied the impact of missing data imputation, using nearest neighbor (NN) strategies, on the structure learning of Bayesian network classifiers for OSA diagnosis, concluding that it can expand the body of evidence for modeling without compromising validity. In this study, we followed the same strategy: (1) variables with more than 80% missing values were removed from the analysis (ie, behavior changes, decreased libido, decreased concentration, pulmonary infarction, glaucoma, and bariatric surgery); (2) remaining variables were ranked by the proportion of missing values; (3) data imputation started using only complete and outcome-wise statistically significant variables (P<.20), imputing incomplete likewise significant variables; and (4) remaining incomplete variables were then imputed stepwise by increasing the proportion of missing values per variable. All imputations were performed using majority voting from the 10 NNs/patients.

Clinical Prediction Algorithm

Aspiring to a more personalized approach to evaluate patients with OSA and targeting to recognize high pretest probability for OSA, cluster analysis (a statistical approach for studying the relationship present among groups of patients or variables [7]) was applied to distinguish whether there are different subgroups of patients with different clinical presentations, that is phenotypes. Clustering has been widely used in health research, particularly in the analysis of gene expression [15], asthma [16], chronic obstructive pulmonary disease [17], fibromyalgia [18], Parkinson disease [19], and sleep apnea [20-22]. The aim is to identify clusters of patients who are similar among themselves, although significantly different from patients of other clusters [7]. As expected, different clusters created from predictive variables express different disease risks, hence defining risk-aligned phenotypes.

Connectivity-Based Clustering

In this study, we applied a hierarchical clustering algorithm to obtain a hierarchy of possible solutions, ranging from one single group with all patients to having every single patient separated from each other. This process, where a cluster hierarchy is created, is based on the distance between data observations (ie, patients), giving as output a dendrogram (a tree diagram that presents different clustering definitions for all possible numbers of clusters, from which the user might choose the desired number of clusters after inspecting the intracluster and intercluster distances of each possible cut point). Therefore, the definition of the distance function is a crucial step in the application of this technique, especially in categorical data, as an incorrect distance can easily lead to biased results with potentially serious consequences to the conclusions drawn.

In this study, we computed the distance measure between 2 patients, a and b, based only on significant variables (univariate significant association with the outcome, for a 20% significance level in both the original and imputed data sets, using chi-square and Fisher exact tests), and each variable was weighted according to the corresponding crude odds ratio for the severe level, as follows:

This distance encoded the similarity between patients weighted by the contribution of each variable toward the outcome, regularized for significant variables only, and was subsequently used in hierarchical clustering with Ward linkage, leading to a complete dendrogram. Afterward, the obtained OSA clusters were defined by inspecting the outcome proportion by cluster and the corresponding 95% CIs.

Phenotypes Consistency

To assess whether predetermined phenotypes would also help in segmenting healthy patients, each healthy patient was assigned to the closest phenotype using the aforementioned distance measure and the same significant variables, determining the distance between each healthy patient and obtained OSA cluster. The resulting clustering definition was then described and analyzed, as was done for the cohort of patients with OSA.

Phenotypes Predictive Value

To assess whether the phenotypes could encode any predictive value, Bayesian network classifiers were built with and without cluster information as a predictive variable. First, a naïve Bayesian network classifier was induced using the selected variables. Then, assigned cluster was also included in the model as a parent node of all independent variables. Validity was then assessed and compared using leave-one-out and 10 times twofold cross-validation strategies, comparing validity measures, such as sensitivity, specificity, accuracy, predictive values, area under the receiver operating characteristic (ROC) curve, likelihood ratios, posttest odds and posttest probabilities, and diagnostic odds ratio.

Statistical Software

R 3.2.2 (R Development Core Team) software [23] was used on every statistical step of this work: discretization of continuous variables (package car [24]), descriptive and comparative analyses (packages gmodels [25] and epitools [26]), missing data analysis (package summarytools [27]), missing data imputation (package DMwR [28]), hierarchical clustering (package stats [23]), Bayesian network inference (packages bnlearn [29] and gRain [30]), and ROC curve analysis (package pROC [31]). Bayesian networks were visually inspected using SamIam software (developed by the University of California, Los Angeles) [32].

Results

Baseline Characteristics

Of the 318 patients included, 207 (65.1%) had OSA. Of these 207 patients, 111 (53.6%) were classified as mild, 50 (24.2%) as moderate, and 46 (22.2%) as severe. Baseline characteristics of patients with OSA and the proportion of missing values for each predictive variable are described below in Table 1 (original data) and in Multimedia Appendix 1 (for the curated data, after missing data imputation).

Patients with OSA had a mean age of 61 (SD 11) years, being slightly older in the moderate subgroup (24/50, 48%; aged >65 years), whereas the proportion of males was higher in the moderate (40/50, 80%) and severe (35/46, 76%) subgroups. Beyond these 2 variables, only sleep efficiency was found to be complete (no missing data), and no differences were found across OSA levels (P=.65). For the remaining variables, distributions were computed before and after data imputation.

The presence of witnessed apneas (109/169, 64.5%), nonrepairing sleep (93/183, 50.8%), nocturia (99/136, 72.8%), stroke (23/44, 52%), arterial hypertension (136/159, 85.5%), diabetes (62/99, 63%), and dyslipidemia (125/148, 84.5%) were more prevalent in patients with OSA than in healthy patients, whereas the opposite was observed in family history (14/77, 18%), pulmonary hypertension (15/117, 12.8%), congestive heart failure (26/138, 18.8%), arrhythmias (17/99, 17%), pacemaker or implantable cardioverter-defibrillator (10/91, 11%), and respiratory changes (81/185, 43.8%). After data imputation, the same variables remained different across OSA levels, except for family history. Only variables significantly associated with the outcome (P<.20) on both the original and curated data sets were further considered for the clustering process.

Table 1. Descriptive analysis of patients with obstructive sleep apnea (absolute and relative frequencies are presented, and *P* values are the results of chi-square tests unless otherwise specified).

Characteristic	Mild (n=111), n (%)	Moderate (n=50), n (%)	Severe (n=46), n (%)	Total (N=207), n (%)	*P* value	Missing, n (%)
Gender (male)	72 (64.9)	40 (80.0)	35 (76.1)	147 (71.0)	*.10* [a]	207 (0.0)
Age (years)					*.18* [b]	207 (0.0)
20-44	7 (6.3)	5 (10.0)	6 (13.0)	18 (8.7)		
45-64	67 (60.4)	21 (42.0)	24 (52.2)	112 (54.1)		
65-90	37 (33.3)	24 (48.0)	16 (34.8)	77 (37.2)		
BMI (kg/m^2)					*.05* [b]	169 (18.4)
Normal weight	14 (15.6)	1 (2.6)	1 (2.5)	16 (9.5)		
Overweight	34 (37.8)	21 (53.8)	16 (40.0)	71 (42.0)		
Obesity	42 (46.7)	17 (43.6)	23 (57.5)	82 (48.5)		
Increased neck circumference	50 (64.1)	23 (67.6)	19 (70.4)	92 (66.2)	.82	139 (32.9)
Increased abdominal circumference	48 (87.3)	23 (95.8)	21 (100.0)	92 (92.0)	.22[b]	100 (51.7)
Modified Mallampati					.44	142 (31.4)
Class I	19 (23.2)	3 (10.0)	5 (16.7)	27 (19.0)		
Class II	29 (35.4)	15 (50.0)	9 (30.0)	53 (37.3)		
Class III	29 (35.4)	9 (30.0)	12 (40.0)	50 (35.2)		
Class IV	5 (6.1)	3 (10.0)	4 (13.3)	12 (8.5)		
Craniofacial and upper airway abnormalities	42 (84.0)	15 (83.3)	6 (66.7)	63 (81.8)	.49[b]	77 (62.8)
Daytime sleepiness	61 (55.5)	27 (60.0)	21 (50.0)	109 (55.3)	.64	197 (4.8)
Snoring	103 (92.8)	43 (93.5)	41 (93.2)	187 (93.0)	>.99[b]	201 (2.9)
Witnessed apneas	55 (58.5)	30 (76.9)	24 (66.7)	109 (64.5)	*.12*	169 (18.4)
Gasping and/or choking	39 (45.3)	12 (36.4)	16 (45.7)	67 (43.5)	.65	154 (25.6)
Sleep fragmentation	55 (73.3)	22 (68.8)	19 (73.1)	96 (72.2)	.88	133 (35.7)
Nonrepairing sleep	47 (47.5)	27 (62.8)	19 (46.3)	93 (50.8)	*.20*	183 (11.6)
Morning headaches	34 (46.6)	14 (48.3)	17 (53.1)	65 (48.5)	.83	134 (35.3)
Body position					.36[b]	201 (2.9)
Decubitus	5 (4.5)	0 (0.0)	1 (2.3)	6 (3.0)		
Left lateral	20 (18.2)	8 (17.0)	8 (18.2)	36 (17.9)		
Right lateral	56 (50.9)	22 (46.8)	16 (36.4)	94 (46.8)		
Supine	29 (26.4)	17 (36.2)	19 (43.2)	65 (32.3)		
Bad sleep efficiency	68 (61.3)	28 (56.0)	30 (65.2)	126 (60.9)	.65	207 (0.0)
Vehicle crashes	7 (20.6)	0 (0.0)	3 (20.0)	10 (16.4)	.28[b]	61 (70.5)
Truck driver	5 (4.7)	5 (10.4)	4 (9.5)	14 (7.1)	.32[b]	197 (4.8)
Driving sleepiness	5 (8.9)	4 (17.4)	4 (18.2)	13 (12.9)	.38[b]	101 (51.2)
Nocturia	47 (64.4)	20 (69.0)	32 (94.1)	99 (72.8)	*.005*	136 (34.3)
Alcohol consumption	61 (66.3)	29 (70.7)	29 (74.4)	119 (69.2)	.64	172 (16.9)
Smoking					.74	204 (1.4)
Yes	11 (10.0)	7 (14.6)	5 (10.9)	23 (10.9)		
Ex-smoker	38 (34.5)	20 (41.7)	17 (37.0)	75 (36.8)		

Characteristic	Mild (n=111), n (%)	Moderate (n=50), n (%)	Severe (n=46), n (%)	Total (N=207), n (%)	*P* value	Missing, n (%)
Coffee intake	77 (87.5)	30 (83.3)	25 (86.2)	132 (86.3)	.85[b]	153 (26.1)
Use of sedatives	23 (22.8)	13 (29.5)	7 (16.3)	43 (22.9)	.34	188 (9.2)
Family history	8 (18.2)	1 (5.9)	5 (31.2)	14 (18.2)	*.18* [b]	77 (62.8)
Epworth Sleepiness Scale	33 (37.5)	17 (44.7)	10 (29.4)	60 (37.5)	.41	160 (22.7)
Stroke	9 (37.5)	8 (80.0)	6 (60.0)	23 (52.3)	*.08* [b]	44 (78.7)
Myocardial infarction	9 (12.7)	3 (9.7)	6 (20.0)	18 (13.6)	.52[b]	132 (36.2)
Arterial hypertension	67 (79.8)	32 (91.4)	37 (92.5)	136 (85.5)	*.09*	159 (23.2)
Pulmonary hypertension	5 (8.3)	3 (10.0)	7 (25.9)	15 (12.8)	*.08* [b]	117 (43.5)
Congestive heart failure	7 (9.9)	6 (17.6)	13 (39.4)	26 (18.8)	*.002*	138 (33.3)
Arrhythmias	5 (10.0)	4 (16.7)	8 (32.0)	17 (17.2)	*.06* [b]	99 (52.2)
Pacemaker and/or cardioverter	3 (6.1)	2 (9.5)	5 (23.8)	10 (11.0)	*.09* [b]	91 (56.0)
Respiratory changes	43 (43.0)	15 (32.6)	23 (59.0)	81 (43.8)	*.05*	185 (10.6)
Diabetes	28 (51.9)	12 (60.0)	22 (88.0)	62 (62.6)	*.008*	99 (52.2)
Dyslipidemia	63 (78.8)	28 (90.3)	34 (91.9)	125 (84.5)	*.11*	148 (28.5)
Renal failure	10 (27.0)	6 (50.0)	7 (36.8)	23 (33.8)	.33	68 (67.1)
Hypothyroidism	12 (25.5)	6 (37.5)	6 (35.3)	24 (30.0)	.58	80 (61.4)
Gastroesophageal reflux	22 (48.9)	10 (71.4)	7 (53.8)	39 (54.2)	.34	72 (65.2)
Anxiety and/or depression	41 (78.8)	23 (92.0)	17 (77.3)	81 (81.8)	.31[b]	99 (52.2)
Insomnia	25 (71.4)	10 (76.9)	10 (90.9)	45 (76.3)	.48[b]	59 (71.5)

[a]*P*<.20 are italicized.

[b]Fisher exact test.

OSA Clusters

Using the 14 variables significantly associated with the outcome, a hierarchical clustering structure was derived, where, given the resulting clustering structure, a 10-cluster cutoff point was chosen (following the hierarchical structure of the clustering in the dendrogram). The resulting clusters had median AHI values of 8, 10 (4 clusters), 12, 13, 14, 31, and 34. As 10 clusters are difficult to interpret in a medical context, we chose to aggregate the 10 created clusters into 3 clusters according to their median values: (1) clusters with median 8 and 10, (2) clusters with median 12, 13, and 14, and (3) clusters with median 31 and 34.

The OSA cluster characteristics of the 14 predictive variables are described below and listed in Table 2. The witnessed apneas variable was also statistically significant in both the original and the curated data but was not considered for the cluster hierarchy, as it depends on third-party reporting, which might create a strong bias in the analysis.

Table 2. Clinical characteristics of the obstructive sleep apnea cohort by the defined clusters (*P* values are the results of chi-square tests unless otherwise specified).

Characteristics	Cluster 1 (n=74)		Cluster 2 (n=104)		Cluster 3 (n=29)		*P* value
	Patient, n (%)	95% CI	Patient, n (%)	95% CI	Patient, n (%)	95% CI	
Gender (male)	51 (68.9)	57-79	72 (69.2)	60-78	24 (82.8)	64-93	.32
Age (years)							<.001[a]
20-44	6 (8.1)	3-17	12 (11.5)	6-20	0 (0.0)	0-15	
45-64	46 (62.2)	50-73	59 (56.7)	47-66	7 (24.1)	11-44	
65-90	22 (29.7)	20-42	33 (31.7)	23-42	22 (75.9)	56-89	
BMI (kg/m^2)							<.001[a]
Normal weight	13 (17.6)	10-29	3 (2.9)	1-9	0 (0.0)	0-15	
Overweight	15 (20.3)	12-32	50 (48.1)	38-58	9 (31.0)	16-51	
Obesity	46 (62.2)	50-73	51 (49.0)	39-59	20 (69.0)	49-84	
Nonrepairing sleep	34 (45.9)	34-58	57 (54.8)	45-65	10 (34.5)	19-54	.13
Nocturia	14 (18.9)	11-30	104 (100.0)	96-100	29 (100.0)	85-100	<.001
Stroke	34 (45.9)	34-58	88 (84.6)	76-91	27 (93.1)	76-99	<.001
Arterial hypertension	54 (73.0)	61-82	96 (92.3)	85-96	29 (100.0)	85-100	<.001[a]
Pulmonary hypertension	9 (12.2)	6-22	2 (1.9)	0-8	6 (20.7)	9-40	.002[a]
Congestive heart failure	4 (5.4)	2-14	1 (1.0)	0-6	23 (79.3)	60-91	<.001[a]
Arrhythmias	4 (5.4)	2-14	1 (1.0)	0-6	12 (41.4)	24-61	<.001[a]
Pacemaker and/or cardioverter	0 (0.0)	0-6	4 (3.8)	1-10	6 (20.7)	9-40	<.001[a]
Respiratory changes	35 (47.3)	36-59	28 (26.9)	19-37	18 (62.1)	42-79	.001
Diabetes	37 (50.0)	39-61	69 (66.3)	56-75	29 (100.0)	85-100	<.001
Dyslipidemia	57 (77.0)	66-86	94 (90.4)	83-95	29 (100.0)	85-100	.003[a]
Apnea-hypopnea index							<.001
Mild	51 (68.9)	57-79	54 (51.9)	42-62	6 (20.7)	9-40	
Moderate	18 (24.3)	15-36	24 (23.1)	16-33	8 (27.6)	13-48	
Severe	5 (6.8)	3-16	26 (25.0)	17-35	15 (51.7)	33-70	

[a]Fisher exact test.

As shown in Table 2, 68.9% (51/74) of the patients in cluster 1 (74/207, 35.7%) were male, 62.2% (46/74) were aged between 45 and 64 years, and 62.2% (46/74) were obese. Nonrepairing sleep was reported in almost half of the patients, and only 18.9% (14/74) reported nocturia. The occurrence of stroke (34/74, 45.9%) did not reach half of the patients, whereas arterial hypertension (54/74, 73.0%) and dyslipidemia (57/74, 77.0%) surpassed it. Pulmonary hypertension, congestive heart failure, arrhythmias, and pacemaker or implantable cardioverter-defibrillator had percentages lower than 15%. The median AHI was 10 (range 7-17), the lowest AHI value, with 69.8% (44/169) reporting witnessed apneas.

Cluster 2 (104/207, 50.2%) had 69.2% (72/104) of males (the same as cluster 1), and only 2.9% (3/104) had normal weight. In contrast to cluster 1, 100.0% (104/104) of patients reported nocturia, 84.6% (88/104) reported stroke, 92.3% (96/104) reported arterial hypertension, and 90.4% (94/104) reported dyslipidemia. Similar to cluster 1, pulmonary hypertension, congestive heart failure, arrhythmias, and pacemaker or implantable cardioverter-defibrillator had percentages lower than 15%. Respiratory changes were reported in 26.9% (28/104) of the patients, and diabetes was reported in 66.3% (69/104) of the patients, compared with cluster 1. Regarding the clinical outcome, this cluster had a median AHI of 14 (range 8-30). Concerning witnessed apneas, cluster 2 had a percentage of 57.8% (52/169), the lowest value of all 3 clusters.

Cluster 3 (29/207, 14.0%) included the highest percentage of men (24/29, 82.8%). None of the patients were aged between 20 and 44 years or had normal weight. This cluster had the lowest proportion of patients aged between 45 and 64 years; nevertheless, it reached the highest proportion of all clusters in patients aged between 65 and 90 years. Although it had one of the lowest proportions of overweight patients, this cluster had the highest percentage (20/29, 69.0%) of patients with obesity. In contrast to cluster 1, but in concordance with cluster 2, nocturia was described in all patients in cluster 3. In addition,

arterial hypertension, diabetes, and dyslipidemia were observed in all the patients. The median AHI was 31 (range 21-60); therefore, it was the highest in all 3 clusters. Witnessed apneas were found with the highest proportion of all clusters (13/169, 81.2%).

Age strata and BMI were found to be different among clusters (P<.001). Comorbidities, such as stroke, arterial hypertension, diabetes (P<.001), and dyslipidemia (P=.003), were increasingly more prevalent from cluster 1 to clusters 2 and 3. Only male sex (P=.32) and nonrepairing sleep (P=.13) were not found to be significantly different.

On the basis of the description of clusters mentioned earlier, the OSA phenotypes can be defined. We classified patients into low (cluster 1), medium (cluster 2), and high (cluster 3) severity phenotypes, as their median AHI corresponded to mild, moderate, and severe levels respectively, defined in PSG for OSA diagnosis. The low severity phenotype includes age >45 years, a fair distribution in normal and overweight patients, accentuating obesity, and low prevalence of symptoms and comorbidities, except for dyslipidemia and arterial hypertension. The medium severity phenotype has almost the same distribution in age as the low severity phenotype, but less normal-weight patients and more overweight patients. Symptoms and comorbidities were higher, with stroke, arterial hypertension, dyslipidemia, and nocturia appearing in more than 85% of the patients with this phenotype. The high severity phenotype presents older and obese patients, with additional comorbidities (congestive heart failure and diabetes) beyond those present in the medium severity phenotype. The foremost difference between our phenotypes and AHI alone is that we considered the risk and diagnostic factors associated with the patient and not only a single value or a counting of events.

Affinity Between Healthy Patients and OSA Phenotypes

Given that our data set included patients who are healthy and with OSA (a total of 318 individuals), we focused our attention on exploring whether the determined OSA phenotypes could also help to segment healthy patients. To do so, we computed the aforementioned distance measure between 2 individuals using the same 14 significant variables. Table 3 describes the baseline characteristics of healthy patients for each OSA phenotype.

As expected, a high severity phenotype was less common in healthy patients (7/111, 6.3%), including older (P<.001), females (P=.49), and obese individuals (P=.50), with a lower proportion of individuals reporting nonrepairing sleep (P=.36). This phenotype also presented the highest proportion of reported nocturia, stroke, arterial hypertension, congestive heart failure, and diabetes (P<.001); pulmonary hypertension and arrhythmias (P=.01); and respiratory changes (P=.11). The medium severity phenotype had the highest proportion of overweight males aged between 45 and 64 years. Although comorbidities such as pulmonary hypertension, congestive heart failure, arrhythmias, and pacemaker or implantable cardioverter-defibrillator do not reach proportions higher than 1%, others such as stroke, arterial hypertension, diabetes, and dyslipidemia present proportions higher than 70%. The low severity phenotype is similar to the medium severity phenotype in terms of the proportion of overweight males, but individuals are younger. Nocturia, pulmonary hypertension, congestive heart failure, arrhythmias, pacemaker or implantable cardioverter-defibrillator, and diabetes have not been reported in this phenotype. Dyslipidemia was the most common comorbidity (16/25, 64%), followed by arterial hypertension (14/25, 56%) and respiratory changes (7/25, 28%).

Table 3. Clinical characteristics of the healthy cohort by the predefined obstructive sleep apnea phenotypes (*P* values are the result of chi-square test, unless otherwise specified).

Characteristics	Low OSA[a] (n=25)		Medium OSA (n=79)		High OSA (n=7)		*P* value[b]
	Patient, n (%)	95% CI	Patient, n (%)	95% CI	Patient, n (%)	95% CI	
Gender (male)	10 (40)	22-61	39 (49)	38-61	2 (29)	5-70	.49
Age (years)							<.001
20-44	16 (64)	43-81	15 (19)	11-30	0 (0)	0-44	
45-64	7 (28)	13-50	47 (59)	48-70	2 (29)	5-70	
65-90	2 (8)	1-28	17 (22)	13-32	5 (71)	30-95	
BMI							.50
Normal weight	4 (16)	5-37	6 (8)	3-16	1 (14)	1-58	
Overweight	11 (44)	25-65	40 (51)	39-62	2 (29)	5-70	
Obesity	10 (40)	22-61	33 (42)	31-53	4 (57)	20-88	
Nonrepairing sleep	18 (72)	50-87	51 (65)	53-75	3 (43)	12-80	.36
Nocturia	0 (0)	0-17	54 (68)	57-78	6 (86)	42-99	<.001
Stroke	1 (4)	0-22	66 (84)	73-91	6 (86)	42-99	<.001
Arterial hypertension	14 (56)	35-75	78 (99)	92-100	7 (100)	56-100	<.001
Pulmonary hypertension	0 (0)	0-17	1 (1)	0-8	2 (29)	5-70	.01
Congestive heart failure	0 (0)	0-17	0 (0)	0-6	7 (100)	56-100	<.001
Arrhythmias	0 (0)	0-17	1 (1)	0-8	2 (29)	5-70	.01
Pacemaker and/or cardioverter	0 (0)	0-17	1 (1)	0-8	0 (0)	0-44	>.99
Respiratory changes	7 (28)	13-50	34 (43)	32-55	5 (71)	30-95	.11
Diabetes	0 (0)	0-17	55 (70)	58-79	7 (100)	56-100	<.001
Dyslipidemia	16 (64)	43-81	75 (95)	87-98	6 (86)	42-99	.001

[a]OSA: obstructive sleep apnea.

[b]Fisher exact test.

Beyond OSA Phenotypes

OSA is a systemic disorder that remains underdiagnosed. Physicians, particularly nonspecialists in sleep disorders, urgently need a simple yet complete tool that allows them to identify a high pretest probability for OSA. This ability, which could enhance current screening, could lead to personalized treatment by additionally improving the understanding of OSA mechanisms and the risk for adverse events.

Our clinical prediction algorithm, that is, previously described OSA phenotypes, is a new way to screen patients, extending traditional approaches. To implement this new strategy, we need a simple, understandable, and updatable tool that can be used daily and that takes into account the knowledge of experts, the literature evidence, and the clinical data.

Belief or Bayesian networks [33] are probabilistic graphical models used to represent knowledge about an uncertain domain; each node represents a random variable, whereas directed edges between the nodes represent probabilistic dependencies among the corresponding variables. Bayesian networks are both mathematically rigorous and intuitively understandable, as they reflect a simple conditional independence statement, that is, each variable is independent of its nondescendants in the graph, given the state of its parents. The Bayesian network thus consists of both a qualitative model (which shows the relationship among variables) and a quantitative model (the joint probability distribution is expressed as conditional probabilities).

Initially, we created the simplest Bayesian classifier (naïve Bayes; Figure 1, Model A), which assumes independence among predictive variables and conditional independence, given the outcome. Subsequently, we extended the model (Figure 2, Model B), adding the defined phenotypes as a parent node of all predictors, thereby adjusting the model by capturing possible interactions among them, expressed by the corresponding phenotype associated with the tested individual. To evaluate the benefits of including OSA phenotypes in the clinical risk assessment tool, it was necessary to estimate the overall performance of each model. The ROC curves of each model (for both leave-one-out and cross-validation estimates) are presented in Figure 3, assessing the discriminative power of both models. As shown in Table 4, the derivation sample (area under the curve [AUC]) improved from 72% (95% CI 66-78) for model A to 84% (95% CI 80-89) for model B. The validity assessment confirmed the improvement achieved by the inclusion of OSA phenotypes, with leave-one-out estimates of 68% to 78%, respectively, from model A to model B and with

10 times twofold cross-validation averaging 67% and 77%, respectively. In addition, the diagnostic odds ratio, as a measure of the effectiveness of a diagnostic test, was 3.55 for model A and 2 times more for model B

Figure 1. Naïve Bayesian network representation of the relationships between the outcome (obstructive sleep apnea) and each of the 14 significant predictive variables. The bars within each variable represent the prior marginal probabilities for the category of each variable. CDI: implantable cardioverter-defibrillator; CHF: congestive heart failure; OSA: obstructive sleep apnea.

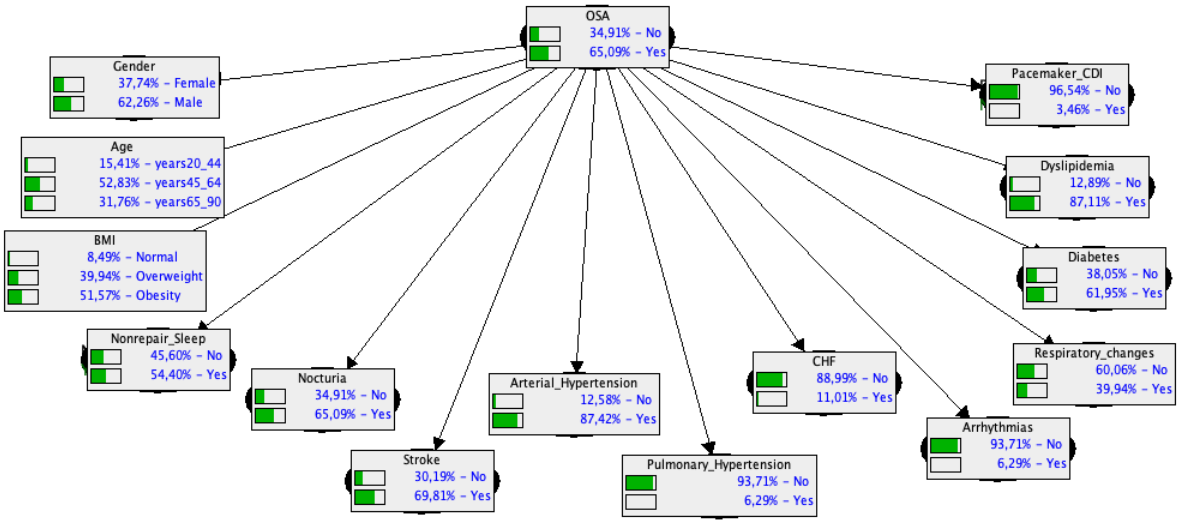


Figure 2. Naïve Bayesian network representation with additional node obtained from predefined obstructive sleep apnea phenotypes. The bars within each variable represent the prior marginal probabilities for the category of each variable. CDI: implantable cardioverter-defibrillator; CHF: congestive heart failure; OSA: obstructive sleep apnea.

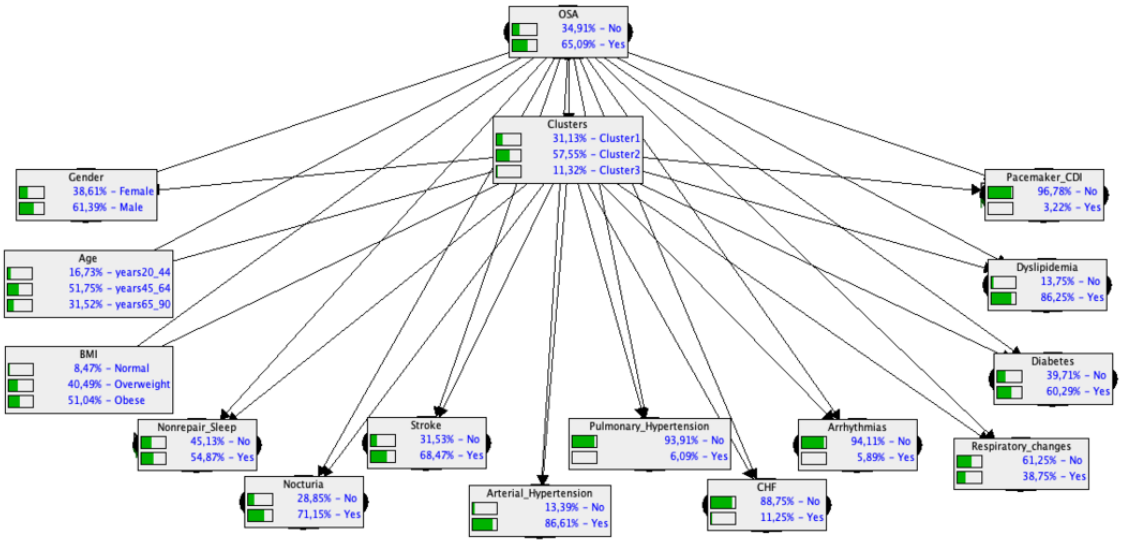


Figure 3. Receiver operating characteristic analyses and AUCs for models A (top) and B (bottom) as well for the internal validation procedures. AUC: area under the curve.

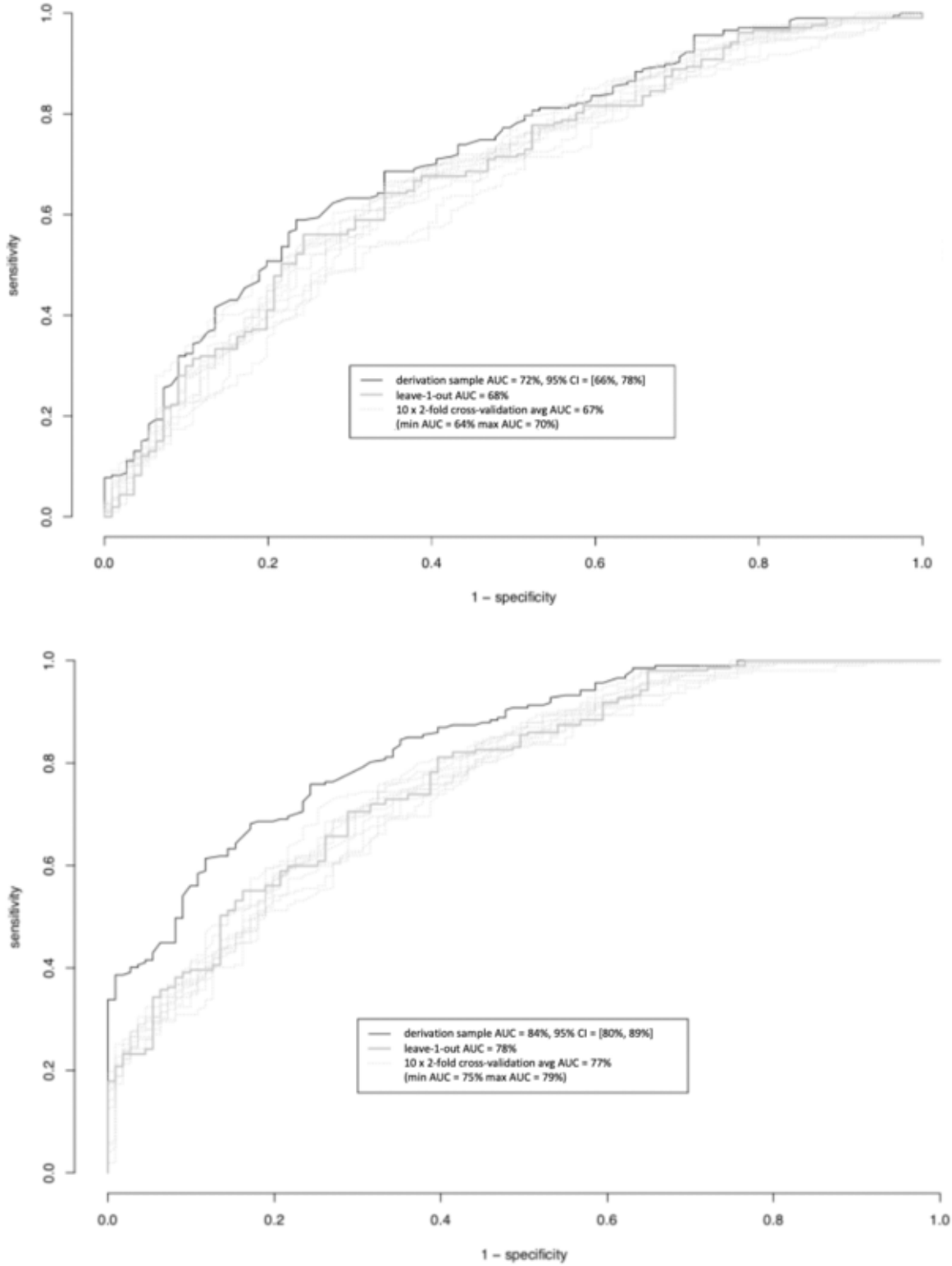


Table 4. Validity assessment estimated from 10 times twofold cross-validation.

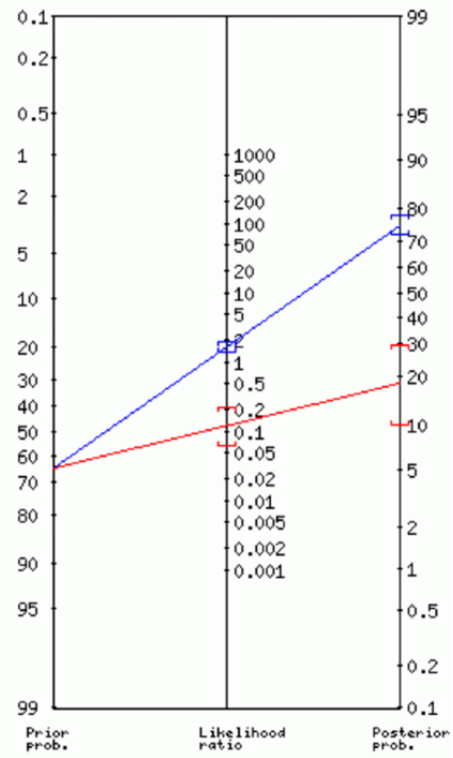
Variables	Obstructive sleep apnea	
	Model A	Model B
Cutoff point	30	22
Accuracy, % (95% CI)	69 (67-70)	74 (72-75)
Sensitivity, % (95% CI)	91 (89-94)	93 (91-95)
Specificity, % (95% CI)	26 (24-29)	38 (35-40)
Positive predictive value, % (95% CI)	70 (69-70)	73 (73-74)
Negative predictive value, % (95% CI)	64 (58-70)	75 (70-80)
Area under the curve, % (95% CI)	67 (67-70)	77 (76-78)
Positive likelihood ratio (95% CI)	1.32 (1.17-1.49)	1.63 (1.39-1.91)
Negative likelihood ratio (95% CI)	0.17 (0.09-0.34)	0.12 (0.06-0.22)
Positive odds posttest (95% CI)	2.45 (2.02-3.01)	3.02 (2.43-3.78)
Negative odds posttest (95% CI)	0.32 (0.19-0.56)	0.22 (0.12-0.38)
Posttest probability (95% CI)	71 (66-76)	75 (70-80)

Aiming at a 95% sensitivity target (screening strategies look for rule-out approaches), cutoff points were defined based on the derivation sample ROC curve, and the corresponding validity assessment results for cross-validation are displayed in Table 4, presenting an increase of specificity (26%-38%) for the desired level of sensitivity and presenting a posttest odds of 3 to 1 for the positive result and almost 1 to 5 for the negative result.

On the basis of the model with OSA phenotypes, OSA probabilities >22% were considered a positive result. The application of this cutoff resulted in a sensitivity value of 93% (95% CI 91-95) and 73% (95% CI 73-74) of positive predictive value, managing to provide a sensitive tool that prevents 1 out of 5 healthy individuals from unnecessarily undergoing PSG.

In our sample, the pretest probability was 65%, whereas the posttest probability increased to 75% using model B, with a posttest negative probability of 18%, as shown in Figure 4. These results highlight the value of using defined OSA phenotypes as predictors of OSA risk in referred individuals.

Figure 4. Fagan nomogram for model B. The blue and red lines represent the positive and negative posttest probability, respectively.



Discussion

Principal Findings

Understanding OSA patterns is important, particularly in the diagnosis of OSA. The AASM task force affirmed that the evaluation with clinical tools, such as clinical prediction algorithms, was less burdensome to the patient and physicians when compared with PSG. However, their low levels of accuracy and the likelihood of misdiagnosis must be weighted. Therefore, they proposed a clinical algorithm for the implementation of clinical practice guidelines for OSA. In the second step of this algorithm, the increased risk of moderate to severe OSA is measured by the presence of excessive daytime sleepiness and at least two of the following 3 criteria: habitual loud snoring, witnessed apnea or gasping or choking, or diagnosed hypertension. When we applied this moderate to severe risk in our data set (n=318), we found a sensitivity of 29%, a specificity of 68%, a positive predictive value of 50%, and a positive likelihood ratio of 0.875, showing possible benefits for a rule-out approach. However, considering the target of moderate to severe OSA identification, this approach revealed a very low level of sensitivity for a rule-in approach, which would be expected in this case.

To the best of our knowledge, this study is the first attempt to explore different clinical phenotypes of patients with OSA using categorical cluster analysis combined with Bayesian networks. We applied a hierarchical clustering procedure using Ward linkage on 14 significant predictive variables (out of the tested 47) that were grouped into 3 clusters: low, medium, and high severity phenotypes. These phenotypes were then used to expand a clinical prediction algorithm based on Bayesian networks, creating a simple but complete and updatable tool for OSA screening that can deal with missing information, based only on clinical and demographic variables, which have the main advantage of being easily available and quickly acquired by physicians.

Cluster analysis has been used in many medical conditions aiming to identify clinical phenotypes, as in the case of patients with asthma [16], where 5 clinical phenotypes illustrated the heterogeneity of the disease and relevant differences in treatment. Regarding OSA, clustering had been discussed as a possible helpful tool back in 1992, where the work of Tsuchiya et al [34] tried to apply cluster analysis in patients with OSA to overcome the stated overemphasis regarding obesity, which may have caused some physicians to overlook other potential factors that predispose this condition. They considered the apnea index (the standard at the time) and applied hierarchical clustering with average linkage, resulting in 2 clusters. The authors highlighted the controversy on the number of clusters, stating that “it should be essential to determine the number of clusters in a realistic way, and also to interpret the structures of clusters from a biologic standpoint.” Ye et al [35] collected demographic and survey data about sleep-related health issues (using numeric predictive variables) identifying 3 clusters: cluster 1 as *disturbed sleep group*, cluster 2 as *minimally symptomatic group*, and cluster 3 as *excessive daytime sleepiness group*. Although we have studied predictive variables related to daytime sleepiness, none were considered statistically significant; therefore, it is difficult to compare the results of the study by Ye et al [35] with the results of this study. Lacedonia et al [7] developed the work of Ye et al [35], enhancing the results using instrumental data, such as blood gas analysis and spirometry parameters (unavailable to us), to identify clinical presentations of patients with OSA. The authors used 2 approaches: a first one with hierarchical clustering revealing 3 clusters and the second one expanding it to 8 clusters with local optimization through principal component analysis.

Other studies are recently being developed, namely, the broad one in sleep apnea from the Sleep Apnea Network or European Sleep Apnea Database (ESADA) group. In 2016, Saaresranta et al [22] hypothesized that distinct OSA phenotypes should be present when discussing comorbidities and adherence to nasal continuous positive airway pressure (CPAP) therapy. This study has 3 main differences from ours: the ESADA database accepted PSG and cardiorespiratory polygraphy, whereas we only accepted PSG results; they accepted CPAP therapy and divided their patients into categories based only on subjective daytime sleepiness and nocturnal complaints. Regarding this last aspect, in our study, both subjective excessive daytime sleepiness and Epworth Sleepiness Scale were not considered in the cluster analysis. In 2020, a study by Bailly et al [21] applied latent class analysis to identify OSA phenotypes while reflecting geographical variations, resulting in 8 distinct clusters that were divided into 2 main categories: gender-based phenotypes (clusters 2 and 6 with only men and clusters 7 and 8 with only women) and men with various combinations (clusters 1, 3, 4, and 5), with which we can compare results. Cluster 3 of the study by Bailly et al [21] is described as obese comorbid patients, being the most similar to our low severity OSA cluster, presenting almost the same percentage of males (69% vs 73%) and higher levels of metabolic comorbidities.

Our results suggest 3 OSA phenotypes that can help in the screening, diagnosis, and later treatment of patients with OSA, capturing the full OSA spectrum of patients, focusing our attention on a detailed description of patients with OSA and not on a stereotypical one, where only a few *typical* symptoms such as snoring or daytime sleepiness are analyzed. To augment awareness of this prevalent disease, we even analyzed healthy patients to determine whether we could use the created phenotypes as identifiers of precursors of OSA.

Strengths and Limitations

This study had a modest number of patients, mainly because of the short period for data collection, which was performed in a small district hospital. Nevertheless, we believe that the procedure and the results are relevant. We also acknowledge that our phenotypes are not fully in accordance with the clinical phenotyping experience, particularly those regarding upper airway morphology. We suppose that the inclusion of other relevant outcome data could create a more robust analysis of the determined phenotypes. The inclusion of more patients and even dissociating variables, such as craniofacial upper airway abnormalities, could benefit future research.

The major strengths of this study are the study of a clinical cohort representing patients with OSA with all levels of severity

and the inclusion of a comprehensive number of risk and diagnostic factors that enhance our understanding of OSA diagnosis, with an overall cross-validated discriminative power of AUC of 77%, improving the specificity of a (designed) 95% sensitivity rule-out clinical prediction algorithm (3 to 1 odds for a positive result and 1 to 5 odds for a negative result). In addition, a diagnostic odds ratio higher than 1 was observed for models A and B, supporting the effectiveness of both models, with model B (inclusion of the disease phenotypes) doubling the diagnostic model performance. To assess the validity of our approach, we evaluated a logistic regression model in the derivation cohort, with and without predefined clusters, which highlighted the added discrimination value of using OSA phenotypes as a predictive variable (81% vs 83%). Moreover, we are aware that several clinical questionnaires (Berlin, STOP-BANG [snoring, tiredness, observed apnea, blood pressure, body mass index, age, neck circumference and gender], and NoSAS [neck, obesity, snoring, age, sex]) are helpful in identifying patients who are at risk of OSA. The Berlin questionnaire, when applied to the general population, reaches values of 37% for sensitivity and 84% for specificity, whereas when applied to primary care patients, the values are 86% and 77% [36], respectively. If we look at the STOP-BANG questionnaire, validation was performed in preoperative patients; the sensitivity and specificity values are 84% and 39%, respectively, for OSA diagnosis [37]. Finally, the NoSAS score was validated for the general population; the sensitivity values varied between 79% and 85%, the specificity varied between 69% and 77%, and AUC varied between 74% and 81% [38]. Comparing these results with our results, we can see that our sensitivity has the highest value, as we aim to establish a rule-out approach. On the other hand, our values for specificity and AUC were lower, only comparable with the value obtained for STOP-BANG.

Conclusions

We can affirm that using OSA phenotypes as predictors allows the creation of sensitive tools, with the defined phenotypes being a reflection of the early expression and the natural history of OSA. Nevertheless, OSA and individual responses are not static and evolve with time, creating the need for further studies on evaluating the phenotyping fluctuations and determining their long-term diagnosis implications.

Acknowledgments

DFS acknowledges Fundacão para a Ciência e Tecnologia under PhD grant (PD/BD/13553/2018) and the PhD Program in Clinical and Health Services Research (PD/00003/2013).

Authors' Contributions

DFS and PPR designed the study. DFS extracted the data. All authors screened the article, analyzed and interpreted the data, produced and revised all important intellectual content, gave their final approval of the version to be published, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Descriptive analysis of patients with obstructive sleep apnea after missing data imputation (absolute and relative frequencies are presented, and <italic>P</italic> values are the result of chi-square tests unless otherwise specified). Footnote a: <italic>P</italic><.20; these values are italicized. Footnote b: Fisher exact test. Footnote c: the odds ratio was calculated for moderate and severe levels combined because of the absence of patients with normal abdominal circumference at the severe level.
[PNG File , 240 KB - medinform_v9i6e25124_app1.png]

References

1. Kapur VK, Auckley DH, Chowdhuri S, Kuhlmann DC, Mehra R, Ramar K, et al. Clinical practice guideline for diagnostic testing for adult obstructive sleep apnea: an American Academy of Sleep Medicine Clinical Practice guideline. J Clin Sleep Med 2017 Mar 15;13(3):479-504 [FREE Full text] [doi: 10.5664/jcsm.6506] [Medline: 28162150]
2. Peppard PE, Young T, Barnet JH, Palta M, Hagen EW, Hla KM. Increased prevalence of sleep-disordered breathing in adults. Am J Epidemiol 2013 May 01;177(9):1006-1014 [FREE Full text] [doi: 10.1093/aje/kws342] [Medline: 23589584]
3. Ravesloot MJ, van Maanen JP, Hilgevoord AA, van Wagensveld BA, de Vries N. Obstructive sleep apnea is underrecognized and underdiagnosed in patients undergoing bariatric surgery. Eur Arch Otorhinolaryngol 2012 Jul 5;269(7):1865-1871 [FREE Full text] [doi: 10.1007/s00405-012-1948-0] [Medline: 22310840]
4. Johnson KG, Johnson DC. Frequency of sleep apnea in stroke and TIA patients: a meta-analysis. J Clin Sleep Med 2010 Apr 15;06(02):131-137. [doi: 10.5664/jcsm.27760]

5. Benjafield AV, Ayas NT, Eastwood PR, Heinzer R, Ip MS, Morrell MJ, et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. Lancet Respir Med 2019 Aug;7(8):687-698. [doi: 10.1016/s2213-2600(19)30198-5]
6. Kapur V, Blough D, Sandblom R, Hert R, de Maine JB, Sullivan SD, et al. The medical cost of undiagnosed sleep apnea. Sleep 1999 Sep 15;22(6):749-755. [doi: 10.1093/sleep/22.6.749] [Medline: 10505820]
7. Lacedonia D, Carpagnano GE, Sabato R, Storto MM, Palmiotti GA, Capozzi V, et al. Characterization of obstructive sleep apnea-hypopnea syndrome (OSA) population by means of cluster analysis. J Sleep Res 2016 May 18;25(6):724-730. [doi: 10.1111/jsr.12429] [Medline: 27191534]
8. Kushida CA, Littner M, Morgenthaler T, Alessi CA, Bailey D, Coleman J, et al. Practice parameters for the indications for polysomnography and related procedures: an update for 2005. Sleep 2005 Apr;28(4):499-521. [doi: 10.1093/sleep/28.4.499] [Medline: 16171294]
9. Campos-Rodriguez F, Martinez-Garcia MA, Cruz-Moron I, Almeida-Gonzales C, Catalan-Serra P, Montserrat JM. Cardiovascular mortality in women with obstructive sleep apnea with or without continuous positive airway pressure treatment. Ann Intern Med 2012 Jun 05;156(2):115-122. [doi: 10.7326/0003-4819-156-2-201201170-00006] [Medline: 22250142]
10. Shi Q, Rodrigues P. Monitoring the effectiveness of clinical guidelines: is the recommendation still valid? In: Proceedings of the IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS). 2018 Jun Presented at: IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS); June 18-21, 2018; Karlstad, Sweden p. 304-309 URL: https://www.computer.org/csdl/proceedings-article/cbms/2018/606001a304/12OmNxGAKWQ [doi: 10.1109/cbms.2018.00060]
11. Kent BD, Grote L, Ryan S, Pépin J, Bonsignore MR, Tkacova R, et al. Diabetes mellitus prevalence and control in sleep-disordered breathing. Chest 2014 Oct;146(4):982-990. [doi: 10.1378/chest.13-2403] [Medline: 24831859]
12. Kakkar RK, Berry RB. Positive airway pressure treatment for obstructive sleep apnea. Chest 2007 Sep;132(3):1057-1072. [doi: 10.1378/chest.06-2432] [Medline: 17873201]
13. Nickerson J, Lee E, Nedelman M, Aurora RN, Krieger A, Horowitz CR. Feasibility of portable sleep monitors to detect obstructive sleep apnea (OSA) in a vulnerable urban population. J Am Board Fam Med 2015 Mar 06;28(2):257-264 [FREE Full text] [doi: 10.3122/jabfm.2015.02.140273] [Medline: 25748767]
14. Ferreira-Santos D, Monteiro-Soares M, Rodrigues PP. Impact of imputing missing data in Bayesian network structure learning for obstructive sleep apnea diagnosis. Stud Health Technol Inform 2018;247:126-130 [FREE Full text] [doi: 10.3233/978-1-61499-852-5-126] [Medline: 29677936]
15. Gallo C, Capozzi V. Clustering techniques for revealing gene expression patterns. In: Encyclopedia of Information Science and Technology, Third Edition. Hershey, PA: IGI Global; 2015:438-447.
16. Moore WC, Meyers D, Wenzel S, Teague WG, Li H, Li X, National Heart, Lung,Blood Institute's Severe Asthma Research Program. Identification of asthma phenotypes using cluster analysis in the Severe Asthma Research Program. Am J Respir Crit Care Med 2010 Feb 15;181(4):315-323 [FREE Full text] [doi: 10.1164/rccm.200906-0896OC] [Medline: 19892860]
17. Garcia-Aymerich J, Gomez FP, Benet M, Farrero E, Basagana X, Gayete A, PAC-COPD Study Group. Identification and prospective validation of clinically relevant chronic obstructive pulmonary disease (COPD) subtypes. Thorax 2011 May;66(5):430-437. [doi: 10.1136/thx.2010.154484] [Medline: 21177668]
18. Docampo E, Collado A, Escaramís G, Carbonell J, Rivera J, Vidal J, et al. Cluster analysis of clinical data identifies fibromyalgia subgroups. PLoS One 2013 Sep 30;8(9):e74873 [FREE Full text] [doi: 10.1371/journal.pone.0074873] [Medline: 24098674]
19. Erro R, Vitale C, Amboni M, Picillo M, Moccia M, Longo K, et al. The heterogeneity of early Parkinson's disease: a cluster analysis on newly diagnosed untreated patients. PLoS One 2013 Aug 1;8(8):e70244 [FREE Full text] [doi: 10.1371/journal.pone.0070244] [Medline: 23936396]
20. Topîrceanu A, Udrescu L, Udrescu M, Mihaicuta S. Gender phenotyping of patients with obstructive sleep apnea syndrome using a network science approach. J Clin Med 2020 Dec 12;9(12):4025 [FREE Full text] [doi: 10.3390/jcm9124025] [Medline: 33322816]
21. Bailly S, Grote L, Hedner J, Schiza S, McNicholas WT, Basoglu OK, ESADA Study Group. Clusters of sleep apnoea phenotypes: a large pan‐European study from the European Sleep Apnoea Database (ESADA). Respirology 2021 Apr;26(4):378-387. [doi: 10.1111/resp.13969] [Medline: 33140467]
22. Saaresranta T, Hedner J, Bonsignore MR, Riha RL, McNicholas WT, Penzel T, ESADA Study Group. Clinical phenotypes and comorbidity in European sleep apnoea patients. PLoS One 2016 Oct 4;11(10):e0163439 [FREE Full text] [doi: 10.1371/journal.pone.0163439] [Medline: 27701416]
23. Core R Team. R: a language and environment for statistical computing. In: R Foundation for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2017.
24. Weisberg S, Fox J. An R Companion to Applied Regression. Thousand Oaks, California: SAGE Publications; 2011:1-472.
25. Warnes GR, Bolker B, Lumley T. gmodels: various R programming tools for model fitting. In: R Foundation for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2018.

26. Aragon TJ. epitools: epidemiology tools. In: R Foundation for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2017.
27. Comtois D. summarytools: tools to quickly and neatly summarize data. In: R Foundation for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2019.
28. Torgo L. Data mining with R - learning with case studies. Minneapolis, USA: Chapman and Hall/CRC; Jun 20, 2020.
29. Scutari M. Learning Bayesian networks with the package. J Stat Soft 2010;35(3):1-22. [doi: 10.18637/jss.v035.i03]
30. Højsgaard S. Graphical independence networks with the grain package for R. J Stat Soft 2012 Feb 28;46(10):12031 [FREE Full text] [doi: 10.18637/jss.v046.i10]
31. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011 Mar 17;12:77 [FREE Full text] [doi: 10.1186/1471-2105-12-77] [Medline: 21414208]
32. Darwiche A. Modeling and Reasoning with Bayesian Networks. California, Los Angeles: Cambridge University Press; 2009.
33. Darwiche A. Bayesian networks. Commun ACM 2010 Dec;53(12):80-90 [FREE Full text] [doi: 10.1145/1859204.1859227]
34. Tsuchiya M, Lowe AA, Pae E, Fleetham JA. Obstructive sleep apnea subtypes by cluster analysis. Am J Orthod Dentofac Orthop 1992 Jun;101(6):533-542. [doi: 10.1016/0889-5406(92)70128-w] [Medline: 1598893]
35. Ye L, Pien GW, Ratcliffe SJ, Björnsdottir E, Arnardottir ES, Pack AI, et al. The different clinical faces of obstructive sleep apnoea: a cluster analysis. Eur Respir J 2014 Sep 03;44(6):1600-1607. [doi: 10.1183/09031936.00032314] [Medline: 25186268]
36. Netzer NC, Stoohs RA, Netzer CM, Clark K, Strohl KP. Using the Berlin Questionnaire to identify patients at risk for the sleep apnea syndrome. Ann Intern Med 1999 Oct 05;131(7):485-491. [doi: 10.7326/0003-4819-131-7-199910050-00002] [Medline: 10507956]
37. Chung F, Yang Y, Brown R, Liao P. Alternative scoring models of STOP-bang questionnaire improve specificity to detect undiagnosed obstructive sleep apnea. J Clin Sleep Med 2014 Sep 15;10(9):951-958 [FREE Full text] [doi: 10.5664/jcsm.4022] [Medline: 25142767]
38. Marti-Soler H, Hirotsu C, Marques-Vidal P, Vollenweider P, Waeber G, Preisig M, et al. The NoSAS score for screening of sleep-disordered breathing: a derivation and validation study. Lancet Respir Med 2016 Sep;4(9):742-748. [doi: 10.1016/s2213-2600(16)30075-3] [Medline: 27321086]

Abbreviations

AASM: American Academy of Sleep Medicine
AHI: apnea-hypopnea index
AUC: area under the curve
CPAP: continuous positive airway pressure
ESADA: European Sleep Apnea Database
NN: nearest neighbor
NoSAS: neck, obesity, snoring, age, sex
OSA: obstructive sleep apnea
PSG: polysomnography
ROC: receiver operating characteristic
STOP-BANG: snoring, tiredness, observed apnea, blood pressure, body mass index, age, neck circumference and gender

Edited by R Kukafka; submitted 20.10.20; peer-reviewed by T Penzel, B Sébastien; comments to author 23.12.20; revised version received 22.01.21; accepted 16.03.21; published 22.06.21.

Please cite as:
Ferreira-Santos D, Rodrigues PP
Enhancing Obstructive Sleep Apnea Diagnosis With Screening Through Disease Phenotypes: Algorithm Development and Validation
JMIR Med Inform 2021;9(6):e25124
URL: https://medinform.jmir.org/2021/6/e25124
doi:10.2196/25124
PMID:34156340

Original Paper

Physicians' Perspectives of Telemedicine During the COVID-19 Pandemic in China: Qualitative Survey Study

Jialin Liu[1,2*], MD; Siru Liu[3*], PhD; Tao Zheng[1], BS; Yongdong Bi[1], BS

[1]Department of Medical Informatics, West China Hospital, Sichuan University, Chengdu, China

[2]Department of Otolaryngology, West China Hospital, Sichuan University, Chengdu, China

[3]Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, United States

[*]these authors contributed equally

Corresponding Author:
Jialin Liu, MD
Department of Medical Informatics
West China Hospital
Sichuan University
No.37 Guoxuexiang street
Chengdu, 610041
China
Phone: 86 2885422306
Email: dljl8@163.com

Abstract

Background: Generalized restriction of movement due to the COVID-19 pandemic, together with unprecedented pressure on the health system, has disrupted routine care for non–COVID-19 patients. Telemedicine should be vigorously promoted to reduce the risk of infections and to offer medical assistance to restricted patients.

Objective: The purpose of this study was to understand physicians' attitudes toward and perspectives of telemedicine during and after the COVID-19 pandemic, in order to provide support for better implementation of telemedicine.

Methods: We surveyed all physicians (N=148), from October 17 to 25, 2020, who attended the clinical informatics PhD program at West China Medical School, Sichuan University, China. The physicians came from 57 hospitals in 16 provinces (ie, municipalities) across China, 54 of which are 3A-level hospitals, two are 3B-level hospitals, and one is a 2A-level hospital.

Results: Among 148 physicians, a survey response rate of 87.2% (129/148) was attained. The average age of the respondents was 35.6 (SD 3.9) years (range 23-48 years) and 67 out of 129 respondents (51.9%) were female. The respondents come from 37 clinical specialties in 55 hospitals located in 14 provinces (ie, municipalities) across Eastern, Central, and Western China. A total of 94.6% (122/129) of respondents' hospitals had adopted a telemedicine system; however, 34.1% (44/129) of the physicians had never used a telemedicine system and only 9.3% (12/129) used one frequently (≥1 time/week). A total of 91.5% (118/129) and 88.4% (114/129) of physicians were willing to use telemedicine during and after the COVID-19 pandemic, respectively. Physicians considered the inability to examine patients in person to be the biggest concern (101/129, 78.3%) and the biggest barrier (76/129, 58.9%) to implementing telemedicine.

Conclusions: Telemedicine is not yet universally available for all health care needs and has not been used frequently by physicians in this study. However, the willingness of physicians to use telemedicine was high. Telemedicine still has many problems to overcome.

(JMIR Med Inform 2021;9(6):e26463) doi:10.2196/26463

KEYWORDS

telemedicine; COVID-19; survey; physician

Introduction

The COVID-19 pandemic has drastically impacted global health care and dramatically changed the practice of health care [1,2]. Pervasive movement restriction and the unprecedented pressure on the health system has disrupted routine care for non–COVID-19 patients. Therefore, the COVID-19 pandemic has rapidly and fundamentally altered the pattern medical

practitioners follow to provide care to patients. To better mitigate and manage the spread of COVID-19, hospitals can replace some routine medical services with telemedicine to improve the efficiency of their health care system [3].

Since telemedicine was first introduced in the late 1950s, it has been used in all aspects of health care with the widespread use of telecommunication technology [4]. In a bibliometric analysis of health technology and informatics, telemedicine was identified as one of the three most common keywords [5]. Now the application of telemedicine has expanded from providing health care services in hospitals, outpatient departments, and specialist offices, as well as between health care providers, to deliver care in patients' homes [6]. One study has shown that achieving instant patient access, overcoming service gaps, and improving quality are important motivators for physicians to implement telemedicine in acute care units, while issues such as licensure, credentialing, malpractice protection, cost, and reimbursement are barriers to successful implementation [7]. Another study identified that the main challenges in establishing telemedicine systems in developing countries are the high cost of telemedicine systems and solutions, slow clinical acceptance of telemedicine and resistance to change, and lack of the required information and telecommunications technology infrastructure for telemedicine. The major recommendations include setting clear goals for the project, selecting the appropriate application of medical areas and priorities, and adopting user-friendly interfaces [8].

Our study focused on the context of COVID-19 to investigate the current usage of telemedicine during the pandemic in China. With the development of telemedicine, the evaluation of telemedicine is particularly important [9]. The selection of statistical methods is a key step in telemedicine evaluation. The following statistical methods have been used extensively in telemedicine evaluation: statistical comparison, agreement evaluation (κ statistic), and the receiver operating characteristic curve [10-14]. Since telemedicine evaluation needs to explore various outcomes, it may be appropriate to evaluate from a multidisciplinary perspective and use various statistical methods [10]. However, there is a lack of empirical research about telemedicine in different specialties [15]. Some researchers have provided theoretical and practical evidence on the significance of using telemedicine and virtual care to treat patients remotely during the COVID-19 pandemic [16]. Major health organizations around the world, including the World Health Organization, the US Centers for Disease Control and Prevention, and the American Medical Association, have advocated for the use of telemedicine during the COVID-19 pandemic and have taken steps to promote its use [17-19]. During the COVID-19 pandemic, telemedicine has been considered a useful tool to relieve pressure on overburdened health systems. Physicians' willingness or unwillingness to use telemedicine is a well-known factor in facilitating or inhibiting telemedicine acceptance [20]. In addition, some studies noted that the adoption of telemedicine systems depends on physicians' and patients' satisfaction with the use of the telemedicine service [21]. However, physicians' perspectives on telemedicine visits have not been fully investigated.

To promote the usage of telemedicine during the COVID-19 pandemic, the current state of telemedicine and physicians' perspectives need to be explored. To better understand the development of telemedicine during the COVID-19 pandemic and to summarize the problems of telemedicine in response to the pandemic, we collected the opinions and suggestions of 148 young and middle-aged physicians regarding the application of telemedicine during the COVID-19 pandemic. These recommendations provide valuable insights for developing and improving telemedicine in the later stages of the COVID-19 pandemic and play an important role in guiding the development of telemedicine.

Methods

Participants

We surveyed all physicians (N=148), from October 17 to 25, 2020, who attended the clinical informatics PhD program at West China Medical School, Sichuan University, China. These physicians passed the program's application and examination process and the hospital academic committee's review. They had high levels of informatics literacy and a certain understanding of information technology and telemedicine at their hospitals. The physicians came from 57 hospitals in 16 provinces (ie, municipalities) across China, 54 of which are 3A-level hospitals, two are 3B-level hospitals, and one is a 2A-level hospital. The Ministry of Health in China categorizes Chinese hospitals into three levels—primary, secondary, and tertiary hospitals—based on the quality of the health care provided, medical education, and research. Each level is further subdivided into three subsidiary levels: A, B, and C. In 2019, there were 1246 hospitals at the 3A level [22], the highest level of hospitals in China.

This study was approved by the Institutional Review Board at West China Medical School, Sichuan University (IRB17-75).

Procedure

We conducted a survey using semistructured and open-ended questions to understand physicians' perspectives of telemedicine during the COVID-19 pandemic in China. Prior to completing the survey, the physicians spent more than 3 hours on coursework related to telemedicine. The questionnaire was derived from the literature on telemedicine satisfaction and experts in telemedicine [23-28]. We conducted a pilot test within our research group. The questionnaire consisted of three sections (Multimedia Appendix 1). The first part included demographic and clinical characteristics (age, gender, clinical specialty, etc). The second part consisted of statements that were rated on a 7-point Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree). Statements were identified from previous literature that related to physicians' perspectives on and attitudes toward telemedicine, such as overall satisfaction, behavioral intention, increasing the burden, safety issues regarding patient data, and hindering communication with patients, among others. In addition, we collected information about the current usage of telemedicine in their hospitals. The final section consisted of open-ended questions that included physician attitudes, concerns, and suggestions about telemedicine and any other

comments related to telemedicine. The questionnaire was administered in a face-to-face manner.

Data Gathering and Analysis

After completing the questionnaire, the data were tabulated and analyzed. All the physicians' responses to the open-ended questions were entered into Microsoft Office Excel 2007 and were subjected to qualitative content analysis by reviewers. The analytical process was conducted by first cleaning the text, followed by extracting themes, and then developing categories. Free-text answers were summarized and assessed independently by two reviewers using a standardized evaluation process. A third reviewer reviewed by adjudication in cases of disagreement. The research team members repeatedly and independently read the answer summaries and validated the accuracy and meaning of the contents. Lastly, the results of the study were confirmed by all researchers in the team. The responses to the Likert scale–based statements were analyzed quantitatively by expressing them as whole numbers. The percentage of respondents who were in agreement with a statement was obtained by dividing the sum of the *strongly agree*, *agree*, and *somewhat agree* responses by the total number of responses to that statement. For questions using a 7-point Likert scale and questions that collected numerical demographic information, we reported mean values with standard deviations. For each clinical specialty, we calculated *P* values to determine the statistical significance of the differences between the scores of usability and willingness. Two-sided *P* values of .01 or less were deemed to meet statistical significance.

Results

Physician Demographics and Characteristics

We received 129 completed survey forms—direct survey handout and return on the day—with a response rate of 87.2% (129/148). Out of 129 respondents, 67 (51.9%) were females and 62 (48.1%) were males. The average age of the respondents was 35.6 (SD 3.9) years (range 23-48 years). The respondents came from 37 clinical specialties in 55 hospitals in China. These hospitals were located in 14 provinces (ie, municipalities) across China, including the three main provincial regions: Western China (n=5), Central China (n=4), and Eastern China (n=5). Among these 55 hospitals, 52 were 3A-level hospitals (ie, the highest level of hospital in China), two were 3B-level hospitals, and one was a 2A-level hospital. Table 1 shows the demographic characteristics of the respondents.

All hospitals in China are divided into three grades, each with three sublevels (ie, A, B, and C), with the highest grade being 3A. In principle, hospitals rated as a 3A-level hospital must meet very high standards in terms of beds, doctors, equipment, and quality of service.

Table 1. Demographic and clinical practice characteristics.

Participant demographics	Value (N=129)
Age (years)	
Mean (SD)	35.6 (3.9)
Range, n (%)	
23-29	4 (3.1)
30-39	105 (81.4)
40-48	20 (15.5)
Sex, n (%)	
Female	67 (51.9)
Male	62 (48.1)
Title, n (%)	
Resident	6 (4.7)
Senior physician	89 (69.0)
Specialist	34 (26.4)
Experience on the job (years)	
Mean (SD)	9.5 (4.5)
Range, n (%)	
1-5	27 (20.9)
6-10	57 (44.2)
11-20	42 (32.6)
21-25	3 (2.3)
Electronic health record use (years)	
Mean (SD)	8.0 (2.8)
Range, n (%)	
0-5	25 (19.3)
6-10	82 (63.6)
11-16	22 (17.1)
Provinces where hospitals were located per region (n=14), n (%)	
Western China[a]	5 (35.7)
Central China[b]	4 (28.6)
Eastern China[c]	5 (35.7)
Hospital level, n (%)	
3A	52 (94.6)
2A	2 (3.6)
3B	1 (1.8)

[a]This includes Sichuan, Chongqing, Guangxi, Xinjiang, and Yunnan.

[b]This includes Shanxi, Henan, Hunan, and Jiangxi.

[c]This includes Beijing, Fujian, Guangdong, Shandong, and Liaoning.

Current Use of Telemedicine

Among the 129 respondents, 94.6% (122/129) of the respondents' hospitals adopted a telemedicine system. Only 5.4% (7/129) of the respondents did not know whether telemedicine was used in the hospital. A total of 34.1% (44/129) of physicians had never used a telemedicine system, 45.0% (58/129) used one occasionally (≤1 time/month), 11.6% (15/129) used one often (>1 time/month – <1 time/week), and only 9.3% (12/129) used one frequently (≥1 time/week). Depending on the question asked, 52% (44/85) of respondents were satisfied

(responses of *strongly satisfied* plus *satisfied* and *somewhat satisfied*) with the telemedicine system (mean 4.7, SD 0.82).

Only 57 out of 129 (44.2%) physicians had participated in telemedicine training. A total of 32% (18/57) of those respondents were satisfied with their training (mean 4.2, SD 0.64). Among physicians who had used telemedicine systems, 11% (9/85) of them believed that electronic medical records were integrated into telemedicine. A total of 32% (27/85) of physicians believed that telemedicine had a decision support system (Table 2).

Table 2. Current use of telemedicine system.

Question or statement	Value (N=129)
Has your hospital adopted a telemedicine system? (yes), n (%)	122 (94.6)
How often do you use the telemedicine system?, n (%)	
Not at all	44 (34.1)
≤1 time/month	58 (45.0)
>1 time/month – <1 time/week	15 (11.6)
≥1 time/week	12 (9.3)
What is your overall satisfaction with the telemedicine system?[a] (n=85)	
Satisfied, n (%)	44 (51.8)
Score, mean (SD)	4.7 (0.82)
Score, range	3-7
Have you taken telemedicine training? (yes), n (%)	57 (44.2)
What is your overall satisfaction with the telemedicine training?[a] (n=57)	
Satisfied, n (%)	18 (31.6)
Score, mean (SD)	4.2 (0.64)
Score, range	2-5
Does the telemedicine system integrate electronic medical records?[b]	
Yes, n (%)	9 (7.0)
Score, mean (SD)	4.2 (0.42)
Score, range	4-5
Does the telemedicine system integrate clinical decision support?[b]	
Yes, n (%)	27 (20.9)
Score, mean (SD)	4.3 (0.67)
Score, range	3-5

[a]Satisfaction scores range from 1 (strongly dissatisfied) to 7 (strongly satisfied).

[b]Agreement scores range from 1 (strongly disagree) to 7 (strongly agree).

Telemedicine During COVID-19

Of the 129 respondents, 60.5% (78/129) indicated that their specialty was suitable (responses of *strongly suitable* plus *suitable* and *somewhat suitable*) for adopting telemedicine during the COVID-19 pandemic (mean 5.0, SD 1.28). A total of 91.5% (118/129) of respondents would be willing to adopt telemedicine during the COVID-19 pandemic (mean 5.7, SD 1.02). In the group with telemedicine-appropriate specialties, obstetrics and gynecology had the highest mean value (mean 6.3, SD 0.97) and dermatology had the lowest mean value (mean 4.2, SD 0.75). Regarding willingness to adopt telemedicine, radiologists had the highest mean value (mean 6.4, SD 0.80) and ophthalmologists had the lowest mean value (mean 4.6, SD 0.49). For each specialty, we calculated *P* values to determine the statistical significance of the differences between the scores of usability and willingness ($P>.01$). The detailed attitudes and opinions about telemedicine on the part of the physicians are shown in Table 3.

Table 3. Physicians' attitudes and opinions on the use of telemedicine in different subspecialties.

Specialty	Is telemedicine suitable for your specialty during the COVID-19 pandemic?[a]			Are you willing to use a telemedicine system during the COVID-19 pandemic?[b]			*P* value
	Score, range	Score, mean (SD)	Suitable (yes), n (%)	Score, range	Score, mean (SD)	Willing (yes), n (%)	
All (N=129)	2-7	5.0 (1.28)	78 (60.5)	3-7	5.7 (1.02)	118 (91.5)	N/A[c]
Dermatology (n=5)	3-5	4.2 (0.75)	__[d]	5-7	6.2 (0.98)	—	.012
Urology (n=6)	2-6	4.2 (1.21)	—	5-7	5.8 (0.90)	—	.03
Laboratory (n=5)	3-7	4.2 (1.47)	—	4-7	6.0 (1.27)	—	.10
Neurosurgery (n=6)	3-7	4.3 (1.25)	—	4-7	5.3 (0.94)	—	.18
Nephrology (n=5)	4-5	4.4 (0.49)	—	5-7	6.0 (0.89)	—	.013
General surgery (n=9)	2-7	4.6 (1.34)	—	4-7	5.4 (1.07)	—	.16
Ophthalmology (n=5)	4-7	4.8 (0.75)	—	4-5	4.6 (0.49)	—	.67
Pediatrics (n=9)	4-6	5.0 (0.67)	—	4-7	5.9 (1.10)	—	.07
Anesthesiology (n=12)	2-7	5.1 (1.38)	—	5-7	6.0 (0.91)	—	.08
Oncology (n=8)	4-7	5.3 (1.09)	—	5-7	5.6 (0.86)	—	.49
Respiratory (n=6)	4-7	5.3 (0.94)	—	4-7	5.8 (1.07)	—	.45
Cardiothoracic surgery (n=7)	4-7	5.5 (1.28)	—	5-7	6.1 (0.83)	—	.50
Orthopedics (n=8)	3-7	5.8 (1.30)	—	5-7	6.0 (1.00)	—	.69
Radiology (n=5)	5-7	6.0 (0.89)	—	5-7	6.4 (0.80)	—	.52
Obstetrics and gynecology (n=8)	4-7	6.3 (0.97)	—	4-7	5.9 (1.05)	—	.50

[a]This includes *strongly suitable* plus *somewhat suitable* and *suitable*. Suitability scores range from 1 (strongly unsuitable) to 7 (strongly suitable).

[b]This includes *strongly willing* plus *willing* and *somewhat willing*. Willingness scores range from 1 (strongly unwilling) to 7 (strongly willing).

[c]N/A: not applicable; *P* values were only calculated for individual specialties.

[d]The number of respondents who found telemedicine to be suitable and were willing to use it was not reported for individual specialties.

Main Concerns of Adopting Telemedicine

Based on the findings of the survey, the major concerns regarding the use of telemedicine included the following: the inability to complete an in-person physical examination (101/129, 78.3%), the inability to communicate well with patients (32/129, 24.8%), the instability of the telemedicine system (30/129, 23.3%), and no assurance of patient medical safety (23/129, 17.8%) (Table 4).

Table 4. Major concerns regarding the use of telemedicine.

Major concerns	Respondents (N=129), n (%)
Cannot communicate well with patients	32 (24.8)
No assurance of patient medical safety	23 (17.8)
Inability to do an in-person physical examination	101 (78.3)
Unstable telemedicine system	30 (23.3)

Barriers to the Use of Telemedicine

Overall, 58.9% (76/129) of respondents agreed that a physician's inability to examine patients will hinder clinical decision making. A total of 44.2% (57/129) of respondents agreed that telemedicine makes it easier for patients' data to be stolen, compromised, or hacked. Approximately one-quarter of the respondents (32/129, 24.8%) agreed that the lack of person-to-person contact in telemedicine can damage the doctor-patient relationship and trust. Only 15.5% (20/129) of respondents agreed that during the COVID-19 pandemic, the use of telemedicine will increase the burden on physicians (Table 5).

Table 5. Barriers to adopting telemedicine.

Barrier	Score, range	Score, mean (SD)	Respondents who agree[a] (N=129), n (%)	Respondents who disagree[a] (N=129), n (%)
The lack of person-to-person contact in telemedicine can damage the doctor-patient relationship and trust.	1-7	3.6 (1.89)	32 (24.8)	62 (48.1)
A physician's inability to examine patients will hinder clinical decision making.	1-7	4.5 (1.02)	76 (58.9)	23 (17.8)
During the COVID-19 pandemic, the use of telemedicine will increase the burden on physicians.	1-6	3.0 (1.20)	20 (15.5)	87 (67.4)
Telemedicine makes it easier for patient data to be stolen, compromised, or hacked.	1-7	4.1 (1.23)	57 (44.2)	42 (32.6)

[a]Agreement includes *strongly agree* plus *somewhat agree* and *agree*. Disagreement includes *strongly disagree* plus *somewhat disagree* and *disagree.* Scores range from 1 (strongly disagree) to 7 (strongly agree).

Physicians' Comments

In the open-ended section of the questionnaire, a total of 127 respondents out of 129 (98.4%) made comments regarding the obstacles to adopting telemedicine and made suggestions for improving telemedicine (Tables 6 and 7). Two respondents did not make comments or suggestions about telemedicine.

The main barriers to implementation cited by physicians included the inability to examine patients personally (48/127, 37.8%), insufficient infrastructure support for telemedicine (40/127, 31.5%), issues concerning the quality of patients' data (28/127, 22.1%), communication issues with patients (18/127, 14.2%), network issues (13/127, 10.2%), and lack of policy support (10/127, 7.9%). Table 6 lists the physicians' comments regarding obstacles to the use of telemedicine.

Physicians believed that telemedicine could be promoted through the following incentives: performance measures (60/127, 47.2%), increased telemedicine equipment (22/127, 17.3%), policy support (21/127, 16.5%), financial support (19/127, 15.0%), technical support (18/127, 14.2%), increased training (18/127, 14.2%), and increased telemedicine publicity (14/127, 11.0%) (Table 7).

Table 6. Physicians' comments regarding obstacles to the use of telemedicine.

Main obstacles to adoption of telemedicine[a]	Respondents (n=127), n (%)
Inability to examine patients personally	48 (37.8)
Insufficient infrastructure support for telemedicine	40 (31.5)
Issues concerning the quality of patients' data	28 (22.1)
Communicating issues with patients	18 (14.2)
Network issues	13 (10.2)
Lack of policy support	10 (7.9)
Others[b]	49 (38.6)

[a]There were a total of 206 comments.

[b]Other comments included low patient acceptance (n=5), lack of funds (n=4), lack of performance measures (n=4), inadequate telemedicine promotion (n=3), etc.

Table 7. Physicians' comments regarding promoting telemedicine.

Suggestions for promoting telemedicine[a]	Respondents (n=127), n (%)
Performance measures[b]	60 (47.2)
Increase telemedicine equipment	22 (17.3)
Policy support	21 (16.5)
Financial support	19 (15.0)
Technical support	18 (14.2)
Increase training	18 (14.2)
Increase telemedicine publicity	14 (11.0)
Others[c]	73 (57.5)

[a]There were a total of 242 comments.

[b]Performance measures included monetary incentives and professional incentives (eg, continuing education credits, facilitating physician promotions, and/or offering time-saving measures for physicians in other aspects of the workday).

[c]Other comments included developing guidelines for telemedicine (n=8), optimization of telemedicine systems (n=7), solving network issues (unable to connect, slow internet performance, etc) (n=5), including telemedicine coverage in health insurance (n=4), increasing the convenience of telemedicine (n=4), harmonious doctor-patient relationships (n=4), etc.

Main Reasons for Being Willing or Unwilling to Use Telemedicine

Physicians' attitudes toward telemedicine were positive, with 88.4% (114/129) of respondents stating that they were willing to adopt telemedicine. Only 8.5% (11/129) of respondents were unwilling to adopt telemedicine, and 4 respondents out of 129 (3.1%) were undecided about whether or not they were willing to adopt telemedicine. The main reasons physicians were willing to adopt telemedicine included convenience for patients (56/114, 49.1%), optimization of medical resources (31/114, 27.2%), and improving the level of medical care (16/114, 14.0%). The main reasons for being willing or unwilling to use telemedicine are given in Table 8.

Table 8. Physicians' attitudes toward telemedicine.

Main reasons physicians were willing or unwilling to use telemedicine[a]	Respondents (N=129), n (%)
Willing (n=114)	114 (88.4)
Convenient for patients	56 (49.1)
Optimized medical resources	31 (27.2)
Improved level of medical care	16 (14.0)
The trend of medical development	8 (7.0)
The COVID-19 pandemic	6 (5.3)
Others[b]	25 (21.9)
Unwilling (n=11)	11 (8.5)
The physician's inability to personally examine a patient will hinder clinical decision making	6 (54.5)
More time spent	3 (27.3)
Low medical fees	2 (18.2)
Concerns about the quality of care	2 (18.2)
Cannot provide valid patient information	2 (18.2)
Others[c]	6 (54.5)
Undecided	4 (3.1)

[a]There were a total of 163 reasons.

[b]Other reasons for being willing to use telemedicine included increased diagnosis and treatment efficiency (n=5), reduced patient burden (n=4), conducive to medical equity (n=2), reduced medical costs (n=1), enhanced patient satisfaction (n=1), etc.

[c]Other reasons for being unwilling to use telemedicine included low economic gain (n=1), patients' distrust of telemedicine (n=1), medical malpractice (n=1), etc.

Discussion

Principal Findings

Although telemedicine has been used in various clinical specialties for decades [29], the emergence of the COVID-19 pandemic has highlighted the importance of telemedicine [30]. In the midst of the global COVID-19 catastrophe, a focus on telemedicine could play a critical role in the provision of global health care and may become a necessity for the general population [31]. In order to make the best use of telemedicine, we need to gain insight into physicians' perceptions of telemedicine.

This study showed that the surveyed physicians had a high willingness to use telemedicine. The reasons for their high willingness were manifold but included the COVID-19 pandemic, telemedicine training courses, as well as young physicians in academic centers. The COVID-19 pandemic forced physicians to quickly adapt and use telemedicine [32]. Physicians' willingness to adopt telemedicine may also be related to the COVID-19 pandemic's movement-restriction policy [33]. Before answering the questionnaire, all the physicians spent more than 3 hours on coursework related to telemedicine. The telemedicine training course increased physicians' awareness of, knowledge about, and attitudes toward telemedicine. There are studies that indicate that the knowledge and perception of health care professionals affect telemedicine adoption [34,35]. Moreover, younger physicians have a greater openness and willingness to adopt telemedicine [36]. One's willingness to use telemedicine may also be influenced by one's attitude toward telemedicine itself, one's level of technology anxiety, and the patient-physician relationship [37]. These factors that were associated with a high willingness to use telemedicine were identified and must be considered in the long-term development of telemedicine.

Although telemedicine has found its way to nearly all clinical specialties, its use is uneven across specialties [38,39]. To promote the development of telemedicine in different specialties, we analyzed the willingness to use, and perceptions of, telemedicine on the part of physicians in different specialties. Due to the uneven distribution of the number of specialists, only specialties that included more than 5 participating physicians were analyzed. Although physicians' willingness to participate in telemedicine was different from the usability of telemedicine in each specialty, there was no correlation between them.

The most obvious concerns and obstacles to telemedicine are limited in-person physical exams and the lack of vital sign assessment. The inability to complete an in-person physical examination was the highest concern for physicians (101/129, 78.3%) and was the main reason physicians cited it as a barrier to implementing telemedicine. This result is consistent with research from the United States [40]. This was mainly due to the concern by physicians that not being able to examine patients in person would affect clinical diagnosis. Whether in the learning stage or late in their careers, physicians want to carefully examine each patient personally. In telemedicine, the inability to examine the patient in person not only affects the physicians' habits, but also sound and light present during telemedicine examinations can affect physicians' diagnoses and treatment recommendations [41]. A well-lit environment and diffuse lighting to reduce glare allow physicians to detect physical examination findings more clearly, such as tremors, convulsions, and subtle facial expressions. Poor sound quality may limit understanding and mutual contact [41-44]. Therefore, health care professionals must be reassured that telemedicine is not a threat to their clinical decision making and that it could allow them to focus on patients who urgently need help. Some authors suggested that telemedicine might be best used in conjunction with face-to-face visits. Physicians can rely on proxies for examination [45].

An important aspect in the application of telemedicine will be the integration of telemedicine with the current health system workflows and the connection to the electronic health record [46]. In order to maximize the benefits of utilizing telemedicine technology, technologies including remote patient monitoring equipment need to be automatically synchronized to the patient's chart, so that physicians can instantly obtain patient data [47]. Clinical decision support in telemedicine should also be enhanced to reduce medical errors.

This study suggests that there are many challenges and risks to telemedicine that need to be addressed before the technology is widely endorsed by physicians. These challenges may be due to regulation, incentives involving telemedicine, effective telemedicine training, malpractice insurance coverage for telemedicine, security and confidentiality of patient data, and telemedicine technology. These are in line with the findings of the other studies [48]. Physicians are less likely to use telemedicine if they are not adequately compensated for their time and effort [49]. Therefore, addressing the barriers to the development of telemedicine will require collaboration and efforts by health care institutions, policy makers, hospital administrators, physicians, and patients.

Limitations of the Study

This study has potential limitations. First, this is a survey-based study and is subject to respondent bias inherent in all survey-based studies. Second, the survey was only about Chinese physicians. Incentive effects may differ in other countries due to cultural differences. Another limitation is the limited sample size and the descriptive nature of the study, which may not be able to reflect the opinions of all physicians in each hospital. However, considering the limited use of telemedicine in China and the lack of knowledge about telemedicine among general physicians, it is difficult to collect opinions through large random sampling. We recruited participants who were physicians and enrolled in a PhD program in clinical informatics. Most of them were also involved with the hospital management team. Therefore, in contrast to general physicians, they have a basic understanding of clinical informatics as well as medical information systems in their own hospital. In addition, the overall response rate was very high (87.2%) and included a variety of clinical specialties. The relatively younger physicians (23 to 48 years old) from the highest-level hospitals represented those who might be more familiar with telemedicine and digital technology. The responses were collected from 55 hospitals in Eastern, Central, and

Western China, as it was a study representing various clinical subspecialties. Moreover, participants spent more than 3 hours on coursework related to telemedicine before completing the survey, so that they had a comprehensive understanding of telemedicine. The survey questions we asked were inherently pragmatic, and the responses to these questions faithfully reflected the physicians' sentiments.

Conclusions

The results of this survey indicate that, although telemedicine cannot yet be used universally for all health care needs and cannot fully replace in-person physical examinations, physicians' willingness to use telemedicine was high. The modality of telemedicine is a tool worthy of careful evaluation and consideration by clinical subspecialties and their medical systems.

Acknowledgments

This work was supported by Sichuan Science and Technology Program (grant 2020YFS0162).

Authors' Contributions

JL and SL conceived the study. JL, SL, TZ, and YB performed the analysis, interpreted the results, and drafted the manuscript. All authors revised the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Telemedicine questionnaire.
[PDF File (Adobe PDF File), 76 KB - medinform_v9i6e26463_app1.pdf]

References

1. Contreras CM, Metzger GA, Beane JD, Dedhia PH, Ejaz A, Pawlik TM. Telemedicine: Patient-provider clinical engagement during the COVID-19 pandemic and beyond. J Gastrointest Surg 2020 Jul;24(7):1692-1697 [FREE Full text] [doi: 10.1007/s11605-020-04623-5] [Medline: 32385614]
2. Liu J, Liu S. The management of coronavirus disease 2019 (COVID-19). J Med Virol 2020 Sep;92(9):1484-1490 [FREE Full text] [doi: 10.1002/jmv.25965] [Medline: 32369222]
3. Bokolo Jnr A. Use of telemedicine and virtual care for remote treatment in response to COVID-19 pandemic. J Med Syst 2020 Jun 15;44(7):132 [FREE Full text] [doi: 10.1007/s10916-020-01596-5] [Medline: 32542571]
4. Whitten P, Holtz B, Laplante C. Telemedicine: What have we learned? Appl Clin Inform 2010 May 05;1(2):132-141 [FREE Full text] [doi: 10.4338/ACI-2009-12-R-0020] [Medline: 23616832]
5. Liu S, Liu J, Zheng T. Current status and trends in health informatics research: A bibliometric analysis by health technology and informatics. Stud Health Technol Inform 2019 Aug 21;264:1960-1961. [doi: 10.3233/SHTI190734] [Medline: 31438428]
6. Almathami HKY, Win KT, Vlahu-Gjorgievska E. Barriers and facilitators that influence telemedicine-based, real-time, online consultation at patients' homes: Systematic literature review. J Med Internet Res 2020 Feb 20;22(2):e16407 [FREE Full text] [doi: 10.2196/16407] [Medline: 32130131]
7. Rogove HJ, McArthur D, Demaerschalk BM, Vespa PM. Barriers to telemedicine: Survey of current users in acute care units. Telemed J E Health 2012;18(1):48-53. [doi: 10.1089/tmj.2011.0071] [Medline: 22082107]
8. Combi C, Pozzani G, Pozzi G. Telemedicine for developing countries. A survey and some design issues. Appl Clin Inform 2016 Nov 02;7(4):1025-1050 [FREE Full text] [doi: 10.4338/ACI-2016-06-R-0089] [Medline: 27803948]
9. Telemedicine: Opportunities and Developments in Member States. Report on the Second Global Survey on eHealth. Geneva, Switzerland: World Health Organization; 2010. URL: https://apps.who.int/iris/bitstream/handle/10665/44497/9789241564144_eng.pdf?sequence=1&isAllowed=y [accessed 2021-05-25]
10. Aoki N, Dunn K, Johnson-Throop KA, Turley JP. Outcomes and methods in telemedicine evaluation. Telemed J E Health 2003;9(4):393-401. [doi: 10.1089/153056203772744734] [Medline: 14980098]
11. Wu T, Parker SA, Jagolino A, Yamal J, Bowry R, Thomas A, et al. Telemedicine can replace the neurologist on a mobile stroke unit. Stroke 2017 Feb;48(2):493-496. [doi: 10.1161/STROKEAHA.116.015363] [Medline: 28082671]
12. Ying G, VanderVeen D, Daniel E, Quinn GE, Baumritter A, Telemedicine Approaches to Evaluating Acute-Phase Retinopathy of Prematurity Cooperative Group. Risk score for predicting treatment-requiring retinopathy of prematurity (ROP) in the Telemedicine Approaches to Evaluating Acute-Phase ROP study. Ophthalmology 2016 Oct;123(10):2176-2182 [FREE Full text] [doi: 10.1016/j.ophtha.2016.06.037] [Medline: 27491396]
13. Serhrouchni S, Malmartel A. Diagnostic agreement between telemedicine on social networks and teledermatology centers. Ann Fam Med 2021;19(1):24-29 [FREE Full text] [doi: 10.1370/afm.2608] [Medline: 33431387]

14. Arbeille P, Provost R, Zuj K, Dimouro D, Georgescu M. Teles-operated echocardiography using a robotic arm and an internet connection. Ultrasound Med Biol 2014 Oct;40(10):2521-2529. [doi: 10.1016/j.ultrasmedbio.2014.05.015] [Medline: 25130450]
15. Kane-Gill SL, Rincon F. Expansion of telemedicine services: Telepharmacy, telestroke, teledialysis, tele-emergency medicine. Crit Care Clin 2019 Jul;35(3):519-533. [doi: 10.1016/j.ccc.2019.02.007] [Medline: 31076051]
16. Vidal-Alaball J, Acosta-Roja R, Pastor Hernández N, Sanchez Luque U, Morrison D, Narejos Pérez S, et al. Telemedicine in the face of the COVID-19 pandemic. Aten Primaria 2020;52(6):418-422 [FREE Full text] [doi: 10.1016/j.aprim.2020.04.003] [Medline: 32402477]
17. Healthcare facilities: Managing operations during the COVID-19 pandemic. Centers for Disease Control and Prevention. 2021. URL: https://www.cdc.gov/coronavirus/2019-ncov/hcp/guidance-hcf.html [accessed 2020-11-26]
18. Harris PA. AMA applauds Medicare telemedicine policy change during pandemic. American Medical Association. 2020 Mar 17. URL: https://www.ama-assn.org/press-center/ama-statements/ama-applauds-medicare-telemedicine-policy-change-during-pandemic [accessed 2020-11-25]
19. Maintaining essential health services: New operational guidance for the COVID-19 context. World Health Organization. 2020 Jun 01. URL: https://www.who.int/news/item/01-06-2020-maintaining-essential-health-services-new-operational-guidance-for-the-covid-19-context [accessed 2020-11-20]
20. Rho MJ, Choi IY, Lee J. Predictive factors of telemedicine service acceptance and behavioral intention of physicians. Int J Med Inform 2014 Aug;83(8):559-571. [doi: 10.1016/j.ijmedinf.2014.05.005] [Medline: 24961820]
21. Kissi J, Dai B, Dogbe CS, Banahene J, Ernest O. Predictive factors of physicians' satisfaction with telemedicine services acceptance. Health Informatics J 2020 Sep;26(3):1866-1880 [FREE Full text] [doi: 10.1177/1460458219892162] [Medline: 31854222]
22. Yu M, He S, Wu D, Zhu H, Webster C. Examining the multi-scalar unevenness of high-quality healthcare resources distribution in China. Int J Environ Res Public Health 2019 Aug 07;16(16):2813 [FREE Full text] [doi: 10.3390/ijerph16162813] [Medline: 31394765]
23. Zachrison KS, Boggs KM, Hayden EM, Espinola JA, Camargo CA. Understanding barriers to telemedicine implementation in rural emergency departments. Ann Emerg Med 2020 Mar;75(3):392-399. [doi: 10.1016/j.annemergmed.2019.06.026] [Medline: 31474481]
24. Uscher-Pines L, Kahn JM. Barriers and facilitators to pediatric emergency telemedicine in the United States. Telemed J E Health 2014 Nov;20(11):990-996 [FREE Full text] [doi: 10.1089/tmj.2014.0015] [Medline: 25238565]
25. Hu PJ, Chau PY, Sheng ORL, Tam KY. Examining the technology acceptance model using physician acceptance of telemedicine technology. J Manag Inf Syst 2015 Dec 02;16(2):91-112. [doi: 10.1080/07421222.1999.11518247]
26. Hu PJ, Chau PY. Physician acceptance of telemedicine technology: An empirical investigation. Top Health Inf Manage 1999 May;19(4):20-35. [Medline: 10387653]
27. Kissi J, Dai B, Dogbe CS, Banahene J, Ernest O. Predictive factors of physicians' satisfaction with telemedicine services acceptance. Health Informatics J 2020 Sep;26(3):1866-1880 [FREE Full text] [doi: 10.1177/1460458219892162] [Medline: 31854222]
28. Paul D, Pearlson K, McDaniel R. Assessing technological barriers to telemedicine: Technology-management implications. IEEE Trans Eng Manag 1999 Aug;46(3):279-288. [doi: 10.1109/17.775280]
29. Whitten P, Holtz B, LaPlante C. Telemedicine. Appl Clin Inform 2017 Dec 20;01(02):132-141. [doi: 10.4338/aci-2009-12-r-0020]
30. Giansanti D. The Italian fight against the COVID-19 pandemic in the second phase: The renewed opportunity of telemedicine. Telemed J E Health 2020 Nov;26(11):1328-1331. [doi: 10.1089/tmj.2020.0212] [Medline: 32579866]
31. Smith AC, Thomas E, Snoswell CL, Haydon H, Mehrotra A, Clemensen J, et al. Telehealth for global emergencies: Implications for coronavirus disease 2019 (COVID-19). J Telemed Telecare 2020 Mar 20;26(5):309-313. [doi: 10.1177/1357633x20916567]
32. Ohannessian R, Duong TA, Odone A. Global telemedicine implementation and integration within health systems to fight the COVID-19 pandemic: A call to action. JMIR Public Health Surveill 2020 Apr 02;6(2):e18810 [FREE Full text] [doi: 10.2196/18810] [Medline: 32238336]
33. Wahezi SE, Kohan LR, Spektor B, Brancolini S, Emerick T, Fronterhouse JM, et al. Telemedicine and current clinical practice trends in the COVID-19 pandemic. Best Pract Res Clin Anaesthesiol 2020 Nov 16(In Press):1-13 [FREE Full text] [doi: 10.1016/j.bpa.2020.11.005]
34. Ayatollahi H, Sarabi FZP, Langarizadeh M. Clinicians' knowledge and perception of telemedicine technology. Perspect Health Inf Manag 2015 Nov 01;12:1c [FREE Full text] [Medline: 26604872]
35. Malhotra P, Ramachandran A, Chauhan R, Soni D, Garg N. Assessment of knowledge, perception, and willingness of using telemedicine among medical and allied healthcare students studying in private institutions. Telehealth Med Today 2020 Nov 27;5(4):1-14 [FREE Full text] [doi: 10.30953/tmt.v5.228]
36. Helou S, El Helou E, Abou-Khalil V, Wakim J, El Helou J, Daher A, et al. The effect of the COVID-19 pandemic on physicians' use and perception of telehealth: The case of Lebanon. Int J Environ Res Public Health 2020 Jul 06;17(13):4866 [FREE Full text] [doi: 10.3390/ijerph17134866] [Medline: 32640652]

37. Zayapragassarazan Z, Kumar S. Awareness, knowledge, attitude and skills of telemedicine among health professional faculty working in teaching hospitals. J Clin Diagn Res 2016 Mar;10(3):JC01-JC04 [FREE Full text] [doi: 10.7860/JCDR/2016/19080.7431] [Medline: 27134899]
38. Aghdam MRF, Vodovnik A, Hameed RA. Role of telemedicine in multidisciplinary team meetings. J Pathol Inform 2019 Nov 18;10:35 [FREE Full text] [doi: 10.4103/jpi.jpi_20_19] [Medline: 31799021]
39. Zulfiqar AA, Hajjam A, Andrès E. Focus on the different projects of telemedicine centered on the elderly in France. Curr Aging Sci 2019;11(4):202-215 [FREE Full text] [doi: 10.2174/1874609812666190304115426] [Medline: 30836931]
40. Salehi PP, Torabi SJ, Lee YH, Azizzadeh B. Telemedicine practices of facial plastic and reconstructive surgeons in the United States: The effect of novel coronavirus-19. Facial Plast Surg Aesthet Med 2020;22(6):464-470. [doi: 10.1089/fpsam.2020.0409] [Medline: 33054375]
41. Cowan KE, McKean AJ, Gentry MT, Hilty DM. Barriers to use of telepsychiatry: Clinicians as gatekeepers. Mayo Clin Proc 2019 Dec;94(12):2510-2523. [doi: 10.1016/j.mayocp.2019.04.018] [Medline: 31806104]
42. Myers K, Nelson E, Rabinowitz T, Hilty D, Baker D, Barnwell S, et al. American Telemedicine Association practice guidelines for telemental health with children and adolescents. Telemed Ehealth 2017 Oct 01;23(10):779-804. [doi: 10.1089/tmj.2017.0177]
43. American Academy of Child and Adolescent Psychiatry (AACAP) Committee on Telepsychiatry and AACAP Committee on Quality Issues. Clinical update: Telepsychiatry with children and adolescents. J Am Acad Child Adolesc Psychiatry 2017 Oct;56(10):875-893. [doi: 10.1016/j.jaac.2017.07.008] [Medline: 28942810]
44. Parish MB, Fazio S, Chan S, Yellowlees PM. Managing psychiatrist-patient relationships in the digital age: A summary review of the impact of technology-enabled care on clinical processes and rapport. Curr Psychiatry Rep 2017 Oct 27;19(11):90. [doi: 10.1007/s11920-017-0839-x] [Medline: 29075951]
45. Roberts LJ, Lamont EG, Lim I, Sabesan S, Barrett C. Telerheumatology: An idea whose time has come. Intern Med J 2012 Oct;42(10):1072-1078. [doi: 10.1111/j.1445-5994.2012.02931.x] [Medline: 22931307]
46. Khoong EC, Rivadeneira NA, Hiatt RA, Sarkar U. The use of technology for communicating with clinicians or seeking health information in a multilingual urban cohort: Cross-sectional survey. J Med Internet Res 2020 Apr 06;22(4):e16951 [FREE Full text] [doi: 10.2196/16951] [Medline: 32250280]
47. Hollander JE, Carr BG. Virtually perfect? Telemedicine for Covid-19. N Engl J Med 2020 Apr 30;382(18):1679-1681. [doi: 10.1056/NEJMp2003539] [Medline: 32160451]
48. Almathami HKY, Win KT, Vlahu-Gjorgievska E. Barriers and facilitators that influence telemedicine-based, real-time, online consultation at patients' homes: Systematic literature review. J Med Internet Res 2020 Feb 20;22(2):e16407 [FREE Full text] [doi: 10.2196/16407] [Medline: 32130131]
49. Pong RW, Hogenbirk JC. Reimbursing physicians for telehealth practice: Issues and policy options. Health Law Rev 2000 Jan;9(1):3-13 [FREE Full text]

Edited by C Lovis; submitted 12.12.20; peer-reviewed by P Li, EM Schomakers, R Cochran; comments to author 28.01.21; revised version received 08.02.21; accepted 03.05.21; published 01.06.21.

Please cite as:
Liu J, Liu S, Zheng T, Bi Y
Physicians' Perspectives of Telemedicine During the COVID-19 Pandemic in China: Qualitative Survey Study
JMIR Med Inform 2021;9(6):e26463
URL: https://medinform.jmir.org/2021/6/e26463
doi:10.2196/26463
PMID:33945493

Original Paper

Physicians' Attitudes Toward Telemedicine Consultations During the COVID-19 Pandemic: Cross-sectional Study

Noora Alhajri[1], MD, MPH; Mecit Can Emre Simsekler[2], PhD; Buthaina Alfalasi[3], BCh, BAO, MB; Mohamed Alhashmi[1], BEng; Majd AlGhatrif[4], MD; Nahed Balalaa[5], MBBCh; Maryam Al Ali[6], BCh, BAO, MB; Raghda Almaashari[7], BCh, BAO, MB; Shammah Al Memari[8], MBBS; Farida Al Hosani[8], MBBS, DrPH, MPH; Yousif Al Zaabi[8], BD, MPH, MSc; Shereena Almazroui[8], MBBS, MPH; Hamed Alhashemi[9], PhD; Ovidiu C Baltatu[1], MD, PhD

[1]Khalifa University College of Medicine and Health Science, Abu Dhabi, United Arab Emirates

[2]College of Engineering, Khalifa University, Abu Dhabi, United Arab Emirates

[3]Department of Family Medicine, Zayed Military Hospital, Abu Dhabi, United Arab Emirates

[4]Department of Medicine, The Johns Hopkins University School of Medicine, Baltimore, MD, United States

[5]Department of General Surgery, Sheikh Shakhbout Medical City, Abu Dhabi, United Arab Emirates

[6]Ambulatory Health Services, Zafarana Clinic, Abu Dhabi Healthcare Company, Abu Dhabi, United Arab Emirates

[7]Department of Dermatology, Sheikh Khalifa Medical City, Abu Dhabi, United Arab Emirates

[8]Abu Dhabi Public Health Center, Department of Health, Abu Dhabi, United Arab Emirates

[9]Department of Health, Abu Dhabi, United Arab Emirates

Corresponding Author:
Ovidiu C Baltatu, MD, PhD
Khalifa University College of Medicine and Health Science
Saadat St, Zone 1
Abu Dhabi
United Arab Emirates
Phone: 971 552277490
Email: ocbaltatu@gmail.com

Abstract

Background: To mitigate the effect of the COVID-19 pandemic, health care systems worldwide have implemented telemedicine technologies to respond to the growing need for health care services during these unprecedented times. In the United Arab Emirates, video and audio consultations have been implemented to deliver health services during the pandemic.

Objective: This study aimed to evaluate whether differences exist in physicians' attitudes and perceptions of video and audio consultations when delivering telemedicine services during the COVID-19 pandemic.

Methods: This survey was conducted on a cohort of 880 physicians from outpatient facilities in Abu Dhabi, which delivered telemedicine services during the COVID-19 pandemic between November and December 2020. In total, 623 physicians responded (response rate=70.8%). The survey included a 5-point Likert scale to measure physician's attitudes and perceptions of video and audio consultations with reference to the quality of the clinical consultation and the professional productivity. Descriptive statistics were used to describe physicians' sociodemographic characteristics (age, sex, designation, clinical specialty, duration of practice, and previous experience with telemedicine) and telemedicine modality (video vs audio consultations). Regression models were used to assess the association between telemedicine modality and physicians' characteristics with the perceived outcomes of the web-based consultation.

Results: Compared to audio consultations, video consultations were significantly associated with physicians' confidence toward managing acute consultations (odds ratio [OR] 1.62, 95% CI 1.2-2.21; *P*=.002) and an increased ability to provide patient education during the web-based consultation (OR 2.21, 95% CI 1.04-4.33; *P*=.04). There was no significant difference in physicians' confidence toward managing long-term and follow-up consultations through video or audio consultations (OR 1.35, 95% CI 0.88-2.08; *P*=.17). Video consultations were less likely to be associated with a reduced overall consultation time (OR 0.69, 95% CI 0.51-0.93; *P*=.02) and reduced time for patient note-taking compared to face-to-face visits (OR 0.48, 95% CI 0.36-0.65; *P*<.001). Previous experience with telemedicine was significantly associated with a lower perceived risk of misdiagnosis (OR 0.46, 95% CI 0.3-0.71; *P*<.001) and an enhanced physician-patient rapport (OR 2.49, 95% CI 1.26-4.9; *P*=.008).

Conclusions: These results indicate that video consultations should be adopted frequently in the new remote clinical consultations. Previous experience with telemedicine was associated with a 2-fold confidence in treating acute conditions, less than a half of the perceived risk of misdiagnosis, and an increased ability to provide patients with health education and enhance the physician-patient rapport. Additionally, these results show that audio consultations are equivalent to video consultations in providing remote follow-up care to patients with chronic conditions. These findings may be beneficial to policymakers of e-health programs in low- and middle-income countries, where audio consultations may significantly increase access to geographically remote health services.

(JMIR Med Inform 2021;9(6):e29251) doi:10.2196/29251

KEYWORDS

audio consultation; clinical decision-making; clinical training; communication; COVID-19; outpatient department; perception; telemedicine; United Arab Emirates; video consultation

Introduction

The COVID-19 pandemic has caused an enormous burden on the health care system and health care delivery worldwide [1-4]. As social distancing and quarantining became the new normal, face-to-face clinical visits plummeted, causing the health care system to rapidly shift to telemedicine to leverage their response to the pandemic [5-8]. Telemedicine created new opportunities for patient care in the context of the COVID-19 pandemic and thus reduced health care disparities [9,10]. Telemedicine is available in various modalities including patient portals, emails, text messages, telemonitoring, store-and-forward, audio consultations, and real-time video consultations [10-13]. The wide variety in communication channels offer different opportunities for providers to manage patients who are in quarantine or live in remote areas, which reduces the risk of disease transmission and improves access to health care services [5,9,14,15].

Owing to the growing concern regarding the risk of workplace transmission, the use of telemedicine services increased globally [16-19], and the United Arab Emirates is no different. In March 2020, Abu Dhabi launched its first Telemedicine Virtual Outpatient Clinic to support the continuity of patient care [20]. It has been estimated that within only 1 month, physicians across Abu Dhabi SEHA hospitals performed over 28,000 virtual consultations [21,22].

Studies conducted on telemedicine during the COVID-19 pandemic, while yielding meaningful insights on its role, have largely been based on physician knowledge of telemedicine in specific subspecialities and have been limited to descriptive data of certain encounters rather than quantifying their association. Currently, the effect of video vs audio consultations on physicians' attitude toward telemedicine is unclear [23,24]. Moreover, barriers against its full implementation beyond the context of the COVID-19 pandemic remain unexplored. Identifying these barriers within each modality, which prevent their successful adoption by health care providers, is essential for directing future infrastructure to modernize the health care system and improve telemedicine utilization and outcomes. This study aimed to describe physicians' attitudes toward the use of telemedicine services in Abu Dhabi during the COVID-19 pandemic. We also aimed to explore the effects of audio vs video consultations and physicians' sociodemographic characteristics on their confidence during the clinical consultation, perceived quality of care, and perceived effects of professional productivity. Future studies are needed to objectively assess the effect of telemedicine modalities on the quality of care and professional productivity and to guide future infrastructure investments to assure embracing this new opportunity to provide high-quality health care to a larger number of patients in the post–COVID-19 era.

Methods

Study Design and Ethics Approval

This was a survey-based study conducted on physicians in outpatient facilities in Abu Dhabi, which provided telemedicine services during the COVID-19 pandemic between November and December 2020. Ethics approval was obtained from the institutional review board of Khalifa University (protocol# H21-006-2020) and of the Abu Dhabi COVID-19 Research Committee of the Department of Health in Abu Dhabi (reference# DOH/CVDC/2020/1747). Surveys were administered through the Department of Health and SEHA, these being the major health authorities in Abu Dhabi. The institutional review board or ethics committee at each participating institution approved the study protocol and survey. Electronic written consent was waived for this data-only study owing to the deidentified nature of this survey. The present study followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) reporting guidelines for cross-sectional studies [25].

Subject Selection and the Inclusion and Exclusion Criteria

The survey was administered to a cohort of 880 physicians at outpatient facilities in Abu Dhabi, who met the following inclusion criteria: being a physician practicing at an outpatient facility in Abu Dhabi and providing audio or video consultations during the COVID-19 pandemic from January to November 2020. Exclusion criteria were being of another allied health care profession such as nurses, pharmacists, and technicians (as our study targeted physicians only) or physicians who did not work at outpatient departments and who did not use telemedicine during the COVID-19 pandemic. From a total of 880 physicians listed, 623 responded to the survey (response rate=70.8%).

Survey Development, Piloting, and Data Collection

A web-based structured survey containing multiple-choice questions was developed by reviewing published telemedicine surveys and their instruments [26-28]. The web-based survey had 6 components, which contained a total of 42 questions related to physicians' perceptions and attitudes toward telemedicine. A pilot survey was conducted, which included a cohort of 25 physicians in Abu Dhabi, who frequently used telemedicine during the COVID-19 pandemic. The main web-based survey was developed using the Microsoft Forms platform (Microsoft Corp) and was sent to the physicians at outpatient facilities via the hospital's internal email system. To reduce the risk of attrition bias, we ensured generating a good rapport between on-site principal investigators and the study participants by sending customized invitations [29,30]. Furthermore, a follow-up email was sent 1 week apart from the initial date of survey distribution to remind nonresponders to participate in the survey.

Study Variables and Outcomes

This was a self-administered survey that gathered data on physicians' sociodemographic characteristics including age, sex, telemedicine modality, clinical specialty, designation, number of years in practice, and past experience with telemedicine. We also gathered data using a 5-point Likert scale to assess (1) physicians' current experience with telemedicine, (2) perceived quality of the web-based clinical consultation, (3) satisfaction with telemedicine, (4) perceived professional productivity compared to traditional face-to-face visits, (5) willingness to use telemedicine after the pandemic, and (6) perceived barriers to telemedicine use. Data on these 6 components were gathered to understand the telemedicine experience better during the COVID-19 pandemic and to gain insights into the preparedness of the digital health care response for any potential crisis. We defined "acute remote care consultation" as any remote consultation made for the first time owing to an urgent medical complaint, the onset of a new disease, or a follow-up case that has not received a consultation for more than 6 months. Furthermore, "chronic remote care consultation" was defined as any remote follow-up consultation within 6 months of the initial in-person visit for a long-term medical condition [31].

Statistical Analysis

Differences between video and audio consultations were investigated using various outcome variables, which included 2 main parts. While the first set of outcomes was related to the perceived quality of clinical consultations, the second set of outcomes tested physicians' professional productivity with telemedicine over face-to-face consultations.

Descriptive statistics characterizing the study cohort were reported as frequency and percentage values for all variables. To compare the responses to our survey questions with regard to video and audio consultations, we performed chi-square analysis at a significance level of .05.

We used ordered logistic regression analyses to investigate the association between outcome variables and the modality, adjusting for confounding factors such as sociodemographic characteristics. A forced-entry approach was adopted to consider the variance inflation factor (VIF) diagnostic to prevent obtaining unreliable estimates of coefficients and odds ratios (ORs) owing to high correlations among predictor variables. Considering the high VIF for the variable of the number of years in practice (VIF>4), we excluded this variable and confirmed that multicollinearity is not a concern in the final models (VIF=1.51). Further, the Akaike information criterion was used to assess the fit of the models after excluding the variable of the number of years in practice. Survey questions were answered on a 5-point Likert scale where 5="strongly agree", 4="agree", 3="neutral", 2="disagree", and 1="strongly disagree". However, owing to limited observations toward the extreme ends of the scale ("strongly agree" and "strongly disagree"), we merged the responses of "strongly agree" and "agree" under positive responses and "strongly disagree" and "disagree" under negative responses together as these 2 statements were found to involve the same attitude continuum toward the question [32]; these were grouped under "disagreement," "neutral," and "agreement." The outcomes of the regression models were reported as ORs with 95% CI values, and $P<.05$ indicated significance. Statistical analyses were performed using STATA (version 16.1, Stata Corp).

Results

Overall, 623 physicians completed the survey, of whom 347 (55.7%) conducted only audio consultations and 276 (44.3%) conducted only video consultations during the COVID-19 pandemic. The sociodemographic descriptive characteristics of the 2 groups are summarized and compared in Table 1.

Table 1. Sociodemographic characteristics of the physicians included in the study (N=623) and descriptive statistics by modality.

Sociodemographic characteristics	Audio consultations (n=347), n (%)	Video consultations (n=276), n (%)	Total, n (%)	*P* value
Sex				.04
Female	163 (46.97)	107 (38.77)	270 (43.34)	
Male	184 (53.03)	169 (61.23)	353 (56.66)	
Age (years)				.52
≤39	87 (25.07)	59 (21.38)	146 (23.43)	
40-49	138 (39.77)	116 (42.03)	254 (40.77)	
50-59	83 (23.92)	62 (22.46)	145 (23.27)	
≥60	39 (11.24)	39 (14.13)	78 (12.52)	
Specialty				.23
Internal medicine	213 (61.38)	186 (67.39)	399 (64.04)	
Surgical specialties	38 (10.95)	22 (7.97)	60 (9.63)	
Family medicine	76 (21.90)	48 (17.39)	124 (19.90)	
Others[a]	20 (5.76)	20 (7.25)	40 (6.42)	
Physician designation				.13
General physician	62 (17.87)	41 (14.86)	103 (16.53)	
Resident	8 (2.31)	3 (1.09)	11 (1.77)	
Specialist	189 (54.47)	175 (63.41)	364 (58.43)	
Consultant	88 (25.36)	57 (20.65)	145 (23.27)	
Number of years in practice				.32
≤4	16 (4.61)	10 (3.62)	26 (4.17)	
5-9	52 (14.99)	33 (11.96)	85 (13.64)	
10-20	132 (38.04)	124 (44.93)	256 (41.09)	
>20	147 (42.36)	109 (39.49)	256 (41.09)	
Past experience with telemedicine				.09
Never used	256 (73.78)	219 (79.35)	475 (76.24)	
Used a few times	75 (21.61)	41 (14.86)	116 (18.62)	
Used frequently	16 (4.61)	16 (5.80)	32 (5.14)	

[a]Other specialties include speech therapy, dentistry, physical medicine and rehabilitation, anesthesiology, emergency medicine, occupational therapy, radiology, aviation and occupational health, periodontics, gynecology center, nutrition, urgent care, prosthodontics, and critical care medicine.

Sociodemographic Characteristics

Compared to physicians who provided audio consultations, those who provided video consultations were predominantly male (61.23% vs 53.03%, respectively; *P*=.04), middle-aged (40-49 years: 42.03% vs 39.77%, 50-59 years: 22.46% vs 23.92%, ≥60 years: 14.13% vs 11.24%; *P*=.52), and had a different specialty distribution with most belonging to internal medicine subspecialities (67.39% vs 61.38%; *P*=.23). Additionally, physicians who provided video consultations were mostly specialists with 10-20 years of experience in practice. In relation to previous experience with telemedicine modalities, there was a variation in responses. The majority of physicians who provided video consultations during the COVID-19 pandemic reported that they had never used this form of telemedicine previously, compared to their counterparts who provided audio consultations (79.35% vs 73.78%, respectively; *P*=.09); conversely, the proportion of physicians who reported frequent provision of video consultations was higher than that of their counterparts who provided audio consultations (5.80% vs 4.61%; *P*=.09).

Perceived Quality of Clinical Care Provided

Physicians' agreement with the following statements was assessed: (1) I was confident in managing acute conditions, (2) I was confident in managing chronic conditions, (3) I was able to answer my patients' questions, (4) I was able to provide health education to patients, and (5) I had an impression of misdiagnosis risk during the teleconsultation. The proportions of physicians who agreed, disagreed, or were neutral about the statements are indicated in Table 2**.** Overall, more than half of the physicians who provided video consultations agreed that they were confident in diagnosing acute conditions (*P*=.01), confident in diagnosing chronic conditions (*P*=.08), and able

to provide patient health education during the clinical consultation, which was significantly higher than that of physicians who provided audio consultations (*P*=.006). However, there was no significant difference in the perceived risk of misdiagnosis (*P*=.41) and the physicians' ability to address the patients' questions (*P*=.26) among those who provided video or audio consultations. Remarkably, the proportion of male physicians who believed that telemedicine raises the likelihood of misdiagnosis was higher than the proportion of female physicians (*P*=.02) (Multimedia Appendix 1).

Table 2. Comparison of survey responses on the perceived quality of clinical care provided by modality.

Perceived quality of clinical care provided	Audio consultations, n (%)	Video consultations, n (%)	Total, n (%)	*P* value
Confidence in managing acute consultations				.01
Disagree and strongly disagree	85 (24.50)	47 (17.03)	132 (21.19)	
Neutral	121 (34.87)	85 (30.80)	206 (33.07)	
Agree and strongly agree	141 (40.63)	144 (52.17)	285 (45.75)	
Confidence in managing chronic conditions and follow-up consultations				.08
Disagree and strongly disagree	20 (5.76)	6 (2.17)	26 (4.17)	
Neutral	50 (14.41)	39 (14.13)	89 (14.29)	
Agree and strongly agree	277 (79.83)	231 (83.70)	508 (81.54)	
Ability to answer patients' questions				.26
Disagree and strongly disagree	9 (2.59)	3 (1.09)	12 (1.93)	
Neutral	36 (10.37)	23 (8.33)	59 (9.47)	
Agree and strongly agree	302 (87.03)	250 (90.58)	552 (88.60)	
Ability to provide patient health education				.006
Disagree and strongly disagree	15 (4.32)	1 (0.36)	16 (2.57)	
Neutral	36 (10.37)	24 (8.70)	60 (9.63)	
Agree and strongly agree	296 (85.30)	251 (90.94)	547 (87.80)	
Perceived risk of misdiagnosis with telemedicine				.41
Disagree and strongly disagree	47 (13.54)	35 (12.68)	82 (13.16)	
Neutral	84 (24.21)	80 (28.99)	164 (26.32)	
Agree and strongly agree	216 (62.25)	161 (58.33)	377 (60.51)	

Perceived Professional Productivity

The overall response to this survey section varied across the entire sample, with no significant difference in the physician-patient rapport among those who provided video or audio consultations compared to face-to-face consultations (*P*=.95) (Table 3). Interestingly, when compared to face-to-face consultations, the proportion of physicians who perceived that telemedicine reduces the overall documentation time (*P*<.001) and increases the total number of patient consultations (*P*=.01) was significantly higher among physicians who provided audio consultations than among those who provided video consultations. The proportion of female physicians who agreed that telemedicine decreases the overall documentation time and increases the total number of patient consultations was substantially higher than that of their male counterparts (*P*=.008 and *P*<.001, respectively) (Multimedia Appendix 1).

Table 3. Comparison of survey responses on perceived professional productivity by modality.

Perceived professional productivity	Audio consultations, n (%)	Video consultations, n (%)	Total, n (%)	*P* value
Patient's rapport rather than face-to-face consultations				.95
Disagree and strongly disagree	228 (65.71)	179 (64.86)	407 (65.33)	
Neutral	83 (23.92)	69 (25.00)	152 (24.40)	
Agree and strongly agree	36 (10.37)	28 (10.14)	64 (10.27)	
Reduced overall consultation time rather than face-to-face consultations				0.066
Disagree and strongly disagree	80 (23.05)	84 (30.43)	164 (26.32)	
Neutral	94 (27.09)	77 (27.9)	171 (27.45)	
Agree and strongly agree	173(49.86	115 (41.67)	288 (46.23)	
Reduced overall documentation time rather than face-to-face consultations				<0.001
Disagree and strongly disagree	84 (24.21)	104 (37.68)	188 (30.18)	
Neutral	77 (22.19)	77 (27.9)	154 (24.72)	
Agree and strongly agree	186 (53.6)	95 (34.42)	281 (45.1)	
Increased total number of consulted patients rather than face-to-face consultations				0.01
Disagree and strongly disagree	89 (25.65)	95 (34.42)	184 (29.53)	
Neutral	112 (32.28)	94 (34.06)	206 (33.07)	
Agree and strongly agree	146 (42.07)	87 (31.52)	233 (37.4)	

Working Experience, Satisfaction, and Barriers to Telemedicine

The majority of physicians who provided video consultations agreed that they received sufficient technological support during the web-based consultation; this proportion was greater than that of physicians who provided audio consultations (76.45% vs 53.60%, respectively; *P*<.001).

There was no significant difference in the satisfaction with the quality of the clinical consultation between physicians who provided video consultations and those who provided audio consultations (*P*=.07).

On assessing the barriers to telemedicine, physicians who provided audio consultations reported that the "inability to see the patient during the consultation" was a significant barrier to the quality of the remote clinical consultations (*P*=.001), and they preferred not to use telemedicine services owing to low payment and reimbursement rates (*P*=.004), were unable to confirm the patient's identity during the audio consultation (*P*=.04), and reported that lack of training is a barrier to the use of telemedicine services to provide remote care to patients (*P*<.001) (Multimedia Appendix 1).

Multivariate Analysis

In the multivariate regression model, video consultations were associated with significantly improved confidence toward the management of acute conditions (OR 1.62, 95% CI 1.2-2.21; *P*=.002) and increased perceived ability to provide patient education (OR 2.21, 95% CI 1.04-4.33; *P*=.04), while male sex was associated with a lower perceived ability to provide patient education during the web-based consultation (OR 0.48, 95% CI 0.27-0.84; *P*=.01). There was no significant difference in physician's confidence in managing chronic conditions or conducting follow-up consultations among those who provided audio or video consultations Table 4. Additionally, previous experience with frequent telemedicine consultations was significantly associated with higher confidence in diagnosing acute conditions (OR=2.12, 95% CI:1.04-4.33 *P*=.039) and with a lower perceived risk of misdiagnosis (OR 0.46, 95% CI 0.31-0.68; *P*<.001). Our analysis also shows that video consultations were significantly associated with a perceived increase in overall consultation time, overall documentation time, and a reduction in the overall number of patients consulted when compared to face-to-face clinical consultations. Previous experience with telemedicine was significantly associated with the perception of an enhanced physician-patient rapport and the perception of an increased total number of patient consultations when compared to face-to-face consultations (Table 5).

Table 4. Adjusted multivariate analysis for the perceived quality of clinical consultations.

Variables	Confidence in managing acute conditions		Confidence in managing chronic conditions and follow-up consultations		Ability to answer patient questions		Ability to provide patient health education		Perceived risk of misdiagnosis with telemedicine	
	OR[a] (95% CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value
Modality (video vs audio consultations)	1.62 (1.2-2.21)	.002	1.35 (0.88-2.08)	.17	1.6 (0.94-2.74)	.08	2.02 (1.19-3.41)	.009	0.81 (0.58-1.12)	.20
Sex (male vs female)	0.84 (0.61-1.16)	.29	0.76 (0.48-1.20)	.24	0.57 (0.32-1.02)	.06	0.48 (0.27-0.84)	.01	1.23 (0.87-1.74)	.25
Age (years)										
40-49 vs <39	0.9 (0.6-1.36)	.62	1.16 (0.66-2.04)	.61	1.06 (0.52-2.13)	.88	1.66 (0.85-3.25)	.14	1.23 (0.8-1.89)	.35
50-59 vs <39	0.95 (0.6-1.51)	.83	1.53 (0.79-2.94)	.20	1.49 (0.67-3.33)	.33	1.34 (0.65-2.76)	.44	1.11 (0.67-1.82)	.68
≥60 vs <39	0.82 (0.46-1.44)	.49	1.7 (0.74-3.94)	.21	1.11 (0.43-2.92)	.83	2.17 (0.81-5.82)	.13	0.85 (0.47-1.54)	.60
Specialty										
Surgical specialties vs internal medicine	1.42 (0.83-2.41)	.20	1.06 (0.51-2.20)	.87	0.95 (0.41-2.19)	.90	1.42 (0.59-3.37)	.43	1.11 (0.62-1.99)	.71
Family medicine vs internal medicine	1.46 (0.92-2.32)	.11	1.38 (0.69-2.74)	.36	1.56 (0.61-3.95)	.35	1.26 (0.56-2.87)	.58	0.67 (0.42-1.09)	.10
Others vs internal medicine	1.52 (0.79-2.94)	.21	0.18 (0.09-0.37)	<.001	0.21 (0.09-0.49)	<.001	0.36 (0.16-0.84)	.02	0.54 (0.29-1.03)	.06
Physician designation										
Resident vs general physician	1.03 (0.33-3.19)	.96	0.74 (0.16-3.34)	.70	0.30 (0.06-1.47)	.14	0.66 (0.11-3.77)	.64	1.00 (0.31-3.23)	.99
Specialist vs general physician	1.34 (0.81-2.20)	.26	0.93 (0.46 - 1.87)	.84	0.70 (0.28-1.74)	.45	0.67 (0.29-1.57)	.36	1.20 (0.72-2.00)	.48
Consultant vs general physician	0.99 (0.56-1.75)	.96	0.48 (0.22-1.06)	.07	0.49 (0.17-1.36)	.17	0.57 (0.21-1.51)	.26	1.18 (0.65-2.15)	.58
Past experience with telemedicine										
Used few times vs never used	1.31 (0.88-1.93)	.18	0.79 (0.48-1.33)	.38	1.36 (0.69-2.70)	.38	1.43 (0.74-2.77)	.29	0.46 (0.31-0.68)	<.001
Used frequently vs never used	2.12 (1.04-4.33)	.04	1.37 (0.46-4.10)	.57	3.65 (0.48-27.63)	.21	1.26 (0.36-4.41)	.71	0.45 (0.22-0.91)	.03

[a]OR: odds ratio.

Table 5. Adjusted multivariate analysis for perceived professional productivity.

Variables	Patient rapport rather than face-to-face consultations		Reduced overall consultation time rather than face-to-face consultations		Reduced overall documentation time rather than face-to-face consultations		Total number of consulted patients rather than face-to-face consultations	
	OR[a] (95% CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value
Modality (video vs audio consultation	1.07 (0.76-1.49)	.71	0.69 (0.51-0.93)	.02	0.48 (0.36-0.65)	<.001	0.66 (0.49-0.89)	.006
Sex (male vs female)	0.69 (0.48-0.99)	.04	1.00 (0.73-1.39)	.98	0.72 (0.52-1.00)	.05	0.60 (0.43-0.82)	.002
Age (years)								
40-49 vs <39	1.00 (0.64-1.56)	.99	0.75 (0.50-1.13)	.17	0.80 (0.53-1.20)	.28	0.81 (0.54-1.22)	.32
50-59 vs <39	1.40 (0.84-2.32)	.20	1.15 (0.72-1.84)	.56	0.89 (0.56-1.43)	.64	1.10 (0.70-1.76)	.67
≥60 vs <39	1.31 (0.70-2.44)	.39	1.25 (0.70-2.21)	.45	1.40 (0.78-2.53)	.26	1.26 (0.72-2.21)	.42
Specialty								
Surgical specialties vs internal medicine	2.03 (1.15-3.59)	.02	1.08 (0.64-1.81)	.78	0.94 (0.56-1.59)	.83	1.20 (0.73-1.97)	.48
Family medicine vs internal medicine	0.93 (0.56-1.53)	.77	1.30 (0.83-2.03)	.26	1.34 (0.84-2.15)	.22	1.19 (0.75-1.89)	.46
Others vs internal medicine	1.45 (0.74-2.84)	.27	0.74 (0.41-1.36)	.34	0.78 (0.42-1.44)	.43	1.22 (0.65-2.28)	.54
Physician designation								
Resident vs general physician	1.45 (0.46-4.63)	.53	0.97 (0.33-2.88)	.96	0.99 (0.3-3.25)	.98	0.76 (0.23-2.51)	.65
Specialist vs general physician	0.85 (0.51-1.43)	.54	0.99 (0.61-1.60)	.97	1.11 (0.67-1.82)	.69	0.91 (0.56-1.48)	.70
Consultant vs general physician	0.46 (0.25-0.86)	.02	0.61 (0.35-1.05)	.08	0.66 (0.37-1.17)	.16	0.59 (0.33-1.03)	.06
Past experience with telemedicine								
Used few times vs never used	1.46 (0.96-2.23)	.08	0.96 (0.66-1.41)	.85	0.96 (0.65-1.41)	.84	1.46 (0.99-2.14)	.05
Used frequently vs never used	2.49 (1.26-4.90)	.008	1.72 (0.84-3.54)	.14	0.80 (0.41-1.57)	.52	2.81 (1.38-5.71)	.004

[a]OR: odds ratio.

Discussion

Principal Findings

This analysis of 623 physicians shows that video consultations are independently associated with a 62% increase in confidence in managing acute conditions, and physicians who provided video consultations were 2-fold more likely to provide patient education during the web-based consultations. Moreover, previous experience with telemedicine was associated with a 2-fold increase in confidence in managing acute conditions and a 55% reduction in the perception of the risk of misdiagnosis. More than one-third (37.68%) of physicians who provided video consultations did not agree that telemedicine reduces the overall consultation time, and approximately one-third (34.42%) did not agree that telemedicine increases the overall number of patient consultations when compared to face-to-face visits. Additionally, those who had previous experience with telemedicine were 2.5-fold more likely to build a rapport with their patients and 2.8-fold more likely to perceive that telemedicine increases the total number of patient consultations when compared to face-to-face consultations.

The COVID-19 pandemic provided sufficient incentive for the health care system to shift to web-based care to minimize the exposure to SARS-CoV-2 [19,33] and ultimately, as reported by Portnoy et al, “the only virus one can get while doing telemedicine is a cyber virus” [34,35]. The presence of these different modalities of telemedicine provided different opportunities for patients to connect with their health care providers, with rapid implementation of video and audio consultations partially owing to the availability of smartphones and the ubiquity of videoconferencing apps, since cameras are now an essential feature of these cellphones [36-42].

Although data on physician experience and outcome quality with each modality are limited, our first key finding suggests that when evaluating a patient for the first time or a patient with an acute condition, there is an added value in using videoconferencing apps to evaluate the patient’s general state of health, which is pivotal to the clinical decision-making process [43,44]. Because medical presentations can vary in

acuity and thus warrant different management approaches, physicians may need a real-time modality to assess the patient better, view the site of pathology, discuss treatment options, address the patient's concerns, and promote compliance with the treatment regimen. Video consultations can proximate real-life visits to a great extent as both the physician and patient can interact with each other simultaneously; this negates the psychological distance by allowing facial expressions and body language to be observed and interpreted, thus promoting empathic communication and the generation of a physician-patient rapport [45]. Therefore, a video consultation may be preferable when consulting a new patient for the first time as physicians would feel more confident in making diagnostic and treatment decisions. However, when evaluating follow-up patients with chronic diseases or for medication refill, video and audio telemedicine may be of equal quality and have similar outcomes as reported here and in previous studies [35,46-48]. These results may also help policymakers in low- and middle-income countries in applying reasonable protocols for selecting either video or audio consultations for patients who live in geographically remote areas or those who require frequent follow-up evaluation [49]. For instance, video consultations could be used for new or mild-to-moderate clinical presentations where real-time evaluation is needed, while audio consultations could be reserved for follow-up patients with chronic medical conditions or those with nonurgent medical problems who need to travel long distances and incur out-of-pocket costs [50]. In this course, a double triage system may be needed where a triage nurse consults with the patient who requests a telemedicine appointment and assess the patient's triage level using the Triage and Acuity Scale before recommending an in-person visit or video or audio consultation for the patient [51].

Our second key finding is that previous experience with telemedicine was associated with a lower perceived risk of misdiagnosis. In this respect, the more physicians were trained on telemedicine, the more confident they were in making a clinical diagnosis and the lesser the impression of a medical malpractice they had. Our results emphasize the need to increase telemedicine competencies in residency training and other clinical programs. For example, it is important to provide a formal education on best practices on how to remotely assess a patient's chief complaint and vital signs and carry out remote physical examination before placing physicians in web-based clinics, as prior experience with telemedicine can increase readiness and preparedness to carry out web-based consultations. This is intuitive specially for physicians who frequently use telemedicine, including those involved in internal medicine and family medicine [52]. Our findings are consistent with those of previous studies [53-55]. Ha et al reported that physicians who had a structured educational program in telemedicine had greater confidence in addressing clinical problems than those who did not receive an educational program [56]. Furthermore, Moore et al reported that the lack of telemedicine training was a barrier to provide telemedicine services among family medicine residents [52].

Our third key finding is that video consultations were associated with a perceived increase in overall consultation time, increased documentation time, and decreased total number of patient consultations. It is plausible that video consultations lasted longer owing to several reasons including technical difficulties related to internet connection, poor audio or speaker quality, disruption to the conversation flow, and difficulties with guiding a remote physical examination. In face-to-face interactions, people see and hear each other's words as they are produced; however, when using videoconferencing platforms, actions and words are heard milliseconds later. These delays, although small, are meaningful and can interfere with the conversation flow and result in miscommunication, thus consuming more time in an attempt to understand patients' problems and physicians' instructions [14,57,58]. Moreover, during video consultations, the physician may guide the patient through remote physical examination, which may increase the duration of the clinical consultation. Subsequently, the total number of daily patient consultations is expected to decrease owing to an increased duration of consultations in a limited clinical schedule.

Our fourth key finding is the identification of elements that represent barriers to telemedicine. A physician's inability to see the patient during the remote consultation could restrict tele-examination of the patient, where a guided remote assessment of the underlying condition is not feasible owing due to limited interaction with the web-based interface and the patient's difficulty to follow clinical instructions without physically seeing the provider's technique [59]. Moreover, the inability to see the patient during the clinical consultation could raise serious security and privacy issues, since the physician may not be able to confirm the patient's identity during the remote consultation, thus emphasizing the need of guidelines on identity management and security considerations to protect the patient's privacy during both audio and video consultations. Additionally, reimbursement issues with audio and video consultations need to be acknowledged, as it does not appear to attract health care providers preferentially for the delivery of telecare services. The current payment plans have been confusing, as the telemedicine provider needs to consider different private and governmental insurance policies when providing a remote consultation [60]. This confusion has been also a major deterrent to the use of telemedicine services. Furthermore, the relative difference in cost between telemedicine visits and a comparable face-to-face visit has been one of the barriers to the use of telemedicine. If a telemedicine visit is remunerated at a lower value than an equivalent face-to-face visit, physicians would be less willing to increase the provision of this service. There is a need to establish standardized regulations and billing rules to control costs. In principle, reimbursement costs for teleconsultation need to be equivalent to those of face-to-face visits to increase the adoption of telemedicine services [60]. A lack of training on how to treat a patient remotely may also be an obstacle that jeopardizes the efficiency of the virtual consultations, which must be overcome by incorporating appropriate training curricula, which can be incorporated through physician training programs.

Limitations and Strengths

This study has several limitations that we intend to address in future studies. First, this was an observational study that reflects outcomes with video and audio telemedicine consultations at a

single point in time. Second, data on what reimbursement challenges are associated with each modality was not captured in detail in this study, which might have biased physicians' attitudes toward each modality. Third, the perception of misdiagnosis was not defined in our survey; hence, it was challenging to understand the association between this outcome and predictors for physicians who used video or audio consultation. Fourth, in this study, patients' preferences for video or audio consultations were not captured and thus could have affected the number of clinical consultations for each modality and might have biased physicians' attitudes toward the mode of remote consultation.

Despite these limitations, our study has several strengths. To our knowledge, this study was one of the first comprehensive telemedicine studies in the Middle Eastern region, which had a nationally representative sample of physicians who used telemedicine and had a high survey response rate. Additionally, our study measured the difference in physicians' attitudes toward telemedicine by modality type, which is informative for policymaking decisions.

Conclusions

The experience with the COVID-19 pandemic has highlighted the important role of telemedicine in emergency responses. While we may not be able to precisely predict the exact diagnostic outcomes with each telemedicine modality, there is, however, a growing body of evidence that suggests that video consultations are associated with improved physician confidence in managing acute conditions and a greater ability to provide patient education during web-based consultations. This study demonstrates that when managing chronic conditions or follow-up patients remotely, audio consultation is as suitable as video consultation to health care providers. These findings may be helpful for health care policymakers in low-to-middle–income countries to provide ample health care access to patients with chronic and noncommunicable diseases. Previous experience with telemedicine was associated with improved physicians' confidence in case management, a lower perceived risk of misdiagnosis, an increased ability to provide patients with health education, and a better physician-patient rapport. Telemedicine services are likely to be retained, and as we build our telehealth system, it is intuitive to prioritize the "new normal" and implement a structured telemedicine curriculum in physician training programs and prepare them for web-based consultations. It is also necessary to acknowledge the barriers to telemedicine and create solutions and regulations to overcome these obstacles and increase the service adoption rate.

Acknowledgments

This study was supported by Khalifa University of Science and Technology (award# FSU-2020-33) and was endorsed by Abu Dhabi Public Health Center. We are grateful to Amina Asghar, clinical research counsellor at SEHA Corporate Academic Affairs, for assisting with the survey dissemination across Abu Dhabi SEHA hospitals. We are deeply thankful to Mandy Chen, senior administrator in medical instructional design at Khalifa University College of Medicine and Health Sciences, for developing the web-based surveys on Microsoft Forms. We also thank Ragheb Hasan Al-Nammari, College of Engineering at Khalifa University, for his efforts in data cleaning.

Authors' Contributions

NA and BA conceptualized and designed the study. NA, BA, NB, MAA, RA, SAM, FAH, YAZ, and HA carried out the investigation and data curation. MCES, NA, and OCB performed the formal analysis. NA, MCES, BA, M Alhashmi, M AlGhatrif, NB, MAA, RA, and OCB drafted the manuscript. NA, MCES, BA, M Alhashmi, M AlGhatrif, NB, MAA, RA, SAM, FAH, YAZ, HA, and OCB critically reviewed and revised the manuscript. NA and BA undertook the administrative tasks related to the study. OCB, SAM, FAH, and HA acquired the funding for the study. All authors have read and agreed to the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables with additional results.
[DOCX File , 52 KB - medinform_v9i6e29251_app1.docx]

References

1. Shirke MM, Shaikh SA, Harky A. Implications of Telemedicine in Oncology during the COVID-19 Pandemic. Acta Biomed 2020 Sep 07;91(3):e2020022 [FREE Full text] [doi: 10.23750/abm.v91i3.9849] [Medline: 32921719]
2. Metzl JM, Maybank A, De Maio F. Responding to the COVID-19 Pandemic: The Need for a Structurally Competent Health Care System. JAMA 2020 Jul 21;324(3):231-232. [doi: 10.1001/jama.2020.9289] [Medline: 32496531]
3. Miller IF, Becker AD, Grenfell BT, Metcalf CJE. Disease and healthcare burden of COVID-19 in the United States. Nat Med 2020 Aug;26(8):1212-1217. [doi: 10.1038/s41591-020-0952-y] [Medline: 32546823]

4. Mercier G, Arquizan C, Roubille F. Understanding the effects of COVID-19 on health care and systems. Lancet Public Health 2020 Oct;5(10):e524 [FREE Full text] [doi: 10.1016/S2468-2667(20)30213-9] [Medline: 33007210]
5. Hollander JE, Carr BG. Virtually Perfect? Telemedicine for Covid-19. N Engl J Med 2020 Apr 30;382(18):1679-1681. [doi: 10.1056/NEJMp2003539] [Medline: 32160451]
6. Frush K, Lee G, Wald S, Hawn M, Krna C, Holubar M, et al. Navigating the Covid-19 Pandemic by Caring for Our Health Care Workforce as They Care for Our Patients. NEJM Catalyst 2021 Jan;2(1). [doi: 10.1056/CAT.20.0378]
7. Perrone G, Zerbo S, Bilotta C, Malta G, Argo A. Telemedicine during Covid-19 pandemic: Advantage or critical issue? Med Leg J 2020 Jul;88(2):76-77. [doi: 10.1177/0025817220926926] [Medline: 32490720]
8. Robinson J, Borgo L, Fennell K. The Covid-19 Pandemic Accelerates the Transition to Virtual Care. NEJM Catalyst 2020;1(5) [FREE Full text]
9. Monaghesh E, Hajizadeh A. The role of telehealth during COVID-19 outbreak: a systematic review based on current evidence. BMC Public Health 2020 Aug 01;20(1):1193 [FREE Full text] [doi: 10.1186/s12889-020-09301-4] [Medline: 32738884]
10. Fischer SH, Ray KN, Mehrotra A, Bloom EL, Uscher-Pines L. Prevalence and Characteristics of Telehealth Utilization in the United States. JAMA Netw Open 2020 Oct 01;3(10):e2022302 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.22302] [Medline: 33104208]
11. Volterrani M, Sposato B. Remote monitoring and telemedicine. Eur Heart J Suppl 2019 Dec;21(Suppl M):M54-M56 [FREE Full text] [doi: 10.1093/eurheartj/suz266] [Medline: 31908618]
12. Callahan CW, Malone F, Estroff D, Person DA. Effectiveness of an Internet-based store-and-forward telemedicine system for pediatric subspecialty consultation. Arch Pediatr Adolesc Med 2005 Apr;159(4):389-393. [doi: 10.1001/archpedi.159.4.389] [Medline: 15809396]
13. Montgomery A, Hunter D, Blair E, Hendricksen M. Telemedicine Today: The State of Affairs. Altarum Institute. 2015 Mar. URL: https://altarum.org/sites/default/files/uploaded-publication-files/TELEMEDICINE-State%20of%20Affairs_030615.pdf [accessed 2021-05-21]
14. Shaw SE, Seuren LM, Wherton J, Cameron D, A'Court C, Vijayaraghavan S, et al. Video Consultations Between Patients and Clinicians in Diabetes, Cancer, and Heart Failure Services: Linguistic Ethnographic Study of Video-Mediated Interaction. J Med Internet Res 2020 May 11;22(5):e18378 [FREE Full text] [doi: 10.2196/18378] [Medline: 32391799]
15. Whitten P, Holtz B, Laplante C. Telemedicine: What have we learned? Appl Clin Inform 2010;1(2):132-141 [FREE Full text] [doi: 10.4338/ACI-2009-12-R-0020] [Medline: 23616832]
16. Ohannessian R, Duong TA, Odone A. Global Telemedicine Implementation and Integration Within Health Systems to Fight the COVID-19 Pandemic: A Call to Action. JMIR Public Health Surveill 2020 Apr 02;6(2):e18810 [FREE Full text] [doi: 10.2196/18810] [Medline: 32238336]
17. Smith AC, Thomas E, Snoswell CL, Haydon H, Mehrotra A, Clemensen J, et al. Telehealth for global emergencies: Implications for coronavirus disease 2019 (COVID-19). J Telemed Telecare 2020 Jun;26(5):309-313 [FREE Full text] [doi: 10.1177/1357633X20916567] [Medline: 32196391]
18. Webster P. Virtual health care in the era of COVID-19. Lancet 2020 Apr 11;395(10231):1180-1181 [FREE Full text] [doi: 10.1016/S0140-6736(20)30818-7] [Medline: 32278374]
19. Kichloo A, Albosta M, Dettloff K, Wani F, El-Amir Z, Singh J, et al. Telemedicine, the current COVID-19 pandemic and the future: a narrative review and perspectives moving forward in the USA. Fam Med Community Health 2020 Aug;8(3):e000530 [FREE Full text] [doi: 10.1136/fmch-2020-000530] [Medline: 32816942]
20. In efforts to curb COVID-19, The Department of Health – Abu Dhabi collaborates with Injazat on launching 'Remote Healthcare Platform'. Department of Health Abu Dhabi. 2020 Apr 13. URL: https://doh.gov.ae/en/news/DoH--Abu-Dhabi-collaborates-with-Injazat-on-launching-Remote-Healthcare-Platform [accessed 2021-03-14]
21. Loeb AE, Rao SS, Ficke JR, Morris CD, Riley LH, Levin AS. Departmental Experience and Lessons Learned With Accelerated Introduction of Telemedicine During the COVID-19 Crisis. J Am Acad Orthop Surg 2020 Jun 01;28(11):e469-e476 [FREE Full text] [doi: 10.5435/JAAOS-D-20-00380] [Medline: 32301818]
22. SEHA completes over 28,000 virtual consultations. SEHA. 2020 Apr 15. URL: https://www.seha.ae/seha-completes-over-28000-virtual-consultations/ [accessed 2021-03-14]
23. Bashshur RL, Reardon TG, Shannon GW. Telemedicine: a new health care delivery system. Annu Rev Public Health 2000;21:613-637. [doi: 10.1146/annurev.publhealth.21.1.613] [Medline: 10884967]
24. Ignatowicz A, Atherton H, Bernstein CJ, Bryce C, Court R, Sturt J, et al. Internet videoconferencing for patient-clinician consultations in long-term conditions: A review of reviews and applications in line with guidelines and recommendations. Digit Health 2019;5:2055207619845831 [FREE Full text] [doi: 10.1177/2055207619845831] [Medline: 31069105]
25. Langan S, Schmitt J, Coenraads P, Svensson A, von Elm E, Williams H, European Dermato-Epidemiology Network (EDEN). The reporting of observational research studies in dermatology journals: a literature-based study. Arch Dermatol 2010 May;146(5):534-541. [doi: 10.1001/archdermatol.2010.87] [Medline: 20479302]
26. Madden N, Emeruwa UN, Friedman AM, Aubey JJ, Aziz A, Baptiste CD, et al. Telehealth Uptake into Prenatal Care and Provider Attitudes during the COVID-19 Pandemic in New York City: A Quantitative and Qualitative Analysis. Am J Perinatol 2020 Aug;37(10):1005-1014 [FREE Full text] [doi: 10.1055/s-0040-1712939] [Medline: 32516816]

27. Vidal-Alaball J, Flores Mateo G, Garcia Domingo JL, Marín Gomez X, Sauch Valmaña G, Ruiz-Comellas A, et al. Validation of a Short Questionnaire to Assess Healthcare Professionals' Perceptions of Asynchronous Telemedicine Services: The Catalan Version of the Health Optimum Telemedicine Acceptance Questionnaire. Int J Environ Res Public Health 2020 Mar 25;17(7):2202 [FREE Full text] [doi: 10.3390/ijerph17072202] [Medline: 32218310]
28. Donelan K, Barreto E, Sossong S, Michael C, Estrada J, Cohen A, et al. Patient and clinician experiences with telehealth for patient follow-up care. Am J Manag Care 2019 Jan;25(1):40-44 [FREE Full text] [Medline: 30667610]
29. Hindmarch P, Hawkins A, McColl E, Hayes M, Majsak-Newman G, Ablewhite J, Keeping Children Safe study group. Recruitment and retention strategies and the examination of attrition bias in a randomised controlled trial in children's centres serving families in disadvantaged areas of England. Trials 2015 Mar 07;16:79 [FREE Full text] [doi: 10.1186/s13063-015-0578-4] [Medline: 25886131]
30. Al Tunaiji H, Al Qubaisi M, Dalkilinc M, Campos LA, Ugwuoke NV, Alefishat E, et al. Impact of COVID-19 Pandemic Burnout on Cardiovascular Risk in Healthcare Professionals Study Protocol: A Multicenter Exploratory Longitudinal Study. Front Med (Lausanne) 2020;7:571057 [FREE Full text] [doi: 10.3389/fmed.2020.571057] [Medline: 33415114]
31. Jayadev C, Mahendradas P, Vinekar A, Kemmanu V, Gupta R, Pradhan ZS, et al. Tele-consultations in the wake of COVID-19 - Suggested guidelines for clinical ophthalmology. Indian J Ophthalmol 2020 Jul;68(7):1316-1327 [FREE Full text] [doi: 10.4103/ijo.IJO_1509_20] [Medline: 32587157]
32. Likert R. A technique for the measurement of attitudes. Arch Psychol 1932;22:5-55 [FREE Full text]
33. Romanick-Schmiedl S, Raghu G. Telemedicine - maintaining quality during times of transition. Nat Rev Dis Primers 2020 Jun 01;6(1):45 [FREE Full text] [doi: 10.1038/s41572-020-0185-x] [Medline: 32483168]
34. Shachar C, Engel J, Elwyn G. Implications for Telehealth in a Postpandemic Future: Regulatory and Privacy Issues. JAMA 2020 Jun 16;323(23):2375-2376. [doi: 10.1001/jama.2020.7943] [Medline: 32421170]
35. Portnoy J, Waller M, Elliott T. Telemedicine in the Era of COVID-19. J Allergy Clin Immunol Pract 2020 May;8(5):1489-1491 [FREE Full text] [doi: 10.1016/j.jaip.2020.03.008] [Medline: 32220575]
36. Ashfaq A, Memon SF, Zehra A, Barry S, Jawed H, Akhtar M, et al. Knowledge and Attitude Regarding Telemedicine Among Doctors in Karachi. Cureus 2020 Feb 09;12(2):e6927 [FREE Full text] [doi: 10.7759/cureus.6927] [Medline: 32190480]
37. Zhang K, Liu W, Locatis C, Ackerman M. Mobile Videoconferencing Apps for Telemedicine. Telemed J E Health 2016 Jan;22(1):56-62 [FREE Full text] [doi: 10.1089/tmj.2015.0027] [Medline: 26204322]
38. Graham YNH, Hayes C, Mahawar KK, Small PK, Attala A, Seymour K, et al. Ascertaining the Place of Social Media and Technology for Bariatric Patient Support: What Do Allied Health Practitioners Think? Obes Surg 2017 Jul;27(7):1691-1696. [doi: 10.1007/s11695-016-2527-z] [Medline: 28054297]
39. Jedamzik S. Digital health and nursing: The future is now. Unfallchirurg 2019 Sep;122(9):670-675. [doi: 10.1007/s00113-019-0672-2] [Medline: 31143981]
40. Potdar R, Thomas A, DiMeglio M, Mohiuddin K, Djibo DA, Laudanski K, et al. Access to internet, smartphone usage, and acceptability of mobile health technology among cancer patients. Support Care Cancer 2020 Nov;28(11):5455-5461. [doi: 10.1007/s00520-020-05393-1] [Medline: 32166381]
41. Iyengar K, Upadhyaya GK, Vaishya R, Jain V. COVID-19 and applications of smartphone technology in the current pandemic. Diabetes Metab Syndr 2020;14(5):733-737 [FREE Full text] [doi: 10.1016/j.dsx.2020.05.033] [Medline: 32497963]
42. Majumder S, Deen MJ. Smartphone Sensors for Health Monitoring and Diagnosis. Sensors (Basel) 2019 May 09;19(9):2164 [FREE Full text] [doi: 10.3390/s19092164] [Medline: 31075985]
43. Mair F, McClusky C, Wilsgaard T, Wootton R. The added value of video for consultations in telemedicine for minor injuries work. J Telemed Telecare 2011;17(8):427-431. [doi: 10.1258/jtt.2011.110318] [Medline: 22036927]
44. McCullough GH, Rangarathnam B. Clinical Decision Making. Semin Speech Lang 2019 Jun;40(3):149-150. [doi: 10.1055/s-0039-1688996] [Medline: 31158899]
45. Zulman DM, Verghese A. Virtual Care, Telemedicine Visits, and Real Connection in the Era of COVID-19: Unforeseen Opportunity in the Face of Adversity. JAMA 2021 Feb 02;325(5):437-438. [doi: 10.1001/jama.2020.27304] [Medline: 33528520]
46. Flodgren G, Rachas A, Farmer AJ, Inzitari M, Shepperd S. Interactive telemedicine: effects on professional practice and health care outcomes. Cochrane Database Syst Rev 2015 Sep 07(9):CD002098 [FREE Full text] [doi: 10.1002/14651858.CD002098.pub2] [Medline: 26343551]
47. Portnoy JM, Waller M, De Lurgio S, Dinakar C. Telemedicine is as effective as in-person visits for patients with asthma. Ann Allergy Asthma Immunol 2016 Sep;117(3):241-245. [doi: 10.1016/j.anai.2016.07.012] [Medline: 27613456]
48. Hammersley V, Donaghy E, Parker R, McNeilly H, Atherton H, Bikker A, et al. Comparing the content and quality of video, telephone, and face-to-face consultations: a non-randomised, quasi-experimental, exploratory study in UK primary care. Br J Gen Pract 2019 Sep;69(686):e595-e604 [FREE Full text] [doi: 10.3399/bjgp19X704573] [Medline: 31262846]
49. Lewis T, Synowiec C, Lagomarsino G, Schweitzer J. E-health in low- and middle-income countries: findings from the Center for Health Market Innovations. Bull World Health Organ 2012 May 01;90(5):332-340 [FREE Full text] [doi: 10.2471/BLT.11.099820] [Medline: 22589566]

50. Hoffer-Hawlik MA, Moran AE, Burka D, Kaur P, Cai J, Frieden TR, et al. Leveraging Telemedicine for Chronic Disease Management in Low- and Middle-Income Countries During Covid-19. Glob Heart 2020 Sep 15;15(1):63 [FREE Full text] [doi: 10.5334/gh.852] [Medline: 33150128]
51. Lin C, Tseng W, Wu J, Tay J, Cheng M, Ong H, et al. A Double Triage and Telemedicine Protocol to Optimize Infection Control in an Emergency Department in Taiwan During the COVID-19 Pandemic: Retrospective Feasibility Study. J Med Internet Res 2020 Jun 23;22(6):e20586 [FREE Full text] [doi: 10.2196/20586] [Medline: 32544072]
52. Moore MA, Jetty A, Coffman M. Over Half of Family Medicine Residency Program Directors Report Use of Telehealth Services. Telemed J E Health 2019 Oct;25(10):933-939. [doi: 10.1089/tmj.2018.0134] [Medline: 30484746]
53. Wong R, Ng P, Spinnato T, Taub E, Kaushal A, Lerman M, et al. Expanding Telehealth Competencies in Primary Care: A Longitudinal Interdisciplinary Simulation to Train Internal Medicine Residents in Complex Patient Care. J Grad Med Educ 2020 Dec;12(6):745-752. [doi: 10.4300/JGME-D-20-00030.1] [Medline: 33391599]
54. Marttos AC, Fernandes Juca Moscardi M, Fiorelli RKA, Pust GD, Ginzburg E, Schulman CI, et al. Use of Telemedicine in Surgical Education: A Seven-Year Experience. Am Surg 2018 Aug 01;84(8):1252-1260. [Medline: 30185295]
55. Afshari M, Witek NP, Galifianakis NB. Education Research: An experiential outpatient teleneurology curriculum for residents. Neurology 2019 Jul 23;93(4):170-175. [doi: 10.1212/WNL.0000000000007848] [Medline: 31332085]
56. Ha E, Zwicky K, Yu G, Schechtman A. Developing a Telemedicine Curriculum for a Family Medicine Residency. PRiMER 2020;4:21 [FREE Full text] [doi: 10.22454/PRiMER.2020.126466] [Medline: 33111048]
57. Ruhleder K, Jordan B. Co-Constructing Non-Mutual Realities: Delay-Generated Trouble in Distributed Interaction. Computer Supported Cooperative Work 2001 Mar;10(1):113-138. [doi: 10.1023/A:1011243905593]
58. Isaacs E, Morris T, Rodriguez T. A comparison of face-to-face and distributed presentations. 1995 Presented at: SIGCHI Conference on Human Factors in Computing Systems; May 1995; Denver, CO. [doi: 10.1145/223904.223950]
59. Annaswamy TM, Verduzco-Gutierrez M, Frieden L. Telemedicine barriers and challenges for persons with disabilities: COVID-19 and beyond. Disabil Health J 2020 Oct;13(4):100973 [FREE Full text] [doi: 10.1016/j.dhjo.2020.100973] [Medline: 32703737]
60. Mehrotra A, Wang B, Snyder G. Telemedicine: What Should the Post-Pandemic Regulatory and Payment Landscape Look Like? The Commonwealth Fund. 2020 Aug 05. URL: https://www.commonwealthfund.org/publications/issue-briefs/2020/aug/telemedicine-post-pandemic-regulation [accessed 2021-05-21]

Abbreviations

OR: odds ratio
VIF: variance inflation factor

Edited by G Eysenbach; submitted 31.03.21; peer-reviewed by K Zhang, J Sturt; comments to author 22.04.21; revised version received 04.05.21; accepted 17.05.21; published 01.06.21.

Please cite as:
Alhajri N, Simsekler MCE, Alfalasi B, Alhashmi M, AlGhatrif M, Balalaa N, Al Ali M, Almaashari R, Al Memari S, Al Hosani F, Al Zaabi Y, Almazroui S, Alhashemi H, Baltatu OC
Physicians' Attitudes Toward Telemedicine Consultations During the COVID-19 Pandemic: Cross-sectional Study
JMIR Med Inform 2021;9(6):e29251
URL: https://medinform.jmir.org/2021/6/e29251
doi:10.2196/29251
PMID:34001497

Original Paper

Extraction of Traditional Chinese Medicine Entity: Design of a Novel Span-Level Named Entity Recognition Method With Distant Supervision

Qi Jia[1,2], PhD; Dezheng Zhang[1,2], PhD; Haifeng Xu[1,2], MA; Yonghong Xie[1,2], PhD

[1]School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China

[2]Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing, China

Corresponding Author:
Yonghong Xie, PhD
School of Computer and Communication Engineering
University of Science and Technology Beijing
30 Xueyuan Road, Haidian District
Beijing, 100083
China
Phone: 86 010 62334547
Email: xieyh@ustb.edu.cn

Abstract

Background: Traditional Chinese medicine (TCM) clinical records contain the symptoms of patients, diagnoses, and subsequent treatment of doctors. These records are important resources for research and analysis of TCM diagnosis knowledge. However, most of TCM clinical records are unstructured text. Therefore, a method to automatically extract medical entities from TCM clinical records is indispensable.

Objective: Training a medical entity extracting model needs a large number of annotated corpus. The cost of annotated corpus is very high and there is a lack of gold-standard data sets for supervised learning methods. Therefore, we utilized distantly supervised named entity recognition (NER) to respond to the challenge.

Methods: We propose a span-level distantly supervised NER approach to extract TCM medical entity. It utilizes the pretrained language model and a simple multilayer neural network as classifier to detect and classify entity. We also designed a negative sampling strategy for the span-level model. The strategy randomly selects negative samples in every epoch and filters the possible false-negative samples periodically. It reduces the bad influence from the false-negative samples.

Results: We compare our methods with other baseline methods to illustrate the effectiveness of our method on a gold-standard data set. The F1 score of our method is 77.34 and it remarkably outperforms the other baselines.

Conclusions: We developed a distantly supervised NER approach to extract medical entity from TCM clinical records. We estimated our approach on a TCM clinical record data set. Our experimental results indicate that the proposed approach achieves a better performance than other baselines.

(JMIR Med Inform 2021;9(6):e28219) doi:10.2196/28219

KEYWORDS

traditional Chinese medicine; named entity recognition; span level; distantly supervised

Introduction

Background

As a complementary medicine with thousands of years history, traditional Chinese medicine (TCM) has received increasing attention and even played an important role in the fight against COVID-19 in China. TCM clinical records contain the symptoms and signs of patient and the diagnosis process of the doctor as unstructured text. These records represent a large number of valuable academic thoughts and clinical experience of TCM experts.

With information technology being applied to TCM modernization, it is essential to discover TCM diagnosis pattern through data mining [1]. While these studies rely on structured data, TCM clinical records are unstructured text. Besides, TCM clinical records are mostly recorded in ancient Chinese. The

narrative is free style and difficult to understand for modern medical practitioners. The cost of manually structuring and maintaining free-text clinical records thus remains very expensive. Therefore, automatically extracting medical entities from TCM clinical records is an urgent need for research and analysis of TCM diagnosis knowledge.

So far, studies on medical entity extraction have mainly concentrated on modern medicine. The research on TCM medical entity extraction is still in early stages and faces more challenges. However, the text expression of TCM medical entity varies substantially and its boundaries are difficult to determine. For example, 牛黄解毒丸 (bezoar detoxicating tablet) is a prescription and includes a medicine 牛黄 (bezoar) in text. Because of these challenges, the cost of annotated corpus becomes very high; additionally, there is a lack of gold-standard data sets for supervised learning methods.

Distantly supervised named entity recognition (NER) is a good approach to deal with the situation where there is a lack of annotated corpus for training. It utilizes the domain entity dictionary and raw text to generate the silver-standard data set for training the NER model. In the TCM field, we can use existing knowledge resources to obtain the domain entity dictionary. However, the domain entity dictionary is always incomplete and cannot cover all entity names in practice. The diversity of TCM medical entities exacerbates this situation. In the silver-standard data set generated with distant supervision, each token that does not match the dictionary will be treated as a nonentity. As a result, it may introduce many false-negative samples and may have a bad influence on the performance of the NER model.

We therefore propose a span-level distantly supervised NER approach to extract TCM medical entity. Our key motivation is that although the entity mention could not be matched by the domain entity dictionary, the expression of the matched entity is correct. At this point, we treat the distantly supervised NER as a span detection task instead of a general sequence tagging task. We first design a simple classifier with a pretrained language model [2] to detect and type text span. It enumerates all possible text spans in a sentence as candidate entity mention and predicts the entity type of each text span independently. Compared with sequence tagging, the span-level method does not rely on the context token label in the sentence and reduces the influence of false-negative samples. The span-level entity extraction model needs to perform negative sampling as a nonentity that is not included in the domain entity dictionary. We then design a negative sampling strategy, which predicts the text span type periodically and evaluates the indeterminacy to filter the possible false-negative samples and reduce the influence on the model performance. We summarize the contribution as follows:

1. We propose a span-level distant supervised NER method to extract the TCM medical entity. It does not rely on the context token label in the sentence and mainly focuses on the feature of the entity span.
2. We also design a negative sampling strategy for our span-level entity extraction model. It filters the possible false-negative samples by measuring the indeterminacy of entity prediction to reduce the bad influence for the model training.
3. We estimate our approach on the TCM clinical record data set. Experimental results indicate that our approach achieves a better performance than other baselines.

Related Work

In recent years, medical entity extraction has become a very popular topic. Early studies in this area mainly focused on the language lexicon pattern and domain dictionary. However, with the rapid development of deep learning, current mainstream research uses deep neural networks to extract medical entities by tagging text sequence. Habibi et al [3] presented a completely generic method based on long short-term memory (LSTM) neural network and statistical word embeddings. The method demonstrated improved recall, and performed better than traditional NER tools, thus confirming the effectiveness of this method over others. Cho and Lee [4] designed a contextual LSTM network with conditional random fields (CRFs), and proved that this method had significantly improved performance on biomedical NER tasks. Li et al [5] proposed an NER method called Bi-LSTM-Att-CRF that integrates the attention mechanism with a bidirection LSTM (Bi-LSTM) neural network to extract Chinese electronic medical records. This method not only captured more useful contextual information, but also introduced medical dictionaries and part-of-speech features to improve the performance of the model. While using the attention-Bi-LSTM-CRF model, Ji et al [6] used the entity auto-correct algorithm to rectify entities based on historical entity information which further improved the performance of the model. Wu et al [7] used the bidirectional encoder representations from transformers (BERT) [2] pretrained language model as a encoder to generate token embedding and incorporated it with several common deep learning models (eg, Bi-LSTM and Bi-LSTM-CRF).

The TCM medical entity extraction has gradually attracted the attention of scholars worldwide. Wang et al [8] used CRF to extract symptom entities from free-text clinical records. Wang et al [9] investigated supervised methods and verified their effectiveness in extracting TCM clinical records and focused on the problems related to symptom entity recognition. Wang et al [10] proposed a supervised method for syndrome segmentation.

However, all these methods relied on high-quality annotation data. In practice, the cost of this gold-standard data set is very high, and therefore, many studies have begun to study the distantly supervised NER method. Ren et al [11] proposed a novel relation phrase-based clustering framework while investigating entity recognition with distant supervision. Their framework uses some linguistic features such as part-of-speech. To use a small amount of labeled data for aspect term extraction, Giannakopoulos et al [12] introduced an architecture that achieves top-ranking performance for supervised aspect term extraction. Shang et al [13] designed a novel and effective neural model (AutoNER) with a new Tie or Break scheme. Through experiments, they proved the effectiveness of AutoNER when only using dictionaries with no additional human effort. Peng et al [14] proposed a novel positive-unlabeled (PU) learning

algorithm to perform NER. The performance of this method was very dependent on the settings of important hyperparameters. Zhang et al [15] followed the corresponding annotation guidelines for clinical records of Chinese medicine and constructed a fine-grained entity annotation corpus. Zhang et al [16] proposed a novel back-labeling approach and integrated it into a tagging scheme, which improved the effectiveness and robustness of distantly supervised methods. These studies were, however, very much dependent on some a priori assumptions and external nature language process toolkits to denoise distantly supervised data set, and limit the task to sequence tagging.

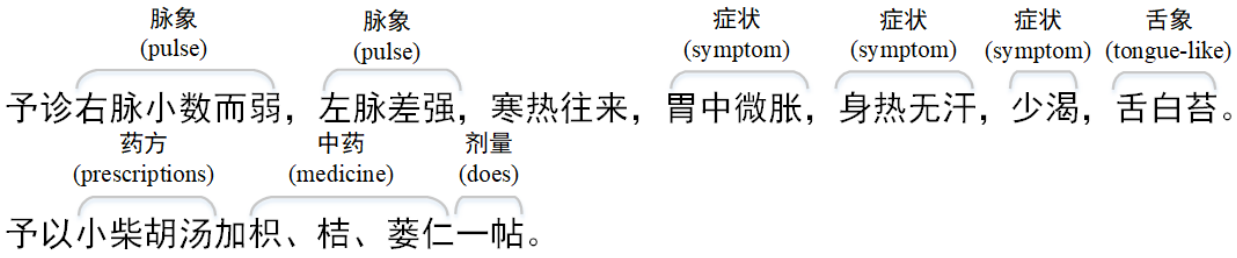
Methods

Data

Distantly supervised NER needs a domain entity dictionary and raw text as the basic data. Therefore, first, we define the TCM medical entity type including 症状 (symptom), 脉象 (pulse), 舌象 (tongue like), 药方 (prescriptions), 中药 (medicine), and 剂量 (dose). We then obtain the domain dictionary from the TCM knowledge graph [17]. The dictionary includes 18,688 symptom entities, 34,594 Chinese medicine entities, 39,640 prescriptions entities, 304 dose entities, 4915 tongue entities, and 5800 pulse entities. We used a book entitled 《中华历代名医医案》 [18] as the raw text. It was compiled by the expert team of Professor Lu Zhaolin, Beijing University of Chinese Medicine, and has been published by the Beijing Science and Technology Publishing House. The book has collected more than 18,000 TCM cases and contains more than 8 million words. Each case introduces the patient's illness involving the symptoms, pulse, and tongue like, and the process of seeking medical treatment. It also introduces the doctor's diagnosis of the patient's condition and a description of how to treat along with the medical prescription (in Chinese) and the corresponding dose of the prescribed medicine. Figure 1 shows an extracted sample record.

We used the maximum matching algorithm [19] to label back the medical entity. Specifically, there are conflicts in the dictionary such as 牛黄解毒丸 (bezoar detoxicating tablet) and 牛黄 (bezoar). For this case, we selected the longest text to match. We filtered out the sentences whose length was less than 15 tokens and did not match their entity in dictionary to generate the silver-standard data set.

Figure 1. A TCM clinical record extraction example.



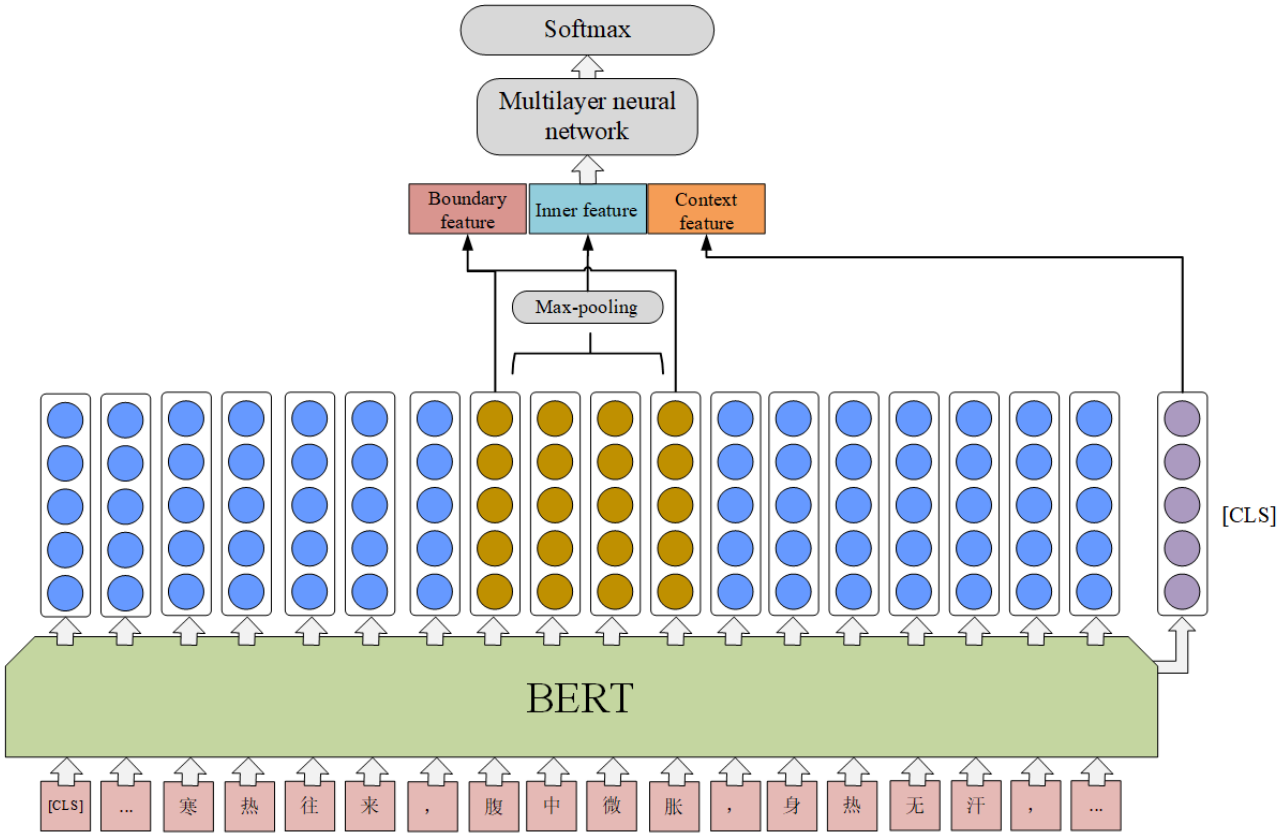
Span-Level NER Model

Overview

In this section, we explicate our span-level NER model in detail. Instead of the general sequence tagging model to extract name entity, the proposed model treats the task as a text-span classification and takes an arbitrary continuous text span as candidate input. For a given sentence $s = [t_1, t_2, \ldots, t_N]$ of n token, there are $n(n + 1)/2$ possible text span. A text span is defined as span = $[t_i, \ldots, t_{i+k}]$, where $1 < i < N$ and $k \geq 0$.

We designed a simple classifier to detect the entity type of text span (Figure 2). It utilizes BERT pretrained language model as text feature encoder and obtains the token embedding representation of text span. The BERT transforms the input token t_i to an embedding vector e_i with a length of 768. The representation of text span is a 2D tensor $[e_i, \ldots, e_{i+k}]$, where k is the length of span. The BERT pretrained language model keeps fine-tuning during training. Then, we model span representation as described in the following subsections.

Figure 2. The span-level named entity extraction model. BERT: bidirectional encoder representations from transformers.



Span Inner Feature

We combine the token embedding of the text span with max-pooling to represent the inner feature of span.

$$R_{inner}(span)=maxpooling([e_i, \ldots, e_{i+k}]) \quad (1)$$

Span Boundary Feature

For TCM medical entity, prefixes and suffixes have strong indications for the type of entity. We concatenate the head and tail token embedding as the boundary representation of span.

$$R_{boundary}(span)=[e_i;e_{i+k}] \quad (2)$$

We concatenate the span inner feature and the span boundary feature. In addition, we concatenate the representation of [CLS] in the BERT as the global sentence representation. The final span presentation is as follows:

$$R_{span}=[R_{inner};R_{boundary};CLS] \quad (3)$$

We feed the span representation to a multilayer neural network (2 layers in our model) and a softmax classifier which yields a posterior for each entity type.

$$R^s=f_{multi}(W \cdot R_{span}+b) \quad (4)$$

 (5)

Negative Sample During Training

The most important problem in distantly supervised NER is the false-negative samples. During the training phase, the proposed span-level method needs to select nonentity text span as negative samples. We thus designed a negative sampling strategy on the silver-standard data set. Instead of using all the possible negative samples for training, the strategy randomly selects a number of negative samples in each epoch. It reduces the bad influence in the training phase from false-negative samples through label smoothing. Meanwhile, the model predicts the silver data set for several epochs periodically. We measure the indeterminacy of the prediction results by information entropy. According to the indeterminacy, we design a negative sample filter mechanism, which filters the possible false-negative samples in the next training period.

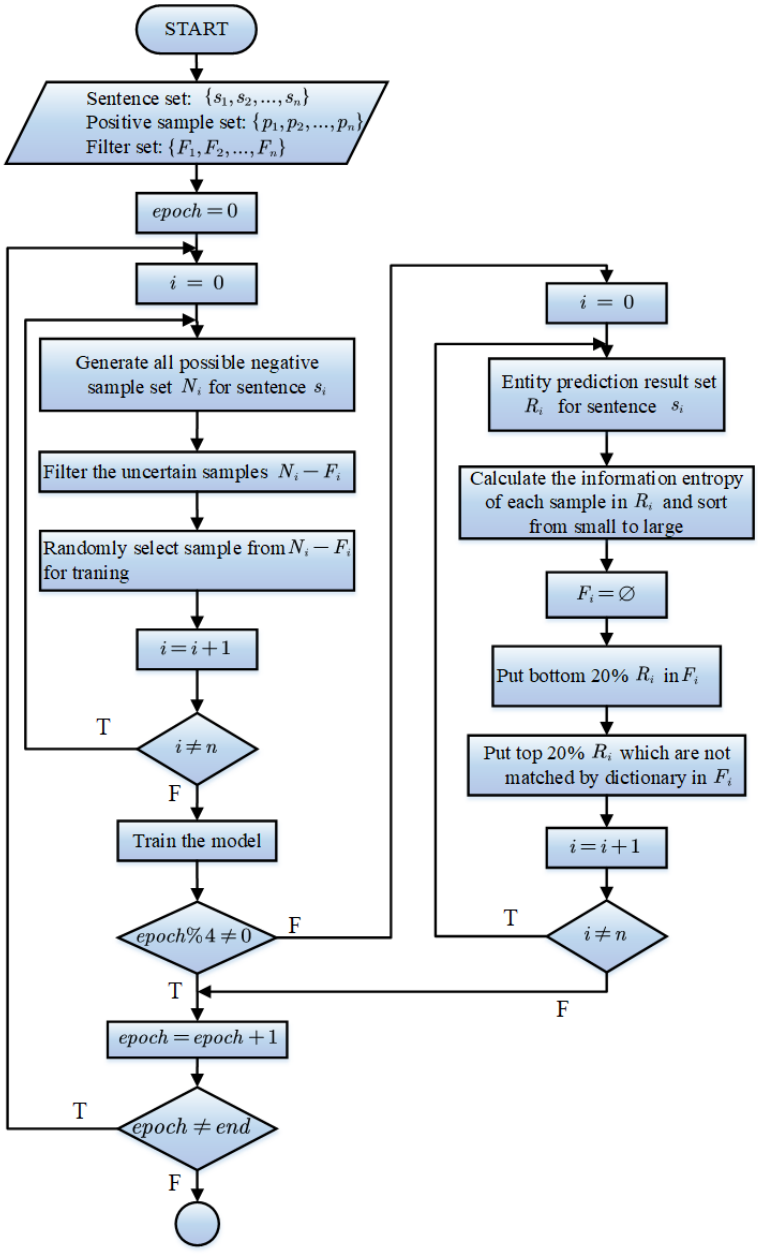
In each epoch, we randomly select the nonentity text span of the silver data set as negative samples. Because there may be entities in these negative samples, we use label smoothing to assign probability of entity types to negative samples:

where K is the number of class and ε is a hyperparameter. During the training, we predict the data set for several epochs periodically. For each sentence s_i in the data set, we predict the entity-type probability of each text span to obtain prediction result R_i. We measure the indeterminacy with information entropy. The greater the information entropy, the greater the indeterminacy of the prediction of the text span.

Then we sort R_i (ie, the prediction result from small to large) according to the indeterminacy and maintain a corresponding negative-sample filtering set F_i for each sentence s_i. We put the bottom 20% sample and the top 20% entity sample, which are

not matched by the dictionary of R_i, in F_i. The samples in F_i are possible false-negative samples. These possible false-negative samples influence the model performance, and so we filter these samples in the model training phase. In the next training period, the samples in F_i would not be selected in negative sampling. The flow diagram of the strategy is shown in Figure 3.

Figure 3. The flow diagram of the negative sampling strategy.



The purpose of this strategy is to avoid the influence of possible false-negative samples on the model. Filtering these samples out helps reduce the influence of incorrect labeling types on training.

Results

Data Set

The silver data set contains 203,485 tokens (10,083 sentences). In order to verify the effectiveness of our approach, we manually annotated 5000 sentences as the test set. It includes about 100,000 tokens with label. We also use the dictionary to label back the entity in the test set. The entity distribution is shown in Table 1. In particular, both the test set and the silver data set are from the same book, and there is no cross between them.

Table 1. The entity distribution on the test set in the dictionary and manual way.

Entity type	Dictionary	Manual
Symptom	5031	6362
Medicine	4854	8187
Prescriptions	450	545
Dose	6078	9276
Tongue-like	322	631
Pulse	668	972

Experiment Set

During the process of model training, we used AdamW as the optimizer, and set the learning rate to 0.001 and the learning rate decay to 0.95. For the negative sampling strategy, we set the label smoothing ε to 0.85, the training epoch to 20, and the period to 4, and randomly select 100 negative samples for a sentence in an epoch. In this case, the filter set of negative sampling will update 4 times in training. According to the max text length in the domain entity dictionary, we limit the max length of span to 12 tokens.

The standard precision (P), recall (R), and F1 score (F1) on the test set are used as evaluation metrics. We compare the following baseline method to illustrate the effectiveness of our method.

Distant-LSTM-CRF [12] introduced domain-term mining and linguistic features such as dependency tree to generate the silver-standard data set. It used the sequence-tagging LSTM-CRF method to recognize entity by training on the silver-standard data set.

AdaPU [14] treated the distantly supervised NER as a positive unlabeled problem. This method generated a silver-standard data set with the maximum matching algorithm and trained LSTM-CRF models separately for each entity type. It designed a loss function depending on 2 hyperparameters: the ratio of entity words and the class weight for the positive class.

The novel label-back AutoNER [16] combined a domain term dictionary generated by AutoPhrase [20]. It designed a label-back strategy according to some prior assumptions to generate the silver-standard data set. The model masked the nonentity term during training to skip the parameter update.

BERT-CRF [2] is a popular supervised method. It is a sequence-tagging method and utilizes the BERT pretrained language model as encoder and CRF as sequence decoder. We used the silver data set in accordance with the proposed method as the training data set.

Evaluation

The performance on the test set of the different methods is presented in Table 2. According to the results, the F1 score of our method is 77.34 and it remarkably outperforms the best baseline (novel label-back AutoNER) by an improvement of 8.16 in the F1 score. This indicates the effectiveness of the proposed method.

Compared with other baseline methods, our method shows substantial improvement in recall and makes a balance between the precision and recall. As a supervised method, BERT-CRF has a better performance than some distantly supervised methods on the silver-standard data set. This illustrates the effectiveness and robustness of the pretrained language model.

Table 2. Experiment results on the test set with comparison.

Method	Precision	Recall	F1 score
Distantly LSTM[a]-CRF[b]	74.03	31.59	53.93
AdaPU	70.15	60.87	65.18
Novel label-back AutoNER	73.06	66.75	69.18
BERT[c]-CRF	75.62	58.73	66.15
Our method	78.28	76.52	77.34

[a]LSTM: long short-term memory.

[b]CRF: conditional random field.

[c]BERT: bidirectional encoder representations from transformers.

Discussion

Principal Findings

In this section, we discuss the influence of negative sampling strategy on performance and the hyperparameter setting. We analyzed the effect of negative sampling strategy through an ablation study. Steps involved in the ablation study are as follows: (1) the random negative sampling is maintained, but the false-negative sample filter is removed; (2) the bottom sampling in the prediction result R_i is removed; (3) the top

sampling in the prediction result R_i is removed; and (4) the label smoothing is removed. The result is presented in Table 3.

Based on the result, the F1 score is reduced by 4.60 without the false-negative sample filter. The false-negative sample filter mechanism avoids the influence of error samples on the model. It proves the validity of the false-negative sample filter mechanism. We also discuss the filter range. The bottom filter affects the performance more than the top filter, which illustrates that samples with larger indeterminacy influence the performance more.

Meanwhile, we also analyzed the influence of the span feature with the ablation study. The span representation includes the inner feature and the boundary feature. The result is shown in Table 4. Both the inner feature and the boundary feature will impact the model performance. In comparison with the boundary feature, the inner feature shows more obvious impact.

Moreover, we discuss the hyperparameters of the negative sampling including the number of random negative samples and the ratio of the false-negative sample filter. The number of random negative samples is set as 50, 100, 150, and 200, and the ratio of false-negative sample filter is set as 10%, 15%, 20%, and 25%. The result is presented in Tables 5 and 6. The obtained result indicates that the hyperparameters need to be set to appropriate values. We also notice that the ratio of the false-negative sample filter has a greater influence on the performance, and we consider this phenomenon to be caused by the coverage of the domain entity dictionary.

However, our study still has some limitations and could be improved. During the process of training and prediction, the method needs to enumerate all possible text spans in the sentence. This step affects the efficiency of the method. We consider introducing a toolkit such as word segmentation to improve the efficiency; otherwise we only consider the nonentity sampling and ignore the possible entity. For a specific domain, an entity name in dictionary has strong uniqueness and is not prone to ambiguity. However, in the open domain, the ambiguity for an entity name is common. Our method did not consider the false-positive samples because of the ambiguity. We intend to introduce some sampling strategies (eg, AdaSampling) and some self-supervised methods to solve the problem in future work.

Table 3. Experimental result without the false-negative sample filter.

Method	Precision	Recall	F1 score
Without the false-negative sample filter	75.15	70.48	72.74
Without bottom	75.93	73.25	74.57
Without top	76.86	75.75	76.30

Table 4. Experimental result without the false-negative sample filter.

Method	Precision	Recall	F1 score
Inner feature only	77.03	75.71	76.36
Boundary feature only	76.12	74.92	75.52

Table 5. Experimental results on different numbers of random negative samples.

Number of random negative samples	Precision	Recall	F1 score
50	77.24	75.17	76.19
100	78.28	76.52	77.34
150	78.42	75.25	76.80
200	79.56	72.91	76.09

Table 6. Experimental results on different ratios of the false-negative sample filter.

Number of random negative samples	Precision	Recall	F1 score
10%	76.25	71.34	73.71
15%	77.73	75.84	76.77
20%	78.28	76.52	77.34
25%	75.04.	75.28	75.15

Conclusions

In this paper, we illustrated a distantly supervised NER approach to extract medical entity from TCM clinical records. Different from general sequence tagging, we propose a span-level model to detect and classify entity. It utilizes the pretrained language model as the text feature extractor, and constructs the span representation containing inner, boundary, and context feature.

The model uses a multilayer neural network and softmax as classifier for the span representation. We designed a negative sampling strategy for the span-level model. The strategy randomly selects negative samples in every epoch and filters the possible false-negative sample periodically. We evaluated the effectiveness of our method by comparing it with other baselines. Meanwhile, we also discussed the influence of different parts of negative sampling strategy on performance. In the future, we intend to extend our method to a wider range of fields and study its generalization. We also will optimize the negative sampling strategy to improve the ability to filter the false-negative samples.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (Grant No. 2017YFB1002304).

Authors' Contributions

QJ lead the method application and performed experiments, result analysis, and wrote the manuscript. HX performed data preprocessing and wrote the manuscript. DZ performed manuscript revision. YX provided theoretical guidance and revised this paper.

Conflicts of Interest

None declared.

References

1. Gu P, Chen H. Modern bioinformatics meets traditional Chinese medicine. Brief Bioinform 2014 Nov 24;15(6):984-1003. [doi: 10.1093/bib/bbt063] [Medline: 24067932]
2. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics; 2019 Presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics; June 2-7, 2019; Minneapolis, MN. [doi: 10.18653/v1/N19-1423]
3. Habibi M, Weber L, Neves M, Wiegandt DL, Leser U. Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics 2017 Jul 15;33(14):i37-i48. [doi: 10.1093/bioinformatics/btx228] [Medline: 28881963]
4. Cho H, Lee H. Biomedical named entity recognition using deep neural networks with contextual information. BMC Bioinformatics 2019 Dec 27;20(1):735 [FREE Full text] [doi: 10.1186/s12859-019-3321-4] [Medline: 31881938]
5. Li L, Zhao J, Hou L, Zhai Y, Shi J, Cui F. An attention-based deep learning model for clinical named entity recognition of Chinese electronic medical records. BMC Med Inform Decis Mak 2019 Dec 05;19(Suppl 5):235 [FREE Full text] [doi: 10.1186/s12911-019-0933-6] [Medline: 31801540]
6. Ji B, Liu R, Li S, Yu J, Wu Q, Tan Y, et al. A hybrid approach for named entity recognition in Chinese electronic medical record. BMC Med Inform Decis Mak 2019 Apr 09;19(Suppl 2):64 [FREE Full text] [doi: 10.1186/s12911-019-0767-2] [Medline: 30961597]
7. Wu J, Shao D, Guo J, Cheng Y, Huang G. Character-based deep learning approaches for clinical named entity recognition: a comparative study using Chinese EHR texts. Cham, Switzerland: Springer; 2019 Presented at: International Conference on Smart Health; July 1-2, 2019; Shenzhen, China. [doi: 10.1007/978-3-030-34482-5_28]
8. Wang Y, Liu Y, Yu Z, Chen L, Jiang Y. A preliminary work on symptom name recognition from free-text clinical records of traditional Chinese medicine using conditional random fields and reasonable features. Stroudsburg, PA: ACL; 2012 Presented at: BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing; June 8, 2012; Montreal, Canada.
9. Wang Y, Yu Z, Chen L, Chen Y, Liu Y, Hu X, et al. Supervised methods for symptom name recognition in free-text clinical records of traditional Chinese medicine: an empirical study. J Biomed Inform 2014 Feb;47:91-104 [FREE Full text] [doi: 10.1016/j.jbi.2013.09.008] [Medline: 24070769]
10. Wang Y, Tang D, Shu H, Su C. An Empirical Investigation on Fine-Grained Syndrome Segmentation in TCM by Learning a CRF from a Noisy Labeled Data. JAIT 2018;9(2):45-50. [doi: 10.12720/jait.9.2.45-50]
11. Ren X, El-Kishky A, Wang C, Tao F, Voss CR, Ji H, et al. ClusType: Effective Entity Recognition and Typing by Relation Phrase-Based Clustering. In: KDD. New York, NY: ACM; 2015 Aug Presented at: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 10-13, 2015; Sydney, Australia p. 995-1004 URL: http://europepmc.org/abstract/MED/26705503 [doi: 10.1145/2783258.2783362]
12. Giannakopoulos A, Musat C, Hossmann A, Baeriswyl M. Stroudsburg, PA: ACL; 2017 Sep 8 Presented at: Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis; September 8, 2017; Copenhagen, Denmark. [doi: 10.18653/v1/w17-5224]

13. Shang J, Liu L, Gu X, Ren X, Ren T, Han J. Learning named entity tagger using domain-specific dictionary. Stroudsburg, PA: ACL; 2018 Presented at: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Nov ; Brussels, Belgium. ACL; November 2-4, 2018; Brussels, Belgium. [doi: 10.18653/v1/d18-1230]
14. Peng M, Xing X, Zhang Q, Fu J, Huang X. Distantly supervised named entity recognition using positive-unlabeled learning. Stroudsburg, PA: ACL; 2019 Presented at: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; July 28 to August 2, 2019; Florence, Italy. [doi: 10.18653/v1/P19-1231]
15. Zhang T, Wang Y, Wang X, Yang Y, Ye Y. Constructing fine-grained entity recognition corpora based on clinical records of traditional Chinese medicine. BMC Med Inform Decis Mak 2020 Apr 06;20(1):64 [FREE Full text] [doi: 10.1186/s12911-020-1079-2] [Medline: 32252745]
16. Zhang D, Xia C, Xu C, Jia Q, Yang S, Luo X, et al. Improving Distantly-Supervised Named Entity Recognition for Traditional Chinese Medicine Text via a Novel Back-Labeling Approach. IEEE Access 2020;8:145413-145421. [doi: 10.1109/access.2020.3015056]
17. Zhang D, Xie Y, Li M, Shi C. Construction of knowledge graph of traditional chinese medicine based on the ontology. Information Engineering 2017;3(1):35-42. [doi: 10.3772/j.issn.2095-915x.2017.01.004]
18. Zhaolin L. Medical Records of Famous Chinese Doctors in the Past (1st edition). Beijing, China: Beijing Science and Technology Publishing House; 2015.
19. Xue N. Chinese word segmentation as character tagging. International Journal of Computational Linguistics & Chinese Language Processing 2003;8(1):29-58. [doi: 10.3115/1119250.1119278]
20. Shang J, Liu J, Jiang M, Ren X, Voss CR, Han J. Automated Phrase Mining from Massive Text Corpora. IEEE Trans. Knowl. Data Eng 2018 Oct 1;30(10):1825-1837. [doi: 10.1109/tkde.2018.2812203]

Abbreviations

CRF: conditional random fields
LSTM: long short-term memory
NER: named entity recognition
POS: part-of-speech
TCM: traditional Chinese medicine

Edited by T Hao; submitted 25.02.21; peer-reviewed by B Hu, G Zhou; comments to author 15.03.21; revised version received 14.04.21; accepted 19.04.21; published 14.06.21.

Please cite as:
Jia Q, Zhang D, Xu H, Xie Y
Extraction of Traditional Chinese Medicine Entity: Design of a Novel Span-Level Named Entity Recognition Method With Distant Supervision
JMIR Med Inform 2021;9(6):e28219
URL: https://medinform.jmir.org/2021/6/e28219
doi:10.2196/28219
PMID:34125076

Original Paper

A Novel Metric to Quantify the Effect of Pathway Enrichment Evaluation With Respect to Biomedical Text-Mined Terms: Development and Feasibility Study

Xuan Qin[1], MD; Xinzhi Yao[1], BA; Jingbo Xia[1], PhD

Hubei Key Lab of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan, China

Corresponding Author:
Jingbo Xia, PhD
Hubei Key Lab of Agricultural Bioinformatics
College of Informatics
Huazhong Agricultural University
1#, Lion Rock Street, Hongshan District
Hubei Province
Wuhan, 430070
China
Phone: 86 02787288509
Email: xiajingbo.math@gmail.com

Abstract

Background: Natural language processing has long been applied in various applications for biomedical knowledge inference and discovery. Enrichment analysis based on named entity recognition is a classic application for inferring enriched associations in terms of specific biomedical entities such as gene, chemical, and mutation.

Objective: The aim of this study was to investigate the effect of pathway enrichment evaluation with respect to biomedical text-mining results and to develop a novel metric to quantify the effect.

Methods: Four biomedical text mining methods were selected to represent natural language processing methods on drug-related gene mining. Subsequently, a pathway enrichment experiment was performed by using the mined genes, and a series of inverse pathway frequency (IPF) metrics was proposed accordingly to evaluate the effect of pathway enrichment. Thereafter, 7 IPF metrics and traditional *P* value metrics were compared in simulation experiments to test the robustness of the proposed metrics.

Results: IPF metrics were evaluated in a case study of rapamycin-related gene set. By applying the best IPF metrics in a pathway enrichment simulation test, a novel discovery of drug efficacy of rapamycin for breast cancer was replicated from the data chosen prior to the year 2000. Our findings show the effectiveness of the best IPF metric in support of knowledge discovery in new drug use. Further, the mechanism underlying the drug-disease association was visualized by Cytoscape.

Conclusions: The results of this study suggest the effectiveness of the proposed IPF metrics in pathway enrichment evaluation as well as its application in drug use discovery.

(JMIR Med Inform 2021;9(6):e28247) doi:10.2196/28247

KEYWORDS

pathway enrichment; metric; evaluation; text mining

Introduction

The rising health issues worldwide and outbreaks of drug resistance have drawn a great amount of attention to new drug development [1]. However, drug development is expensive and time-consuming, and an average of US $800 million [2] to US $1.8 billion [3] and more than 10 years is invested in the development of 1 drug [4]. Improving the efficiency of drug discovery has long been one of the most important research directions and goals of medical research. As per the data in the 2018 edition of the World Health Organization's International Classification of Diseases and related health problems, there are 31,055 diseases [5]. Direct drug-disease pairing validation will have 85,214,920 drug-disease treatment validations. This highlights the importance of understanding the mechanisms of disease pathology and the action mechanisms of the existing drugs. According to the data released by the US National Food and Drug Administration in 2018, 35,283 types of drugs and

2744 types of effective ingredients have been approved [6]. Therefore, drug repositioning is recommended as a low-cost drug discovery method based on the clinical use of the drug, by which new indications of the marketed drug are discovered and an old drug is repurposed [4,7]. The linking of drugs to diseases via enriched gene sets is the basis of the drug use strategy under pathway enrichment analysis, which has long been an investigative way to unveil the functional interpretation of known gene sets [8,9]. The enrichment analysis mainly relies on the evaluation of the overexpressed gene set in a specific pathway, thereby leading to functional interpretation [10]. Technically, for a given disease or drug, relevant pathway information is publicly available in pathway databases [11]. For humans, the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [12] contains 38,680 *Homo sapiens* genes, and the abundance of data makes the correlation of disease-related genes or drug-related genes possible. In addition, there are multiple ways to identify a relevant gene set for a given disease [13]. While genome-wide association studies [14] or mRNA analysis [15] is the typical method for drug-related knowledge discovery, biomedical natural language processing is an alternative [16]. However, evaluating pathway enrichment in terms of a chosen gene set exclusively generated by a text mining system is still an unsolved issue [17]. The text mining system extracts the drug-related genes from drug-related literature, and pathway enrichment is then subsequently performed upon the text-mined genes. Although it is believed that text mining takes advantage of the abundant information from text resources [18], the diversity rooted from the various text mining systems leads to diversified results and effects in subsequent pathway enrichment. As representatives of the text mining system, PubTator [19] in a co-occurrence manner and the Turku Event Extraction System (TEES) [20] in a more semantic and syntactic manner play an important role in the biomedical named entity recognition and pathway enrichment.

The framework of this study was as follows. First, we used various biomedical text mining strategies to investigate the drug-related gene sets. Second, we designed novel metrics for pathway enrichment of text-mined genes. Here, 7 novel inverse pathway frequency (IPF) metrics were proposed and they were compared with the traditional *P* values. Finally, we performed a case study to show the effectiveness of the IPF metrics in pathway enrichment as well as the promising application of the text mining pipeline for new drug use discovery.

Methods

Collection of Rapamycin-Centric Resources

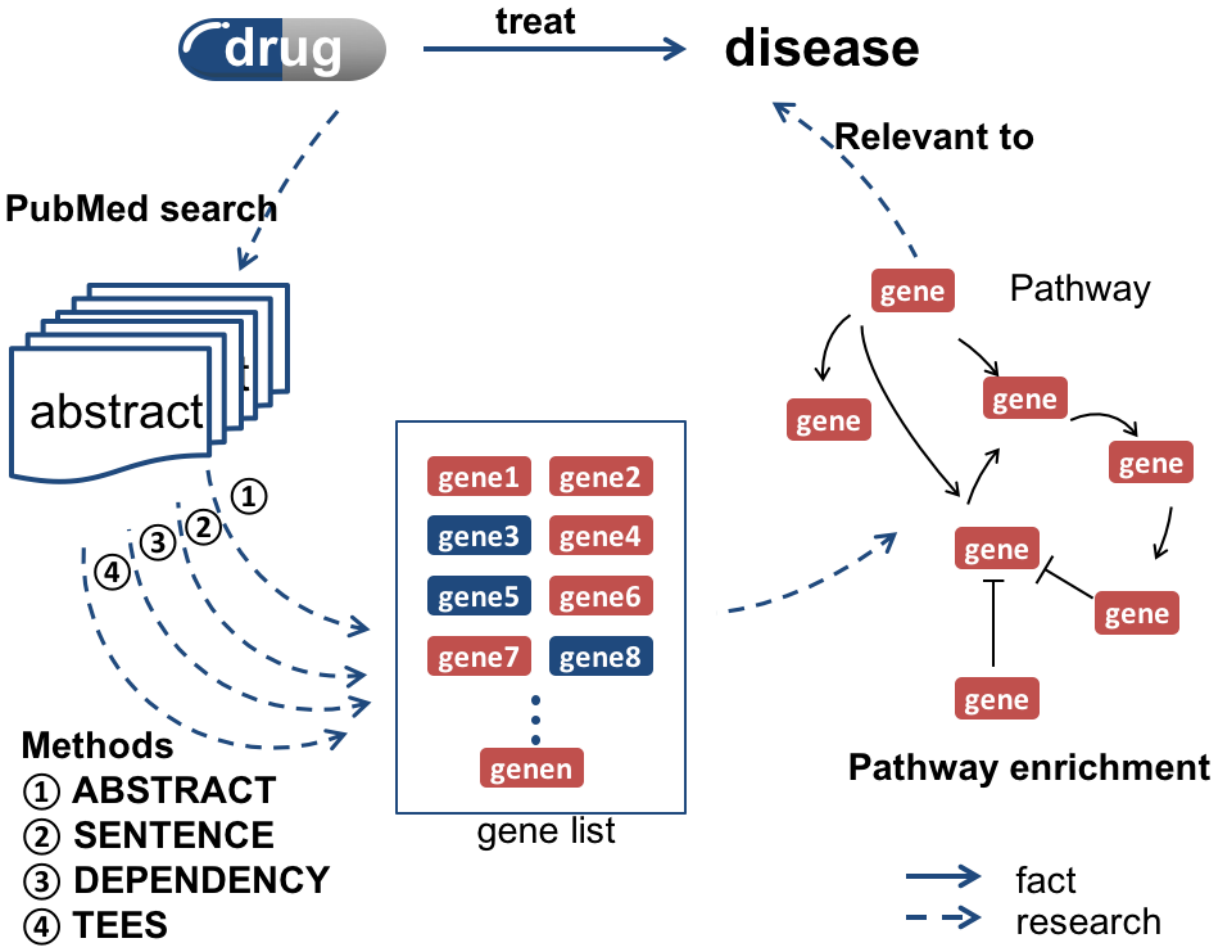
In this paradigm, a drug-centric text resource was obtained to extract the related genes. We set the drug as rapamycin, also known as sirolimus, as the target drug, which is used for the treatment of renal cell carcinoma and malignant lymphoma. Relevant texts and pathway data were collected targeting rapamycin as follows:

1. Text resources: 31,118 abstracts reporting rapamycin were downloaded from PubMed.
2. Rapamycin-related pathway data set: The drug pathway was retrieved from the comparative toxicogenomic database (CTD) [21], in which the KEGG pathway is enriched significantly among genes that interact with the drug or its downstream entity with a significant *P* value. In total, there are 166 pathways that are related to rapamycin.

Pathway Enrichment Evaluation in Terms of Text-Mined Genes

As shown in Figure 1, 4 text mining methods were applied to extract the gene pairs in rapamycin-related PubMed texts. They were (1) Method 1: *ABSTRACT* (co-occurrence in abstract) [19], (2) Method 2: *SENTENCE* (co-occurrence in sentence), (3) Method 3: *DEPENDENCY* (under consideration of dependency tree structure) [22], and (4) Method 4: *TEES* (Turku Event Extraction System) [20]. By taking co-occurrence or relation from the above methods, genes were linked to form an undirected pathway. We then proposed 7 types of novel pathway enrichment metrics by introducing various weights to the mined genes. Since the genes were extracted from 4 types of text mining systems, metrics evaluation was compared with respect to different text mining systems. For a given gene set, the candidate pathway is derived from 329 pathways in KEGG. Therefore, the sorted pathways based on *P* values in KEGG enrichment are regarded as the ground truth of pathway enrichment without using the text-mined knowledge. Furthermore, the feasibility of the text mining system for drug mechanism prediction was investigated.

Figure 1. Text mining systems for gene extraction and pathway construction. TEES: Turku Event Extraction System.

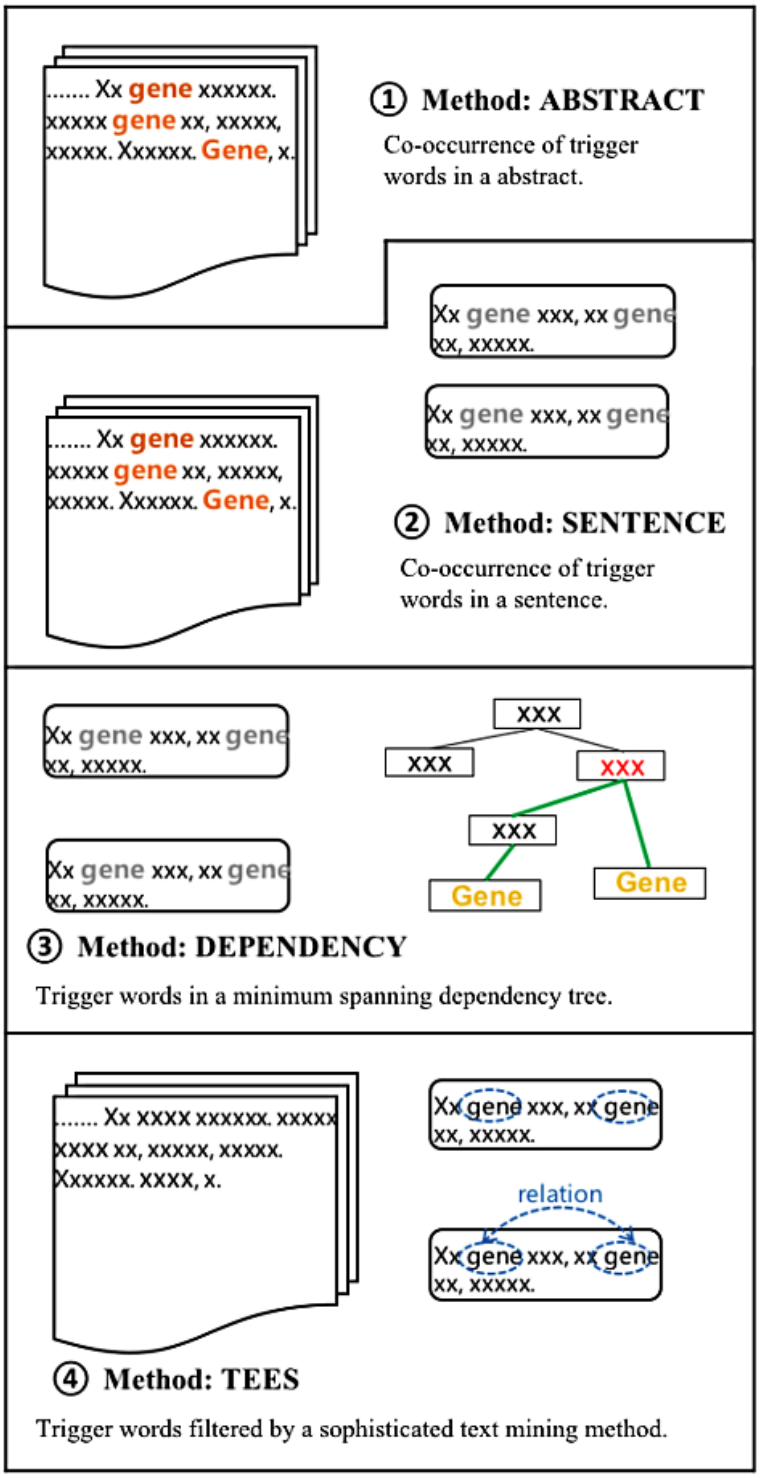


State-of-the-art Text Mining Methods

To extract gene pairs from the abstracts of papers, PubTator [19] and TEES [23] were selected as the 2 baseline text mining tools, which contribute to the following 4 text mining systems (Figure 2):

1. Method 1: *ABSTRACT.* Only abstracts containing the specific drug name were collected. If more than 2 genes showed up in one collected sentence, these genes were extracted and regarded as drug-related genes.
2. Method 2: *SENTENCE.* Similar to the abstract-level extraction rule, gene pairs were extracted based on a sentence co-occurrence rule.
3. Method 3: *DEPENDENCY.* Being stricter than sentence-level gene-pair extraction, the syntactic rule was introduced to restrict the co-occurrence filtering rule. Here, the Stanford parser was used to identity the gene subject or the gene object in a sentence. The gene pair is maintained only when the 2 genes act as sub or obj in the syntactic tree.
4. Method 4: *TEES.* TEES [20] is a sophisticated biomedical relation extraction system, which has been trained over 400,000 linguistics features. TEES is used to extract the genes that have interactions with other genes in drug-related abstracts. Thus, the TEES method provides a set of genes, which shows interaction information in drug-related abstracts.

Figure 2. Gene pair extraction rule for the text mining systems.



Traditional Metrics for Pathway Enrichment

Based on the drug-related abstract text file, 1 text mining tool extracts 1 group of genes. This group of genes is considered to be associated with the drug. For the sake of new drug use discovery, a group of drug-related genes is obtained using a text mining tool. Meanwhile, in the KEGG database, 1 pathway contains a group of genes, which are related to the disease the pathway correlates with. Thus, the matching degree between the drug-related gene group and the disease-related pathway represents the potential of the matching degree between the drug and the disease. ClusterProfile [24] is a known pathway enrichment tool, which applies the *P* value setting for the significance test of the relevant pathway for a given gene set.

Assuming in total that there are *N* background genes related to a specific pathway and there is a given gene set with *k* genes, pathway enrichment is performed to evaluate the significance for the given gene set to be relevant to the specific pathway. The significance value is obtained via chance computation for the given gene set in comparison to a randomly sampled gene set. In random sampling, *k* genes are sampled and *x* out of *k*

genes are related to the pathway. Then, the probability for this instance is as follows:

The *P* value used to address the significance of the pathway for the gene set is as follows:

The *P* value as a traditional enrichment metric reflects solid statistical concern in terms of chance computation. It relies on the hypothesis that the chance for each gene belonging to a given gene set is equal. However, this prerequisite is in some cases not met, for example, housekeeping genes have higher chances to appear in any given pathway, while on the contrary, certain specific genes only appear in a specific pathway.

Proposed Metrics for Pathway Enrichment

IPF for a Gene in a Given Pathway

The 4 text mining methods extracted 4 different sets of drug-related genes. Through these gene-drug relations, a bridge between the genes and the drug was established. The aim of this study was to investigate how a drug is associated with its indication through the gene. The next part was to establish the bridge between these genes and the indication. Mature gene-disease relations were easily accessed through KEGG in the form of the KEGG pathway. The KEGG pathway is a collection of manually drawn pathway maps representing the knowledge on the molecular interaction, reactions, and relation network. Thus, a bridge between the genes and the drug was established via KEGG. The whole path in that mechanism was addressed by finding a gene bridge between the drug and its indication. The next step was to evaluate this strategy. We paid attention to which text mining method is more suitable in this strategy. We focused on the drug-related gene set extracted by the text mining method in terms of the quantity and importance. Thus, we needed to define the importance standard of the gene to the indication. The standard of the gene to the indication in this case is based on the KEGG pathway information. One gene specifically shows up in a specific pathway, which means that this pathway can be identified with this gene. In other words, the less pathways a gene appears in, the more important it is to its related pathway. To calculate this situation, we give a value IPF.

Where $P=\{p_1,p_2.....p_M\}$ refers to all KEGG pathways, where $M=\#\{P\}$ is the number of pathways in the KEGG database.

$\{p_m|gene_i \in p_m\}$ refers to a pathway that contains the *i*-th gene, denoted as $gene_i$. Thus, every gene in the KEGG database receives a basic score. Simply adding all the gene scores together is unfair. Because all pathways show up in KEGG in the form of a map, each map consists of a set of node boxes and severe edges instead of genes and edges. Therefore, we need to figure out how to calculate that score that one text mining method receives from all node boxes in a specific pathway.

Enrichment of Text Mining–Based Gene Sets in a Pathway in View of a Gene

Assume T_t is a gene set that contains all of the genes mined by the *t*-th text mining method. In order to evaluate how T_t genes are enriched in a specific pathway, P_m, we define

Where $IPF\_gene_{T_t,P_m}$ considers the number of genes that exist in a pathway as well as the weight of each gene. The sum of the IPFs can be used to evaluate the association of the group of genes to a pathway. By doing this, cumulative associations along with gene weights are represented.

Enrichment of Text Mining–Based Gene Sets in a Pathway in View of a Node

In KEGG, a node box in some cases represents 1 set of homologous genes, instead of 1 separate gene. Generally, although there exists more than 1 gene, these genes play the same role. Therefore, even the text mining method digs more than one gene belonging to this pathway but they play the same role in the same node box. We only applied the max gene score to represent the score that this text mining method receives in this node box in this pathway. If $node_j$ is a single node,

where

If $node_j$ has E subnodes,

Where $g_i \in \{N_{node_j} \cap T_t\}$, $g_i = g_{max}$

For each $gene_i$, which belongs to gene set $node_j$ as well as T_t, the maximum IPF$gene_i$ is assigned, which means $gene_i$ belongs to gene set N_{node_j}.

It is noted that a node box sometimes represents 1 set of protein complex genes that need to work together to play a role in the pathway. Therefore, we applied the sum of all the gene scores that the text mining method received in this node and multiplied it with a coefficient to represent the score that this text mining method receives in this node box in this pathway.

where $|N_{node_j}|$ means the gene number of gene set N_{node_j}, while $|g_i \in \{node_j \cap T_t\}|$ means the gene number of the union of gene set N_{node_j} and gene set T_t.

Enrichment of Text Mining-Based Gene Sets in a Pathway in View of the Shortest Path

Besides the inclusion of genes in 1 node, the graph theory of the node in the pathway should be taken into consideration. In

graph theory, the degree of a vertex is the number of edges associated with the vertex. In a pathway graph, one node holding a high degree indicates that this node connects with more vertices. In term of gene, this gene is associated with many genes. Mutations and regulation of the gene affect more genes. In 1 pathway, the more a node shows up in the shortest path between the 2 genes, the more important this gene is in this pathway.

First, assume $SP_{node_r,node_s}$ refers to the shortest path between 2 arbitrary nodes, that is, $node_r$ and $node_s$ in pathway P_m, then, we count the occurrence of $node_j$ in $SP_{node_r,node_s}$ with respect to P_m.

$$Count_{node_j,P_m} = \#\{SP_{node_r,node_s} | node_j \in SP_{node_r,node_s}\} \quad \textbf{(9)}$$

In addition, $NShortPath_{node_r,node_j,node_k}$ is a binary value, which denotes whether or not $node_j$ appears in the shortest path between $node_r$ and $node_k$.

Thus, each node in the pathway holds a "count" score. To compare the importance of a node among all the nodes in one pathway, softmax function is applied to $NShortPath_{node_r,node_j,node_k}$. Here, the softmax function is the gradient logarithmic normalization of the discrete probability distribution of finite terms. The result of softmax is suitable for describing the importance of 1 node in 1 pathway.

Then, we added all IPF_{node_j} to represent the total score that the text mining method receives in this pathway,

where

Based on the above discussion on *IPF_gene* (equation 4), *IPF_node* (equation 8), and *IPF_shortpath* (equation 11), we formulate a generalized formula for $IPF\_node_{T_t,P_m}$.

Here, equation (13) summarizes all the above metric considerations and proposes a generalized form of IPF metrics. For instance, *IPF_gene* in equation (4) holds if 1 is assigned to $Weight_{node_j,P_m}$. Equation (12) is assigned to $Score_{T_t,node_j}$ Score(T_t, $node_j$) and equation (3) to $Weight_{gene_i}$. The full list of generalized IPF metrics is shown in Table 1.

Table 1. The complete inverse pathway frequency metrics list.

Inverse pathway frequency (IPF) metrics	$Weight_{node_j,P_m}$	$Score_{T_t,node_j}$	$Weight_{gene_i}$
IPF_gene	1	Equation (12)	Equation (3)
IPF_node	1	Equations (5) and (7)	1
IPF_shortpath	Equation (10)	Equation (12)	1
IPF_shortpath_gene	Equation (10)	Equation (12)	Equation (3)
IPF_shortpath_node	Equation (10)	Equations (5) and (7)	1
IPF_gene_node	1	Equations (5) and (7)	Equation (3)
IPF_gene_node_shortpath	Equations (5), (7), and (10)	Equations (5) and (7)	Equation (3)

Results

IPF Metric Comparison Under the Evaluation of Relevance Gene Ranking

We evaluated the effectiveness of IPF metrics by observing the rank counts of topic-related genes obtained from the 4 text mining methods. First, the 4 baseline text mining methods, that is, ABSTRACT, SENTENCE, DEPENDENCY, and TEES, were used to filter the vital genes in rapamycin-related texts. Afterwards, for each gene set obtained by the various text mining methods, 7 IPF metrics and traditional *P* values were used to map to obtain vital pathways and their pathway ranks. We then evaluated the pathway ranks by counting the occurrences of the key CTD pathways depicted in the Methods section. As shown in Figure 3 and Table 2, the x-axis refers to the rank of the enriched pathways and the y-axis refers to the cumulative percentage (CumPer), which is the ratio of the vital CTD pathway among the top i-th enriched pathways.

Figure 3. Comparison of the pathway-enrichment metrics based on the rapamycin-related gene set. CumPer: cumulative percentage; IPF: inverse pathway frequency; TEES: Turku Event Extraction System.

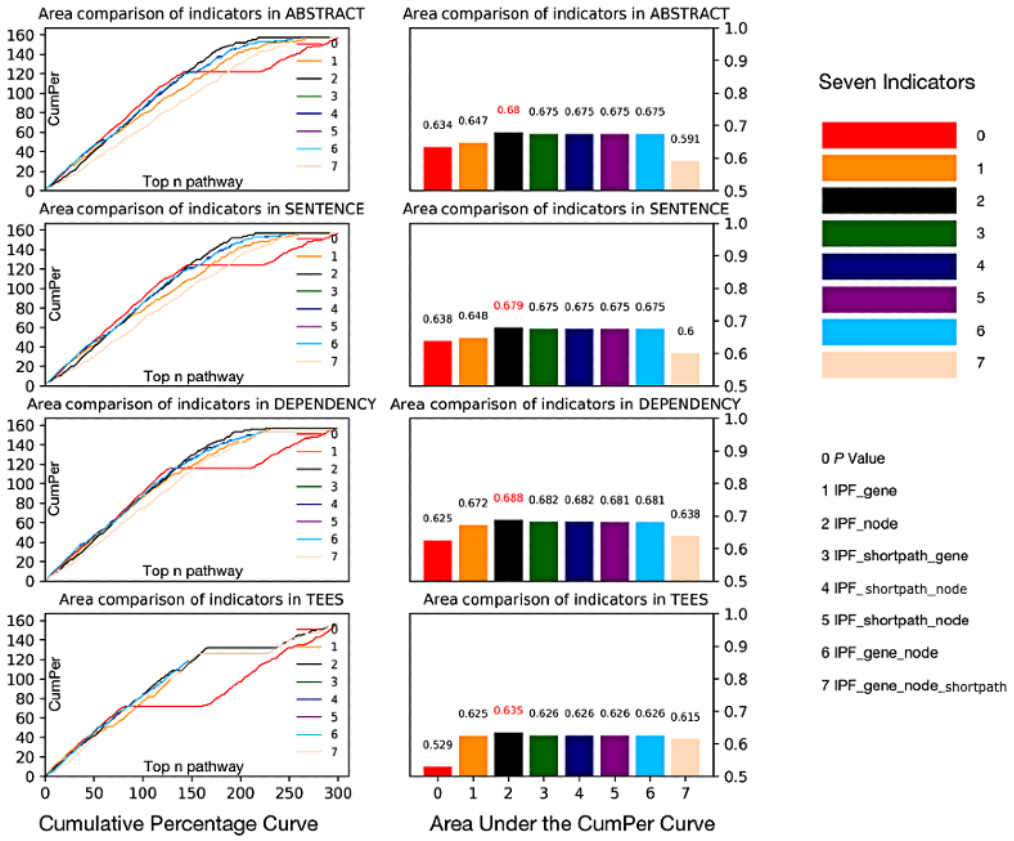


Table 2. Comparison of the areas under the cumulative percentage curve for the pathway-enriched methods based on the known rapamycin-related pathway.

Inverse pathway frequency metrics	ABSTRACT	SENTENCE	DEPENDENCY	Turku Event Extraction System
IPF_gene	0.634	0.638	0.628	0.529
IPF_node	0.647	0.648	0.672	0.625
IPF_shortpath	0.680[a]	0.679[a]	0.688[a]	0.635[a]
IPF_shortpath_gene	0.675	0.675	0.682	0.626
IPF_shortpath_node	0.675	0.675	0.682	0.626
IPF_gene_node	0.675	0.675	0.682	0.626
IPF_gene_node_shortpath	0.675	0.675	0.681	0.626
P value	.59	.60	.64	.62

[a]Indicates that the area is significantly superior to this text mining method in terms of the pathway enrichment indicator.

The bars from 0 to 8 in the bar plot represent the *P* value and 7 IPF metrics in Table 1, respectively. The results show that genes ranked with *P* values map to less vital pathways than genes from IPF metrics. In detail, the cumulative percentage curves of *P* values are given in the left 4 plots, and it is straightforward to observe that the *y* obtained by the *P* value grades the lowest in all the text mining cases. If computing the area under the cumulative percentage curve, the areas are 0.634, 0.638, 0.625, and 0.529 for *P* values for each case, which are as well the least in all cases. In all, the consistency of the poor performance of the *P* value positively shows the effectiveness of the IPF metric in support of the key pathway enrichment. Furthermore, in all the 7 IPF metrics, the black bar, which represents *IPF_node*, performs the best with the highest value of area under the cumulative percentage curve. It achieves 0.68, 0.679, 0.688, and 0.635 in *ABSTRACT*, *SENTENCE*, *DEPENDENCY*, and *TEES*, respectively.

Artificial Intelligence in Pathway Enrichment

Although the area values among IPF metrics do not differ substantially from each other, the *IPF_node* prevails over the rest of all in a consistent manner. The results show that the *IPF_node* represents the best semantic feature from the view of the natural language processing method and it is the most supportive for vital pathway enrichment.

Replication in the Discovery of Efficacy of Rapamycin for Breast Cancer

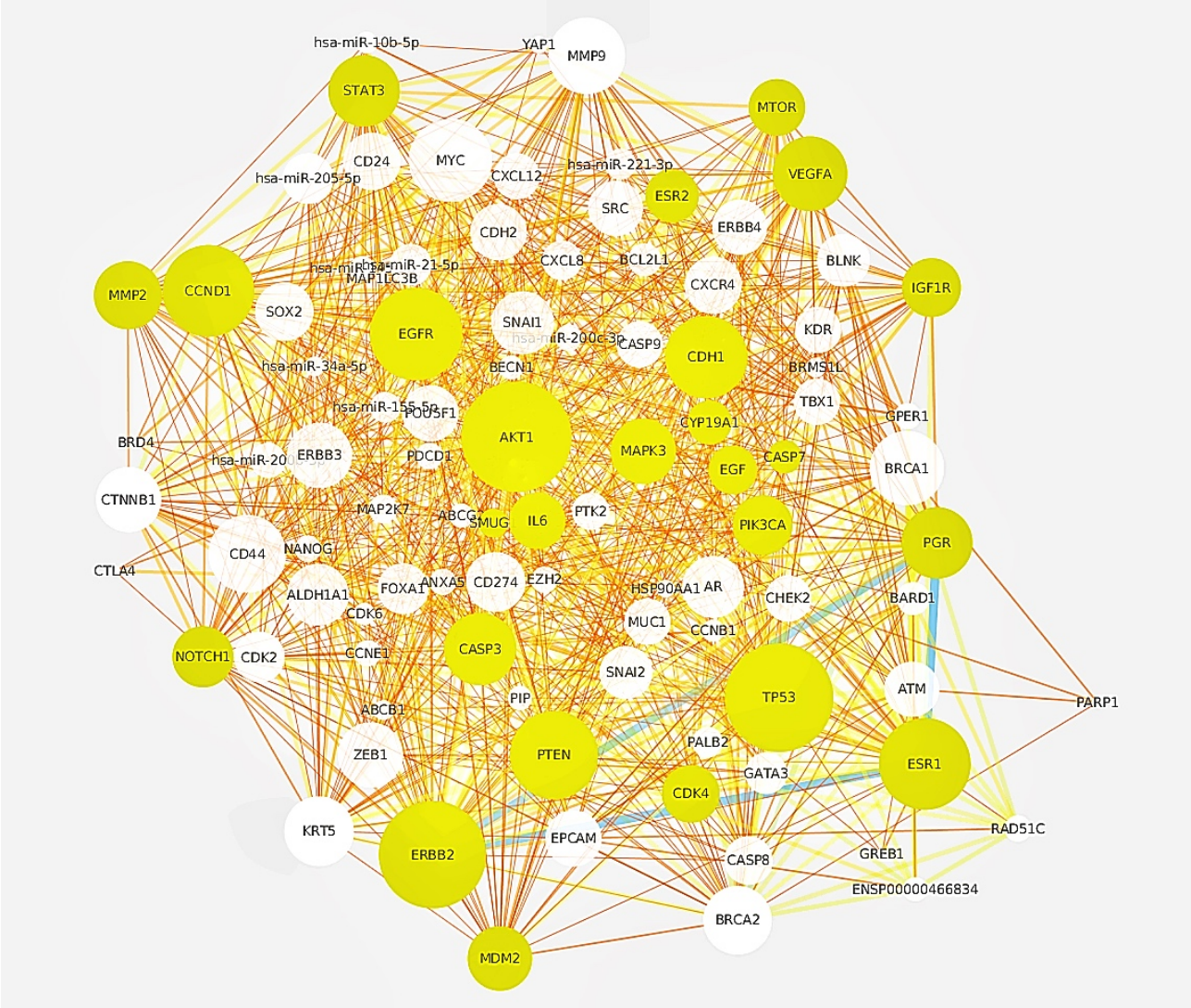
The discovery of the efficacy of rapamycin was replicated via a pathway enrichment experiment. PubReminer was used to retrieve the research trend of rapamycin and breast cancer drugs. A total of 1502 abstracts were obtained, and the starting time was the year 2000. The experiment was designed to test if the gene interaction of rapamycin could be excavated by the text mining method from literature without reporting the relevance of breast cancer and rapamycin. All the gene pairs in the literature related to rapamycin from the years 1978 to 2000 were excavated, the active genes of rapamycin were obtained, and the enrichment analysis of the strategic gene pathway in this study was carried out. After applying the *IPF_node,* 1640 abstracts of rapamycin prior to the year 2000 were obtained and 243 genes were obtained. Afterwards, a standard pathway enrichment was obtained, and the top 0.5% of the pathways under each enrichment path index was statistically analyzed. As expected, the breast cancer pathway was listed in the enrichment results, and the results indicated that the potential activity of rapamycin can be obtained by enriching the gene pathway by text mining interaction genes.

Visualization of the Pharmaceutical Mechanism

The text mining system was investigated to bridge the drug, protein, and disease pathway in order to explore the pharmaceutical mechanism of rapamycin. Starting from the Literature Network application, the disease-related gene network was constructed, and 480 genes obtained by rapamycin-centric text mining were used to highlight the overlapping parts in the breast cancer gene network. All the breast cancer–related genes were collected from the STRING database. According to all the existing databases and text information, each gene was sorted for rapamycin correlation, and in this verification section, 100 breast cancer–related genes from STRING were selected. The breast cancer gene network was constructed according to the gene interaction mentioned in more than 40,000 papers, and the network was constructed using the literature network application program. After gene pathway enrichment analysis, the drug was associated with the pathway and Cytoscape was used for network visualization. In view of the relation between the pathway information and the disease, the drug was further associated with the disease. In order to further analyze the relationship between drugs and diseases, the distribution of the drug-active genes excavated in the disease gene network was analyzed.

In order to construct a disease-specific gene network, the genetic relationship of this network in nature was obtained from disease-related abstracts. Since Cytoscape is a high-quality visualization platform for network analysis, a literature network application program based on Cytoscape was applied to address the drug disease associations obtained after pathway enrichment. Figure 4 highlights 38 vital genes plotted as yellow circles, namely, *STAT3*, *TP53*, *CDK4*, *CTLA4*, *AR*, *MYC*, *NOTCH1*, *IL6*, *ERBB2*, *CXCL12*, *BECN1*, *IGF1R*, *CDK2*, *EGF*, *ERBB4*, *MMP9*, *PIK3CA*, *CXCL8*, *ABCB1*, *EZH2*, *CDK6*, *SOX2*, *AKT1*, *CDH1*, *SRC*, *MTOR*, *ABCG2*, *KDR*, *CCND1*, *VEGFA*, *EGFR*, *ZEB1*, *ATM*, *PTEN*, *CXCR4*, *ERBB3*, *MDM2*, and *GATA3*. These 38 genes are based on the intersection of the breast cancer text network and the drug rapamycin-active gene obtained in this strategy. The size of the point in the graph represents the degree of the point, the greater the degree, the larger the point, and the degree in this network is the number of proteins that interact with the protein. The edge thickness in the figure represents the number of sentences that support the protein-protein relationship. The edge color in the figure also represents the number of sentences that support the protein-protein relationship. It can be seen from the figure that the yellow bright spot covers the vast majority of breast cancer gene networks with moderately large spots. The 38 genes were enriched by the *P* value pathway, and 16 of them, that is, *EGFR*, *IL6*, *TP53*, *CDK6*, *CDK4*, *PTEN*, *CDK2*, *KDR*, *AKT1*, *IGF1R*, *CCND1*, *VEGFA*, *PIK3CA*, *MDM2*, *MTOR*, and the *MYC* signaling pathway belong to one of the *MTOR* signaling pathways. Among them, *MTOR* is an important gene targeted by rapamycin. The *MTOR* pathway plays an important role in multiple activities of rapamycin and is therefore linked to breast cancer. The reason that literature network is used to construct breast cancer–related network is that the protein interaction involved in constructing the network is obtained from the literature related to breast cancer, and it is the programmed realization of protein interaction based on sentence coexpression in this study. It is convenient for users to quickly construct interactive protein interaction networks based on text relationships. In this study, the breast cancer–related genes obtained from the STRING database were rearranged according to the text information, and the protein interaction information excavated from the text was reflected in the size of the protein gene points. Thus, breast cancer genes were given different weights. It is more convenient to give priority to the location of the active genes under the active conditions defined by the interaction. The overlap of disease and drug-active genes was observed and the possible mechanism of action was speculated.

Figure 4. Visualization of the extracted gene pairs from literature.



Discussion

In this study, all text resources were obtained from a rapamycin-centric literature data set prior to the year 2000, and all predicted drug efficacies for rapamycin were based on knowledge ahead of this timeline. Therefore, it was interesting to "replicate" and evaluate a novel pathway-discovery method in our case study and to investigate the research paradigm based on pathway enrichment. Several studies after the year 2000 provide evidences to show that the mined rapamycin-centric pathway make sense. For example, after Liu et al [25] reported the effect of rapamycin in effectively inhibiting the growth of breast cancer in preclinical and clinical trials, the mechanism of action of rapamycin was elucidated. Rapamycin controls the growth, metabolism, and senescence of cells, as well as cells' reactions to nutrients, energy levels, and growth factors. *MTOR*, the target of rapamycin, is often upregulated in a variety of cancers, while rapamycin is extremely selective in blocking *MTOR*. Interestingly, our case study pinpointed *MTOR* correctly and made our pathway enrichment method conceivable in the study of breast cancer. Hopefully, the investigation of rapamycin action in the treatment of breast cancer will be propelled by further extensive and abundant text mining results in the future.

In conclusion, this research proposed a group of new pathway enrichment metrics by combining protein-interaction mechanisms, graph theories, information retrieval, and data mining weighting technology and by providing a new idea on pathway enrichment analysis. Moreover, the effectiveness of the best new enrichment metric for rapamycin was analyzed and the new activity of the drug shown by our method is supported by evidence from the literature. This research strategy sheds light on the investigation of the mechanism of action of drugs on diseases by using text-mined genes that are enriched in signaling pathways.

Acknowledgments

The authors would like to express their gratitude to Prof Lars Juhl Jensen, Dr Marc Legeay, and Ms Yuxing Wang for many valuable discussions. Data and codes are available in https://github.com/RuringQinXuan/PathwayEnrichmentMetric.

Authors' Contributions

XQ was responsible for the coding, performed the whole text mining experiments, implemented the IPF metric evaluation, and drafted the manuscript. JX formulated the whole mathematical analysis, performed the TEES experiments, and modified the manuscript. XY performed PubMed term extraction.

Conflicts of Interest

None declared.

References

1. Wright JA, Lewis WH, Parfett CLJ. Somatic cell genetics: a review of drug resistance, lectin resistance and gene transfer in mammalian cells in culture. Can J Genet Cytol 1980;22(4):443-496. [doi: 10.1139/g80-056] [Medline: 7016268]
2. Alfiya A, Paulius V, Clara D, Nicole B, Jochen Z, Georg S. Treatment with imatinib prevents fibrosis in different preclinical models of systemic sclerosis and induces regression of established fibrosis. Arthritis & Rheumatology 2010;60(1):219-224. [doi: 10.3410/f.1148053.605164] [Medline: 19116940]
3. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, et al. Nat Rev Drug Discov 2010 Mar;9(3):203-214. [doi: 10.1038/nrd3078] [Medline: 20168317]
4. Dudley JT, Sirota M, Shenoy M, Pai RK, Roedder S, Chiang AP, et al. Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. Sci Transl Med 2011 Aug 17;3(96):96ra76 [FREE Full text] [doi: 10.1126/scitranslmed.3002648] [Medline: 21849664]
5. Gaebel W, Zielasek J, Reed G. Mental and Behavioural Disorders in the ICD-11: Concepts, Methodologies, and Current Status. Psychiatr Pol 2017;51(2):169-195. [doi: 10.12740/pp/69660]
6. Cohen MH, Williams GA, Sridhara R, Chen G, McGuinn WD, Morse D, et al. United States Food and Drug Administration Drug Approval summary: Gefitinib (ZD1839; Iressa) tablets. Clin Cancer Res 2004 Feb 15;10(4):1212-1218 [FREE Full text] [doi: 10.1158/1078-0432.ccr-03-0564] [Medline: 14977817]
7. Kingsmore KM, Grammer AC, Lipsky PE. Drug repurposing to improve treatment of rheumatic autoimmune inflammatory diseases. Nat Rev Rheumatol 2020 Jan 12;16(1):32-52. [doi: 10.1038/s41584-019-0337-0] [Medline: 31831878]
8. Wadi L, Meyer M, Weiser J, Stein LD, Reimand J. Impact of outdated gene annotations on pathway enrichment analysis. Nat Methods 2016 Aug 30;13(9):705-706 [FREE Full text] [doi: 10.1038/nmeth.3963] [Medline: 27575621]
9. Lu P, Zhang H, Liu Y, Wu Y, Qin X, Xia J. Parameter searching in attractor algorithm for community detection—an application in pathway enrichment analysis. In: Journal of Physics: Conference Series. 2018 Aug 30 Presented at: 3rd Annual International Conference on Information System and Artificial Intelligence (ISAI2018); 22-24 June 2018; Suzhou p. 012051. [doi: 10.1088/1742-6596/1069/1/012051]
10. Hung J, Yang T, Hu Z, Weng Z, DeLisi C. Gene set enrichment analysis: performance evaluation and usage guidelines. Brief Bioinform 2012 May;13(3):281-291 [FREE Full text] [doi: 10.1093/bib/bbr049] [Medline: 21900207]
11. Deng X, Tavallaie MS, Sun R, Wang J, Cai Q, Shen J, et al. Drug discovery approaches targeting the incretin pathway. Bioorg Chem 2020 Jun;99:103810. [doi: 10.1016/j.bioorg.2020.103810] [Medline: 32325333]
12. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000 Jan 28;28(1):27-30. [doi: 10.1093/nar/28.1.27] [Medline: 10592173]
13. Masoudi-Sobhanzadeh Y, Omidi Y, Amanlou M, Masoudi-Nejad A. Drug databases and their contributions to drug repurposing. Genomics 2020 Mar;112(2):1087-1095. [doi: 10.1016/j.ygeno.2019.06.021] [Medline: 31226485]
14. Yan Y, Burbridge C, Shi J, Liu J, Kusalik A. Comparing four genome-wide association study (GWAS) programs with varied input data quantity. 2018 Presented at: 2018 IEEE international conference on bioinformatics and biomedicine (BIBM); Dec 3, 2018; Madrid p. 1800-1802. [doi: 10.1109/bibm.2018.8621425]
15. Yeganeh P, Mostafavi M, Lu P, Zhang H, Liu Y, Wu Y. Use of machine learning for diagnosis of cancer in ovarian tissues with a selected mRNA panel. 2018 Presented at: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); December 3, 2018; Madrid p. 2429-2434. [doi: 10.1109/bibm.2018.8621371]
16. Gachloo M, Wang Y, Xia J. A review of drug knowledge discovery using BioNLP and tensor or matrix decomposition. Genomics Inform 2019 Jun;17(2):e18. [doi: 10.5808/gi.2019.17.2.e18]
17. Qin X, Wang S, Wu Y, Xia J. Evaluation of the Performance of BioNLP Tools for Discovering Causal Genes in Terms with Pathway Enrichment. J. Phys.: Conf. Ser 2018 Aug 30;1069:012037. [doi: 10.1088/1742-6596/1069/1/012037]
18. Percha B, Altman RB. A global network of biomedical relationships derived from text. Bioinformatics 2018 Aug 01;34(15):2614-2624 [FREE Full text] [doi: 10.1093/bioinformatics/bty114] [Medline: 29490008]
19. Wei CH, Kao HY, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. Nucleic Acids Res 2013 Jul;41(Web Server issue):W518-W522 [FREE Full text] [doi: 10.1093/nar/gkt441] [Medline: 23703206]
20. Deng X, Tavallaie MS, Sun R, Wang J, Cai Q, Shen J, et al. Drug discovery approaches targeting the incretin pathway. Bioorg Chem 2020 Jun;99:103810. [doi: 10.1016/j.bioorg.2020.103810] [Medline: 32325333]

21. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, McMorran R, Wiegers J, et al. The Comparative Toxicogenomics Database: update 2019. Nucleic Acids Res 2019 Jan 08;47(D1):D948-D954 [FREE Full text] [doi: 10.1093/nar/gky868] [Medline: 30247620]
22. Debusmann R, Kuhlmann M. Dependency grammar: classification and exploration. In: Resource-Adaptive Cognitive Processes. Berlin: Springer; 2010:365-388.
23. Xia J, Fang AC, Zhang X. A novel feature selection strategy for enhanced biomedical event extraction using the Turku system. Biomed Res Int 2014;2014:205239 [FREE Full text] [doi: 10.1155/2014/205239] [Medline: 24800214]
24. Yu G, Wang L, Han Y, He Q. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 2012 May;16(5):284-287 [FREE Full text] [doi: 10.1089/omi.2011.0118] [Medline: 22455463]
25. Liu J, Li H, Zhou F, Yu J, Sun L, Han Z. Targeting the mTOR pathway in breast cancer. Tumour Biol 2017 Jun;39(6):1010428317710825 [FREE Full text] [doi: 10.1177/1010428317710825] [Medline: 28639903]

Abbreviations

CTD: comparative toxicogenomic database
IPF: inverse pathway frequency
KEGG: Kyoto Encyclopedia of Genes and Genomes
TEES: Turku Event Extraction System

Edited by T Hao; submitted 25.02.21; peer-reviewed by B Hu, W Heng; comments to author 30.03.21; revised version received 05.04.21; accepted 19.04.21; published 18.06.21.

Please cite as:
Qin X, Yao X, Xia J
A Novel Metric to Quantify the Effect of Pathway Enrichment Evaluation With Respect to Biomedical Text-Mined Terms: Development and Feasibility Study
JMIR Med Inform 2021;9(6):e28247
URL: https://medinform.jmir.org/2021/6/e28247
doi:10.2196/28247
PMID:34142969

Original Paper

Drug-Drug Interaction Predictions via Knowledge Graph and Text Embedding: Instrument Validation Study

Meng Wang[1,2*], PhD; Haofen Wang[3*], PhD; Xing Liu[4*], PhD; Xinyu Ma[1*], MA; Beilun Wang[1*], PhD

[1]School of Computer Science and Engineering, Southeast University, Nanjing, China

[2]Key Laboratory of Computer Network and Information Integration, Southeast University, Nanjing, China

[3]College of Design and Innovation, Tongji University, Shanghai, China

[4]Third Xiangya Hospital, Central South University, Changsha, China

[*]all authors contributed equally

Corresponding Author:
Haofen Wang, PhD
College of Design and Innovation
Tongji University
No. 281 Fuxin Road
Yangpu District
Shanghai, 200092
China
Phone: 86 139 1858 6855
Email: carter.whfcarter@gmail.com

Abstract

Background: Minimizing adverse reactions caused by drug-drug interactions (DDIs) has always been a prominent research topic in clinical pharmacology. Detecting all possible interactions through clinical studies before a drug is released to the market is a demanding task. The power of big data is opening up new approaches to discovering various DDIs. However, these data contain a huge amount of noise and provide knowledge bases that are far from being complete or used with reliability. Most existing studies focus on predicting binary DDIs between drug pairs and ignore other interactions.

Objective: Leveraging both drug knowledge graphs and biomedical text is a promising pathway for rich and comprehensive DDI prediction, but it is not without issues. Our proposed model seeks to address the following challenges: data noise and incompleteness, data sparsity, and computational complexity.

Methods: We propose a novel framework, Predicting Rich DDI, to predict DDIs. The framework uses graph embedding to overcome data incompleteness and sparsity issues to make multiple DDI label predictions. First, a large-scale drug knowledge graph is generated from different sources. The knowledge graph is then embedded with comprehensive biomedical text into a common low-dimensional space. Finally, the learned embeddings are used to efficiently compute rich DDI information through a link prediction process.

Results: To validate the effectiveness of the proposed framework, extensive experiments were conducted on real-world data sets. The results demonstrate that our model outperforms several state-of-the-art baseline methods in terms of capability and accuracy.

Conclusions: We propose a novel framework, Predicting Rich DDI, to predict DDIs. Using rich DDI information, it can competently predict multiple labels for a pair of drugs across numerous domains, ranging from pharmacological mechanisms to side effects. To the best of our knowledge, this framework is the first to provide a joint translation-based embedding model that learns DDIs by integrating drug knowledge graphs and biomedical text simultaneously in a common low-dimensional space. The model also predicts DDIs using multiple labels rather than single or binary labels. Extensive experiments were conducted on real-world data sets to demonstrate the effectiveness and efficiency of the model. The results show our proposed framework outperforms several state-of-the-art baselines.

(JMIR Med Inform 2021;9(6):e28277) doi:10.2196/28277

KEYWORDS

drug-drug interactions; knowledge graph; natural language processing

Introduction

An increasing amount of research in clinical studies is focusing on drug-drug interactions (DDIs) because the majority of adverse drug reactions (ADRs) occur between pairs of drugs. ADRs may lead to patient morbidity and mortality, accounting for 3% to 5% of all in-hospital medication errors [1]. Furthermore, patients with 2 or more diseases (eg, older adult patients with chronic diseases) have a higher risk of an ADR if they take 5 or more different drugs simultaneously [2,3]. Detecting DDIs based on experimentation is a time-consuming and laborious process for clinicians. This signals the need for a more comprehensive and automated method of predicting unknown DDIs before a new drug can be released.

Traditional experimental approaches in vitro [4], in vivo [5], and in populo [6] focus on small sets of specific drug pairs and have laboratory limitations. Many machine learning approaches, such as similarity or feature-based approaches [7-9], have been proposed to predict DDIs. Recently, several graph neural networks and long short-term memory methods based on knowledge graphs (KGs), such as KG neural network [10] and KG-DDI [11], have significantly outperformed traditional shallow machine learning methods. The superior performance of these proposed methods can be attributed to their use of the prior knowledge and learning of higher-level representations for DDI detection. However, as these approaches only predict binary DDIs or those that have been predefined in structured databases, they may be hampered by robustness caused by data sparsity and vast computation requirements. Although several approaches [12-14] have used natural language processing techniques to extract DDIs from biomedical text, to the best of our knowledge, they have not employed drug KGs to improve performance.

With the increasing emergence of biomedical data, many world-leading biomedical researchers are now focusing on automatically populating and completing biomedical KGs using the huge volume of structured databases and text available to the public. HKG [15], Knowlife [16], and DrugBank [17] are just a few examples. Efforts such as Bio2RDF [18] and Linked Open Drug Data [19] have mapped similar entities in different KGs and built large heterogeneous graphs that contain an abundance of basic biomedical facts about drugs. SPARQL [20], a query language for KGs, supports the retrieval and manipulation of drug-related facts distributed over different KGs. Unfortunately, these biomedical KGs are affected by incomplete and inaccurate data that impede their application in the field of safe medicine development.

Existing KGs already include thousands of relation types, millions of entities, and billions of facts [19]. As noted, KG applications based on conventional graph-based algorithms are restricted by data sparsity and computational inefficiency. To address these problems, graph embedding techniques [9,21-26] based on representation learning for KGs have been proposed that embed both entities and relations into a continuous low-dimensional vector space. Among these methods, translation-based models [9,22,24] are the most simple and effective. Currently, they represent the state-of-the-art in knowledge acquisition and inference and link prediction [9]. In light of these analogies, DDIs can be treated as a category of relations in a drug KG, and KG embedding techniques can be used to predict unknown DDIs. However, most translation-based methods only concentrate on predefined relations or unstructured text and fail to exploit the link between existing relations and rich unstructured text.

Leveraging both drug KGs and biomedical text is a promising pathway for rich and comprehensive DDI prediction, but it is not without issues. Our proposed model seeks to address the following challenges: data noise and incompleteness—real-world KGs are known to be inaccurate, incomplete, and unreliable for direct use; data sparsity—the potential DDI information in both KGs and biomedical text is sparse, and estimating the potential DDIs in such a long-tailed distribution is difficult; computational complexity—undoubtedly, this will be precluded from practice if graph-based algorithms are employed to process large-scale KGs or represent data objects with simple one-hot feature vectors.

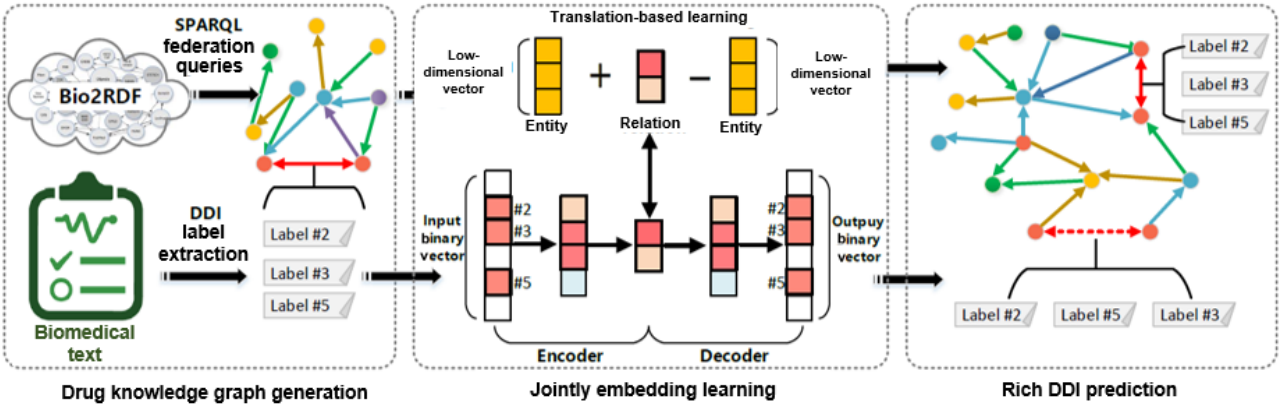
Given these challenges, we propose a novel framework called Predicting Rich DDI (PRD). The framework is based on graph embedding techniques and treats specific DDI predictions as a linked prediction process. The proposed framework proceeds as follows: A large, high-quality drug KG is generated from distributed drug resources, which includes data on drug-target interactions, the impact of drugs on gene expression, the outcomes of drugs in clinical trials, and so on. A novel translation-based embedding model embeds the entities and relations in the drug KG into a low-dimensional space, and an autoencoder incorporates the descriptions of the DDIs from biomedical text as representations into the same semantic space. The decoder predicts the corresponding labels for potential DDIs based on the learned embeddings.

To the best of our knowledge, our PRD approach is the first method that is able to predict comprehensive and specific DDIs based on large-scale drug KGs and comprehensive biomedical text on pharmacology and ADRs. Our method further includes a joint translation-based embedding model that encodes the KG and rich DDI information from biomedical text into a shared low-dimensional space. The DDI predictions are then translated into a linked prediction process from the learned embeddings. Extensive experiments on real-world data sets were conducted to evaluate the framework. The results show that the framework can be powerful in predicting rich DDIs and outperforms several state-of-the-art baselines in terms of both capability and accuracy.

Methods

Figure 1 shows the architecture of the proposed framework. It consists of 3 key phases: drug KG generation, joint embedding learning, and DDI relations prediction.

Figure 1. Overview of the framework. DDI: drug-drug interaction.



Drug KG Generation

A typical KG usually arranges knowledge as a triple set of facts that indicates the relation between 2 entities, and thus comprises a head entity, a relation, and a tail entity. These are denoted as (h, r, t).

First, a basic drug KG is constructed by collecting drug-related entities and relations among these entities. We follow the data model of drug-related extraction settings defined in the work of Kamdar and Musen [27], in which the types of entities or relations are summarized in the fashion depicted in Table 1. Specifically, we use SPARQL federation queries [20] to extract triples that contain 4 types of drug-related entities (E_1~E_4) and 5 types of biological relations (R_1~R_4) from a variety of biomedical sources (eg, Bio2RDF [18]). These extracted triples are defined as basic triples in our drug KG according to definition 1: (basic triple) $B = (E, R)$ is a set of basic triples in the form (h, r, t), where $E = E_1 \cup E_2 \ldots \cup E_4$ is a set of entities; and $R = R_1 \cup R_2 \ldots \cup R_5$ is a set of relations, $h, t \in E$, and $r \in R$.

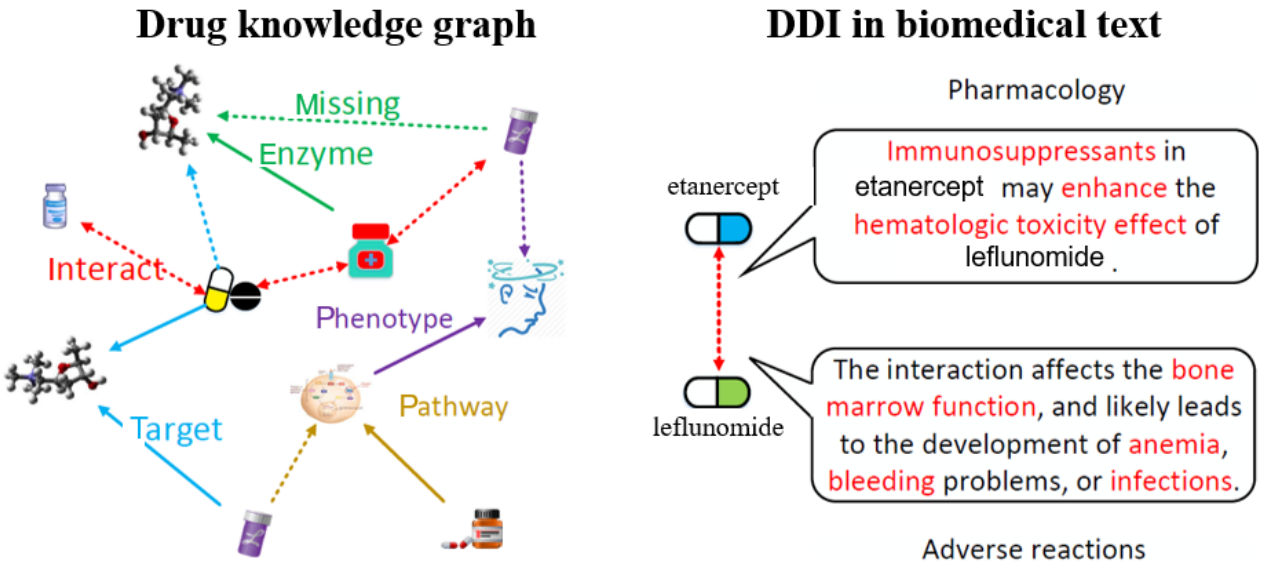
For instance, we can extract "(etanercept, hasTarget, lymphotoxin-alpha)" as a basic triple in our drug KG, which indicates that there is a relationship "hasTarget" linking etanercept to lymphotoxin-alpha, meaning that lymphotoxin-alpha is one of the targets of etanercept.

Table 1. Entities and relations of basic triples in Kamdar and Musen [27].

Variable	Entity or relation interpretation
E	
E_1	Drugs
E_2	Drugs
E_3	Pathways
E_4	Phenotypes
R	
R_1	Drug, hastarget, protein
R_2	Drug, hasenzyme, protein
R_3	Drug, hastransporter, protein
R_4	Protein, ispresentin, pathway
R_5	Pathway, isimplicatedin, phenotype

A specific DDI between 2 drugs can be captured by multiple key phrases extracted from biomedical text, as shown in Figure 2. Hence, we collect biomedical DDI text documenting drug pairs (eg, DDI corpus [28], MEDLINE abstracts, and DrugBank DDI documents). We remove all stop words from raw text and use an entity linking method [29] to align the drug names in the biomedical text with the KG. The top-n labels (n=5) are then selected from the biomedical text for each DDI based on the term frequency-inverse document frequency (TF-IDF) features (some other textual features can be used to rank the labels instead).

Figure 2. A drug knowledge graph is shown on the left with missing relations represented as dotted lines. There is usually no direct DDI relation between drugs. DDI descriptions from the biomedical text are shown on the right. The words in red represent concerns regarding DDI information in terms of both adverse DDIs and in-depth ways drugs can interact in pharmacology. DDI: drug-drug interaction.



Based on this, the DDI relations between drug entities are defined as a set of labels rather than as a single label according to definition 2: (rich DDI triple) $T = (E_1, L)$ is a set of rich DDI triples in the form (u, l, v), where E_1 is a set of drug entities; L is a fixed label vocabulary from biomedical text; and u, v $\in E_1$ and $l = \{n_1, n_2, \ldots\} \subseteq L$ is the set of labels to describe the DDI information.

For instance, the following is an example of a rich DDI triple: (etanercept, {immunosuppressants, enhancetoxicity, anemia, infections}, leflunomide), where "enhancetoxicity" means etanercept can enhance the toxicity of leflunomide. Note that the DDI relations between 2 drugs are bidirectional; hence, our method replaces each rich DDI relation with 2 directed triples of opposing directions in the drug KG.

Formally, the generated drug KG is defined according to definition 3 (drug KG): the drug KG, G, is denoted as (E, B, T), where $E = E_1 \cup E_2 \ldots \cup E_4$ is a set of entities, B is a set of basic triples, and T is a set of rich DDI triples.

Joint Embedding Learning

KG embedding mainly consists of 3 steps: representation of entities and relations, definition of a scoring function, and encoding of the entity and relation into dense vectors. This section introduces the translation-based KG embedding model that learns representations from the drug KG, $G = (E, B, T)$ and the optimization described in the following sections.

Basic Triple Encoder

For a set of basic triples, B, the method aims to encode entities and relations into a continuous vector space. This paper, without loss of generality, uses the bold letters **h**, **r**, **t** to denote the embedding vectors h, r, t. We adopt the translation-based mechanism **h+r≈t** to capture the correlations between entities and relations. Translation in this context refers to a translation operation **r** between 2 entity vectors **h** and **t** in the low-dimensional space. We follow the TransR model in Lin et al [22] to represent entities and relations in distinct vector spaces bridged by relation-specific matrices so as to learn more thorough graph representations. Specifically, for each triple, (h, $r, t) \in B$, h and t are embedded into h, $t \in R^k$, and r is embedded into $r \in R^d$. For each relation r, a projection matrix $M_r \in R^{(k \times d)}$ ×projects entities from the entity space to the relation space. The energy function z_{bte} (**h**, **r**, **t**) is then defined as follows:

$$z_{bte}(\mathbf{h}, \mathbf{r}, \mathbf{t}) = b_1 - \| \mathbf{h}\mathbf{M}_r + \mathbf{r} - \mathbf{t}\mathbf{M}_r \|_{(L1/L2)} \quad (\mathbf{1})$$

where b_1 is a bias constant.

The conditional probability of a triple h, r, t is defined as follows:

(**2**)

$P(t|h, r)$, $P(r|h, t)$ can be defined in an analogous manner. The likelihood of observing a triple (h, r, t) is defined as follows:

$$L(h, r, t) = \log P(h \mid r, t) + \log P(t \mid h, r) + \log P(r \mid h, t) \quad (\mathbf{3})$$

By maximizing the conditional likelihoods of all existing triples in B, the objective function is defined as follows:

(**4**)

It is worth mentioning that other graph embedding models, such as HOLE [23], can also be easily adopted for basic triple encoding. In the interest of brevity, this paper only explores the effectiveness of TransR.

Rich DDI Triple Encoder

The interaction l between 2 drug entities, u and v, in rich DDI triples $(u, l, v), \in T$, can also be represented as translations in

low-dimensional space. We set **u**, **v**$\in R^k$, **l**$\in R^d$. The energy function z_{dte} (u, l, v) is defined as follows:

$$z_{dte}(\mathbf{u}, \mathbf{l}, \mathbf{v}) = b_2 - \| \mathbf{u}\mathbf{M}_r + \mathbf{l} - \mathbf{v}\mathbf{M}_l \|_{(L1/L2)} \quad (5)$$

where b_2 is a bias constant and $\mathbf{M}_l = R^{\times d}$ is the projection matrix. Following the analogous method in the basic triple encoder, the conditional likelihoods of all existing triples are maximized as follows:

(**6**)

Note, in equation 5, l is the relation representation obtained from $l = \{n_1, n_2, \ldots\}$. This will be introduced in-depth next.

A deep autoencoder is employed to construct the relation representation $l \in R^d$ for a rich DDI triple (u, l, v) $\in$ T. Specifically, a DDI relation, l, is described by a set of labels $l = \{n_1, n_2, \ldots\} \subseteq L$. The corresponding binary vector for l is initialized as **s** =  , where $\mathbf{s}_i = 1$ if $n_i \in l$, and $\mathbf{s}_i = 0$ otherwise. The deep autoencoder then takes the binary vector **s** as input and uses the following nonlinear transformation layers to transform the label set into the low-dimensional space R^k:

$$h^{(1)} = f(\mathbf{W}^{(1)} s + \mathbf{b}^{(1)})$$

$$h^{(i)} = f(\mathbf{W}^{(i)} h^{(i-1)} + b^{(i)}), I = 2, \ldots, K \quad (7)$$

where f is the activation function and K is the number of layers. Here, $h^{(i)}$, $\mathbf{W}^{(i)}$, and $\mathbf{b}^{(i)}$ represent the hidden vector, transformation matrix, and the bias vector in the i-th layer, respectively.

There are 2 parts to the autoencoder: an encoder and a decoder. The encoder employs the *tanh* activation function to obtain the DDI relation representation $\mathbf{l} = h^{(K/2)}$. The decoder deciphers the embedding vector of **l** to obtain a reconstructed vector  . Intuitively, PRD should then minimize the distance  because the reconstructed vector  should be similar to **s**. However, the number of zero elements in **s** is usually much larger than that of nonzero elements due to data sparsity. This leads the decoder to tend to reconstruct zero elements rather than nonzero elements, which conflicts with our purpose. To overcome this obstacle, different weights are set for different elements, and the following objective function is maximized:

(**8**)

where b_3 is a bias constant, **x** is a weight vector, and $\odot$ is denoted as the Hadamard product. For **x** =  , $\mathbf{x}_i = 1$, if $\mathbf{s}_i = 0$, and $\mathbf{x}_i = \beta > 1$ otherwise. According to equation 8, the probability of P  can be defined as follows:

(**9**)

where S is the set of binary vectors of all DDI relations. The likelihood of reconstructing the binary vector s of a relation l can be defined as follows:

$$L(l) = logP \quad (10)$$

By maximizing the likelihoods of the encoding and the decoding for all described relations l, the objective function can be defined as follows:

(**11**)

Joint Learning and Optimization

The goal of the framework PRD is to not only represent the basic triples (drug KG B) but also the rich DDI triples (biomedical text T) in a unified joint embedding model. Considering the above 3 objective functions (equations 4, 6, and 11) together, the optimization function is defined as follows:

$$O(X) = L_{bte} + L_{dte} + L_{rcl} + \gamma C(X) \quad (12)$$

where X represents the embeddings of entities and relations, and γ is a hyper-parameter that weights the regularization factor $C(X)$, which is defined as follows:

$$C(X) = \quad (13)$$

where $[x]_+ = max(0, x)$ denotes the positive part of x. The regularization factor will normalize the embeddings during learning. We adopted the approach by Srivastava et al [30] to prevent deep neural networks from overfitting and used the Adam algorithm [31] to maximize the objective function.

It is impractical to directly compute the normalizers in $P(h \mid r, t)$, $P(t \mid h, r)$, $P(r \mid h, t)$, and P  , as calculating them would require summing the complete set of entities and relations. To address this problem, we use the negative sampling method from Mikolov et al [32] to transform the objective functions. Taking $P(h \mid r, t)$ as an example, the following objective function is maximized instead of using its original form:

$$log\sigma(z_{bte}(\mathbf{h}, \mathbf{r}, \mathbf{t}))$$

(**14**)

where c is the number of negative examples, $\sigma(x) = 1/(1 + exp(-x))$ is the sigmoid function,  is the invalid triple set, and z_{neg} is a function randomly sampling instances from  . When a positive triple (h, r, t) $\in B$ is selected to maximize equation 14, c-negative triples are constructed by sampling entities from a uniform distribution over E and replacing the head of (h, r, t). The objective functions of $P(r \mid h, t)$, $P(t \mid h, r)$, $logP$  , and $L(u.\ l.\ v)$ are transformed and maximized in an equivalent manner. Finally, PRD iteratively selects random mini-batches from the training set to learn the embeddings efficiently until convergence.

DDI Relations Prediction

The DDI prediction task can be defined as a link prediction problem on KG; that is, with the learned deep autoencoder and the embedding vectors of all entities and relations, the framework PRD can leverage the translation mechanism to predict the missing DDI relations between 2 drug entities. To be more specific, given 2 drug entities u, $v \in E_1$, the following equation predicts the potential relation embedding **l** between **u** and **v.**

$$\mathbf{l} = \mathbf{v}\mathbf{M}_l - \mathbf{u}\mathbf{M}_l \quad (15)$$

Finally, with the decoder part of the learned deep autoencoder, PRD can obtain the label set l by decoding the embedding vector l.

Results

To examines the effectiveness of the DDI prediction framework PRD, we performed 2 types of experiments. First, we compared the performance of our model to several baseline methods on binary-type DDI predictions. We then investigated PRD's strengths in modeling rich DDI relations between drug entities. The results demonstrate that PRD significantly outperformed the baselines in terms of both accuracy and capability.

Data Construction

Experiments in this paper were performed on 2 real drug-related data sets, Bio2RDF [18] and DDI Corpus [28].

Bio2RDF (version 4) is an open-source project that provides 11 billion triples from 35 biological and pharmacological KGs across a wide variety of drug-related entities, such as proteins, targets, and diseases. It is accessible online via the SPARQL endpoint.

DDI Corpus (2013 version) is a semantically annotated corpus of documents describing DDIs from the DrugBankdatabase and MEDLINE abstracts. It contains 233 MEDLINE abstracts and 784 DrugBank texts on the DDIs subjects. There are a total of 5021 annotated DDIs in 18,491 pharmacological sentences.

Following the federation queries in Kamdar and Musen [27], we extracted basic triples for our drug KG from 4 different KGs in Bio2RDF: (1) DrugBank [17] provides comprehensive data about drug, disease, and target information; (2) Kyoto Encyclopedia of Genes and Genomes [33] offers pathways, proteins, and drugs information; (3) PharmGKB [34] contains protein-drug-disease relations; (4) Comparative Toxicogenomics Database ([35] provides data about protein interactions and pathway-disease relations.

For the rich DDI triples, we collected 4694 DrugBank DDI sentences about 8197 drugs from the DDI corpus. The top 5 labels from each sentence were selected based on TF-IDF to construct rich DDI triples and build the DDI label vocabulary. To overcome the issue of inconsistent drug names between basic triples and rich DDI triples, we applied the entity linking method [29] to align the drug aliases.

The drug KG we constructed contains 71,460 basic triples, 4694 rich DDI triples, 8197 drug entities, 305,642 other entities, and 1053 distinct labels in the DDI vocabulary.

Baselines

For the baseline approaches, DDI prediction and state-of-the-art KG embedding methods were used. Three DDI methods were used:

1. Tiresias [8] is a large-scale similarity-based framework that predicts DDIs through link prediction. It takes various sources of drug-related data and knowledge as inputs and generates binary DDI predictions as outputs.
2. Syntax Convolutional Neural Network (SCNN) [36] represents a DDI extraction method based on a SCNN to extract 4 predefined DDI types (ADVICE, EFFECT, INT, and MECHANISM) from the biomedical literature.
3. Multitask dyadic DDI prediction (MDDP) [37] defines the DDI type prediction problem as a multitask dyadic regression problem. It can predict the specific DDI type between 2 drugs.

Two state-of-the-art KG embedding methods were used:

1. TransE [9] is the most representative translational distance model to embed components of a KG, including entities and relations, into continuous vector spaces. These embeddings can also be used for link prediction.
2. TransR [22] shares a similar approach with TransE, but represents entities and relations in distinct vector spaces bridged by relation-specific matrices.

Evaluation Method and Metrics

Given a drug KG with some DDI relations removed, rich DDI prediction aims to predict the occurrence of DDI relations among drug entities. DDI relations with a rate of 0.3 chosen randomly as the ground truths for the test set were removed, and the remaining KG was used as the training set. We also randomly sampled an equal number of drug pairs with no DDI relations to serve as the negative sample in the test set.

To make an unbiased comparison, we first treated DDI prediction as a binary classification task. Tiresias is already a binary classification model. For SCNN and MDDP, we defined the 2 DDI types as yes and no in the training model. For TransE, TransR, and our PRD method, we concatenated the representations of the entities of a candidate drug pair to form the feature vector and used logistic regression to train classifiers. We then treated multiple DDI type predictions as a multilabel classification task. For Tiresias, SCNN, and MDDP, we used their feature representation methods and adopted one-versus-rest logistic regression to train a multilabel classifier. For TransE and TransR, we separated each training triple (u, l, v) where $l = \{n_1, n_2, \ldots\}$ into several triples (ie, $[u, n_i, v]$ for $n_i \in l$), which could be directly used to train the models.

We used 10-fold cross-validation on the training set to tune PRD's embedding model. We determined the optimal parameters using a grid search strategy. The search ranges for the various parameters were as follows: the learning rate λ for the Adam algorithm {0.1, 0.01, 0.001}; γ for the soft constraints {0.1, 0.01, 0.001}; the vector dimension k {20, 50, 80, 100};

and all bias constants b_1, b_2, b_3, c were 10 to 10. The training instances were conducted over 1000 iterations. The running time per iteration was 391 seconds. The best configurations for the joint model were λ=0.001, γ=0.01, k=100, b_1=5, b_2=5, b_3=1, c=10, and K=3, with L_1 being used as a dissimilarity metric.

We used receiver operator characteristic curves and precision-recall curves to evaluate the proposed method on binary DDI-type predictions. For multiple DDI- type predictions, we followed the setting in TransE [9] and report the 2 measures as evaluation metrics: the average rank of all correct relations (MeanRank) and the proportion of correct relations ranked in top k (Hits@k). The above metrics may be biased for methods that rank other correct labels higher in the same label set. Hence, all other correct labels were filtered out before ranking. The filtered version is denoted as "Filter," and the unfiltered version is denoted as "Raw."

Experiment Results

As shown in Figure 3a and Figure 3b, the proposed framework PRD outperformed all baselines. In terms of the receiver operator characteristic curve, PRD outperformed Tiresias by 6.69%, TransR by 7.13%, and MDDP and TransE by 12%; meanwhile, SCNN had a relatively low predictive ability. According to the precision-recall curve, PRD learned 14.2% better than did Tiresias (which was at the top among the 3 DDI prediction baselines), 16.8% better than TransR, 21.57% better than MDDP, 25.33% better than TransE, and 37.89% better than SCNN.

Table 2 shows the evaluation results for rich DDI relation predictions according to the different evaluation metrics for both the raw and filter tests.

Figure 3. ROC and PR results of binary drug-drug interaction-type predictions. MDDP: multitask dyadic drug-drug interaction (DDI) prediction; ROC: receiver operator characteristic; PR: precision-recall.

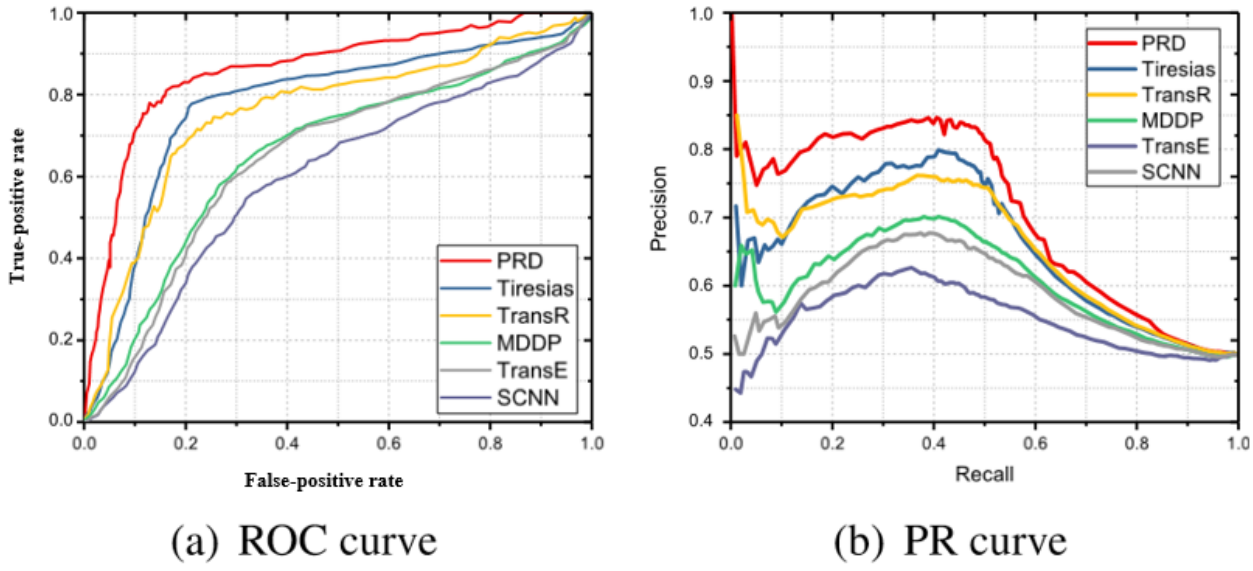


Table 2. Evaluation results for multiple drug-drug interaction relation predictions (×100 for Hits@k).

Framework	Raw				Filter			
	Hits@1[a]	Hits@5	Hits@10	MeanRank[b]	Hits@1	Hits@5	Hits@10	MeanRank
Tiresias	14.23	33.18	50.61	21.89	19.21	45.29	52.94	17.93
SCNN[c]	12.19	26.31	39.02	37.91	16.82	27.03	40.78	37.06
MDDP[d]	20.95	58.66	79.48	13.53	43.19	68.57	84.12	7.85
TransE	26.61	70.23	83.97	8.01	57.88	79.99	87.27	7.02
TransR	31.33	75.80	87.63	6.89	69.58	84.01	89.01	6.25
PRD[e]	45.11	85.57	91.01	6.11	75.11	88.60	92.85	5.45

[a]Hits@x: accuracy of real values contained in the top x rank.

[b]MeanRank: the average rank of all correct relations.

[c]SCNN: Syntax Convolutional Neural Network.

[d]MDDP: multitask dyadic drug-drug interaction prediction.

[e]PRD: Predicting Rich Drug-Drug Interaction.

Case Study

To further demonstrate PRD's ability for rich DDI predictions, we selected the drug acetylsalicylic acid (aspirin) as a test case. The DDI predictions and rich labels relations are shown in Table 3. According to the usefulness and diversity of the predicted labels, a professional pharmacist evaluated and annotated the practical useful predictions (labels in italics in Table 3). Observe that both TransR and PRD were able to recommend reasonable DDI labels for the drug interactions, representative of detailed DDI information. However, TransR sometimes recommended similar labels for a specific drug because it is based on a similarity method. Conversely, PRD was able to recommend discriminative labels because it uses a decoder.

We also present a case study to visualize the effectiveness of binary DDI types of prediction on a DDI network sample. We constructed drug-drug networks to indicate whether any 2 drugs would result in a binary DDI. A node in the network denotes a drug. An edge between 2 nodes denotes the existence of a DDI. Intuitively, the more drugs interact, the more risk there is. In the network, the size of the node specifies the degree of risk of a drug. We classified the degree of risk into various levels using different colors (ie, high risk is shown in dark green, and low risk is shown in light green). The red nodes denote forecasting errors of DDI drugs. As shown in Figure 4a to Figure 4f, PRD predicts DDIs mostly accurately. The ID of the drug with a high risk is shown on the node.

Table 3. Rich drug-drug interaction predictions for acetylsalicylic acid.

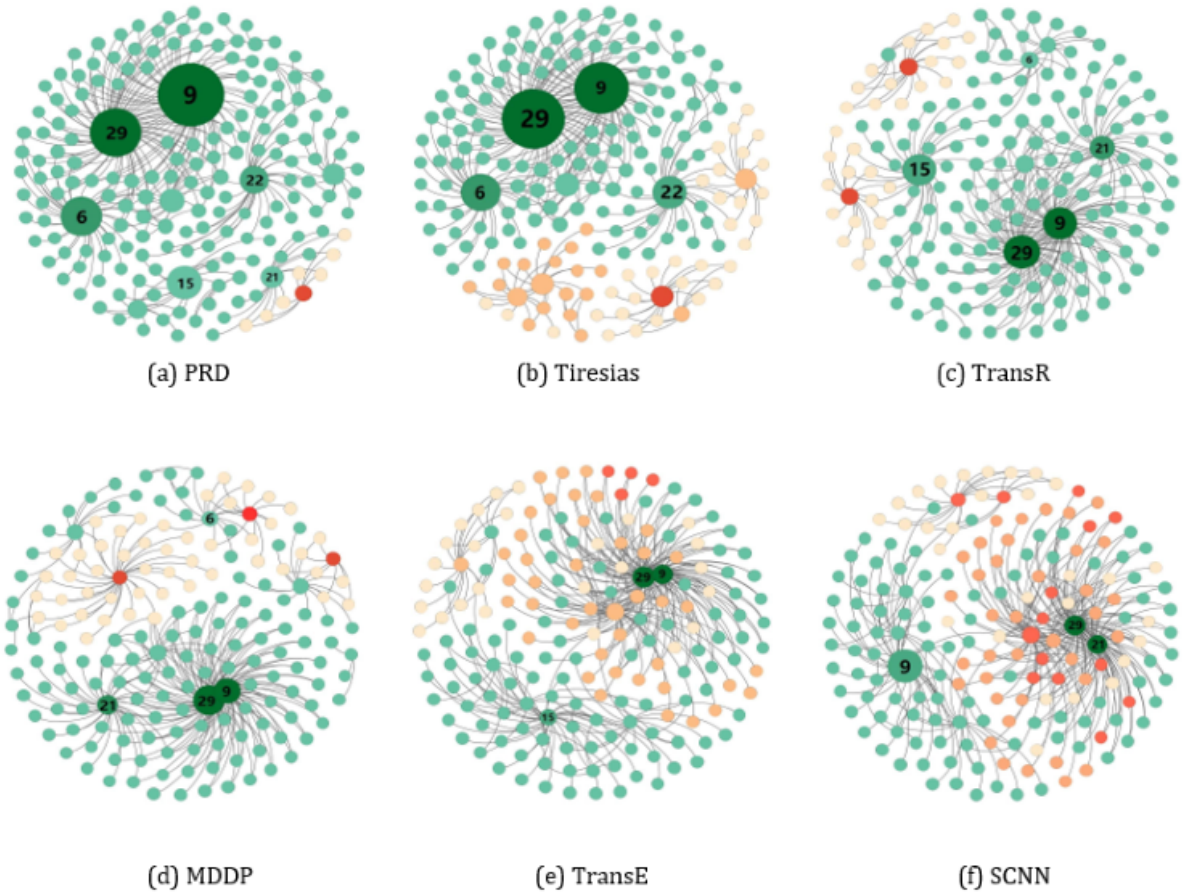
Interacted drug	TransR[a]	PRD[b]
Ibritumomab	*enhance*[c]*adverse*, *toxic*, risk, bleeding	*enhance*, *toxic*, *bleeding*, *platelet*, *antiplatelet*
Alteplase	*enhance*, increase, *adverse*, toxic, effect	*enhance*, *toxic*, *bleeding*, *thrombolytic*, *adverse*
Anistreplase	*enhance*, effect, *thrombolytic*, *agents*, *anticoagulant*	*enhance*, *anticoagulant*, *antiplatelet*, *thrombolytic*, *agents*
Ramipril	*diminish*, *antihypertensive*, effect, treatment, affect	*diminish*, *antihypertensive*, *inhibitor*, *doses*, affect

[a]TransR: a knowledge graph embedding model, which performs translation in the corresponding relation space.

[b]PRD: Predicting Rich Drug-Drug Interaction.

[c]Labels in italics indicate those annotated by a professional pharmacist.

Figure 4. Case visualization of the binary drug-drug interaction-type prediction on a drug-drug interaction network sample. MDDP: multitask dyadic drug-drug interaction prediction; PRD: Predicting Rich Drug-Drug Interaction; SCNN: Syntax Convolutional Neural Network.



Discussion

Principal Findings

PRD achieved a significant improvement over all baselines. Specifically, PRD outperformed MDDP by around 10%. MDDP is currently considered to be the best DDI prediction baseline for multiple DDI type predictions. Tiresias and SCNN performed poorly because they neglect various types of semantic information concerning DDIs. These results demonstrate the effectiveness of PRD to predict rich DDI relations among drug entities.

Compared to TransR and TransE, PRD also performed better, as it incorporates binary DDI types into the relation representation learning and also models multiple DDI labels of a DDI relation simultaneously. This accounts for its promising results in rich DDI prediction.

Conclusions

PRD is unlike other existing models. Using rich DDI information, it can competently predict multiple labels for a pair of drugs across numerous domains, ranging from pharmacological mechanisms to side effects. To the best of our knowledge, this framework is the first to provide a joint translation-based embedding model that learns DDIs by integrating drug KGs and biomedical text simultaneously in a common low-dimensional space. The model also predicts DDIs using multilabels, rather than single or binary labels. Extensive experiments were conducted on real-world data sets to demonstrate the effectiveness and efficiency of the model. The results show PRD outperforms several state-of-the-art baselines. In future work, we intend to incorporate a convolutional neural network to encode the rich DDI text to improve the performance of the embedding model. Another direction for our research is to have the embedding model consider subgraph features composed in the generated drug KG during learning. This may make it possible to predict DDIs among 3 or more drugs.

Acknowledgments

Grant support was received from the National Natural Science Foundation of China with (grant 61906037), the Fundamental Research Funds for the Central Universities with (grants 4309002159 and 22120210109), and the CCF-Baidu Open Fund.

Conflicts of Interest

None declared.

References

1. Leape LL. Systems analysis of adverse drug events. JAMA 1995 Jul 05;274(1):35. [doi: 10.1001/jama.1995.03530010049034]
2. Juurlink DN, Mamdani M, Kopp A, Laupacis A, Redelmeier DA. Drug-drug interactions among elderly patients hospitalized for drug toxicity. JAMA 2003 Apr 02;289(13):1652-1658. [doi: 10.1001/jama.289.13.1652] [Medline: 12672733]
3. Wang S, Li X, Chang* X, Yao L, Sheng QZ, Long G. Learning multiple diagnosis codes for ICU patients with local disease correlation mining. ACM Trans. Knowl. Discov. Data 2017 Apr 14;11(3):1-21. [doi: 10.1145/3003729]
4. Huang S, Strong JM, Zhang L, Reynolds KS, Nallani S, Temple R, et al. New era in drug interaction evaluation: US Food and Drug Administration update on CYP enzymes, transporters, and the guidance process. J Clin Pharmacol 2008 Jun;48(6):662-670. [doi: 10.1177/0091270007312153] [Medline: 18378963]
5. Quinney S, Zhang X, Lucksiri A, Gorski JC, Li L, Hall SD. Physiologically based pharmacokinetic model of mechanism-based inhibition of CYP3A by clarithromycin. Drug Metab Dispos 2010 Feb;38(2):241-248 [FREE Full text] [doi: 10.1124/dmd.109.028746] [Medline: 19884323]
6. Schelleman H, Bilker W, Brensinger C, Han X, Kimmel S, Hennessy S. Warfarin with fluoroquinolones, sulfonamides, or azole antifungals: interactions and the risk of hospitalization for gastrointestinal bleeding. Clin Pharmacol Ther 2008 Nov;84(5):581-588 [FREE Full text] [doi: 10.1038/clpt.2008.150] [Medline: 18685566]
7. Vilar S, Uriarte E, Santana L, Lorberbaum T, Hripcsak G, Friedman C, et al. Similarity-based modeling in large-scale prediction of drug-drug interactions. Nat Protoc 2014 Sep;9(9):2147-2163 [FREE Full text] [doi: 10.1038/nprot.2014.151] [Medline: 25122524]
8. Abdelaziz I, Fokoue A, Hassanzadeh O, Zhang P, Sadoghi M. Large-scale structural and textual similarity-based mining of knowledge graph to predict drug–drug interactions. Journal of Web Semantics 2017 May;44:104-117. [doi: 10.1016/j.websem.2017.06.002]
9. Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O. Translating embeddings for modeling multi-relational data. 2013 Presented at: The 26th Annual Conference on Neural Information Processing Systems; 2013 Dec 3-10; Lake Tahoe, USA.
10. Lin X, Quan Z, Wang ZJ, Ma T, Zeng X. KGNN: Knowledge graph neural network for drug-drug interaction prediction. 2021 Presented at: The 29th International Joint Conference on Artificial Intelligence; 2012 Jan 7-15; Yokohama, Japan. [doi: 10.24963/ijcai.2020/380]

11. Karim MR, Cochez M, Jares JB, Uddin M, Beyan O, Decker S. Drug-Drug Interaction Prediction Based on Knowledge Graph Embeddings and Convolutional-LSTM Network. 2019 Presented at: The 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics; 2019 Sept 7-10; Niagara Falls, USA.
12. Huang D, Jiang Z, Zou L, Li L. Drug–drug interaction extraction from biomedical literature using support vector machine and long short term memory networks. Information Sciences 2017 Nov;415-416:100-109. [doi: 10.1016/j.ins.2017.06.021]
13. Bui Q, Sloot PMA, van Mulligen EM, Kors JA. A novel feature-based approach to extract drug-drug interactions from biomedical text. Bioinformatics 2014 Dec 01;30(23):3365-3371. [doi: 10.1093/bioinformatics/btu557] [Medline: 25143286]
14. Liu S, Tang B, Chen Q, Wang X. Drug-drug interaction extraction via convolutional neural networks. Comput Math Methods Med 2016;2016:6918381-6918388 [FREE Full text] [doi: 10.1155/2016/6918381] [Medline: 26941831]
15. Rotmensch M, Halpern Y, Tlimat A, Horng S, Sontag D. Learning a health knowledge graph from electronic medical records. Sci Rep 2017 Jul 20;7(1):5994 [FREE Full text] [doi: 10.1038/s41598-017-05778-z] [Medline: 28729710]
16. Ernst P, Siu A, Weikum G. KnowLife: a versatile approach for constructing a large knowledge graph for biomedical sciences. BMC Bioinformatics 2015 May 14;16:157 [FREE Full text] [doi: 10.1186/s12859-015-0549-5] [Medline: 25971816]
17. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, et al. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res 2014 Jan;42(Database issue):D1091-D1097 [FREE Full text] [doi: 10.1093/nar/gkt1068] [Medline: 24203711]
18. Belleau F, Nolin M, Tourigny N, Rigault P, Morissette J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. J Biomed Inform 2008 Oct;41(5):706-716 [FREE Full text] [doi: 10.1016/j.jbi.2008.03.004] [Medline: 18472304]
19. Samwald M, Jentzsch A, Bouton C, Kallesøe CS, Willighagen E, Hajagos J, et al. Linked open drug data for pharmaceutical research and development. J Cheminform 2011 May 16;3(1):19 [FREE Full text] [doi: 10.1186/1758-2946-3-19] [Medline: 21575203]
20. Harris S, Seaborne A. SPARQL 1.1 Query Language. W3C Working Draft. 2013 Mar 21. URL: https://www.w3.org/TR/2009/WD-sparql11-query-20091022/ [accessed 2021-05-01]
21. Chang X, Yang Y. Semisupervised feature analysis by mining correlations among multiple tasks. IEEE Trans. Neural Netw. Learning Syst 2017 Oct;28(10):2294-2305. [doi: 10.1109/tnnls.2016.2582746]
22. Lin Y, Jones P, Samatova NF. Learning entity and relation embeddings for knowledge graph completion. 2017 Presented at: ACM on Conference on Information and Knowledge Management; 2017 Nov 6-10; Singapore. [doi: 10.1145/3132847.3133095]
23. Nickel M, Rosasco L, Poggio T. Holographic embeddings of knowledge graphs. 2016 Presented at: Proceedings of the 30th AAAI Conference on Artificial Intelligence; 2016 Feb 12-17; Phoenix, USA.
24. Tu C, Zhang Z, Liu Z, Sun M. TransNet: translation-based network representation learning for social relation extraction. 2017 Presented at: International Joint Conference on Artificial Intelligence; 2017 August 19-25; Melbourne, Australia.
25. Wang S, Chang X, Li X, Long G, Yao L, Sheng QZ. Diagnosis code assignment using sparsity-based disease correlation embedding. IEEE Trans. Knowl. Data Eng 2016 Dec 1;28(12):3191-3202. [doi: 10.1109/tkde.2016.2605687]
26. Zhu X, Li X, Zhang S, Ju C, Wu X. Robust joint graph sparse coding for unsupervised spectral feature selection. IEEE Trans. Neural Netw. Learning Syst 2017 Jun;28(6):1263-1275. [doi: 10.1109/tnnls.2016.2521602]
27. Kamdar M, Musen MA. PhLeGrA: Graph analytics in pharmacology over the web of life sciences linked open data. 2017 Presented at: The 26th International Conference on World Wide Web; 2017 Apr 3-7; Perth, Australia. [doi: 10.1145/3038912.3052692]
28. Herrero-Zazo M, Segura-Bedmar I, Martínez P, Declerck T. The DDI corpus: an annotated corpus with pharmacological substances and drug-drug interactions. J Biomed Inform 2013 Oct;46(5):914-920 [FREE Full text] [doi: 10.1016/j.jbi.2013.07.011] [Medline: 23906817]
29. Wang M, Zhang J, Liu J, Hu W, Wang S, Li X, et al. PDD graph: bridging electronic medical records and biomedical knowledge graphs via entity linking. 2017 Presented at: ISWC 2017: 16th International Semantic Web Conference; 2017 Oct 21–25; Vienna, Austria.
30. Srivastava N, Hinton J, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning 2014;15(1):1929-1958.
31. Kingma D, Ba JL. Adam: a method for stochastic optimization. 2015 Presented at: The 3rd International Conference for Learning Representation; 2015 May-79; San Diego, USA.
32. Mikolov T, Sutskever I, Chen K, Corrado K, Dean J. Distributed representations of words and phrases and their compositionality. 2013 Presented at: The 26th Annual Conference on Neural Information Processing Systems; 2013 Dec 5-10; Lake Tahoe, USA. [doi: 10.4324/9780203776506-14]
33. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000 Jan 01;28(1):27-30 [FREE Full text] [doi: 10.1093/nar/28.1.27] [Medline: 10592173]
34. Hewett M, Oliver DE, Rubin DL, Easton KL, Stuart JM, Altman RB, et al. PharmGKB: the Pharmacogenetics Knowledge Base. Nucleic Acids Res 2002 Jan 01;30(1):163-165 [FREE Full text] [doi: 10.1093/nar/30.1.163] [Medline: 11752281]

35. Davis A, Murphy CG, Johnson R, Lay JM, Lennon-Hopkins K, Saraceni-Richards C, et al. The Comparative Toxicogenomics Database: update 2013. Nucleic Acids Res 2013 Jan;41(Database issue):D1104-D1114 [FREE Full text] [doi: 10.1093/nar/gks994] [Medline: 23093600]
36. Zhao Z, Yang Z, Luo L, Lin H, Wang J. Drug drug interaction extraction from biomedical literature using syntax convolutional neural network. Bioinformatics 2016 Nov 15;32(22):3444-3453 [FREE Full text] [doi: 10.1093/bioinformatics/btw486] [Medline: 27466626]
37. Jin B, Yang H, Xiao C, Zhang P, Wei X, Wang F. Multitask dyadic prediction and its application in prediction of adverse drug-drug interaction. 2017 Presented at: The 31st AAAI Conference on Artificial Intelligence; 2017 Feb 4-9; San Francisco, USA.

Abbreviations

ADR: adverse drug reaction
DDI: drug-drug interaction
KG: knowledge graph
MDDP: multitask dyadic drug-drug interaction prediction
PRD: Predicting Rich Drug-drug Interaction
SCNN: Syntax Convolutional Neural Network
TF-IDF: term frequency-inverse document frequency

Edited by T Hao; submitted 28.02.21; peer-reviewed by H Ye, Y Gao, Z Zhang, Z Yang; comments to author 30.03.21; revised version received 29.04.21; accepted 05.05.21; published 19.06.21.

Please cite as:
Wang M, Wang H, Liu X, Ma X, Wang B
Drug-Drug Interaction Predictions via Knowledge Graph and Text Embedding: Instrument Validation Study
JMIR Med Inform 2021;9(6):e28277
URL: https://medinform.jmir.org/2021/6/e28277/
doi:10.2196/28277
PMID:34185011

Original Paper

Document Retrieval for Precision Medicine Using a Deep Learning Ensemble Method

Zhiqiang Liu[1], BA; Jingkun Feng[1], MD; Zhihao Yang[1], PhD; Lei Wang[2], PhD

[1]College of Computer Science and Technology, Dalian University of Technology, Dalian, China

[2]Beijing Institute of Health Administration and Medical Information, Beijing, China

Corresponding Author:
Zhihao Yang, PhD
College of Computer Science and Technology
Dalian University of Technology
No. 2 Ling Gong Road
Gan Jing Zi District
Dalian
China
Phone: 86 131 9011 4398
Email: yangzh@dlut.edu.cn

Abstract

Background: With the development of biomedicine, the number of biomedical documents has increased rapidly bringing a great challenge for researchers trying to retrieve the information they need. Information retrieval aims to meet this challenge by searching relevant documents from abundant documents based on the given query. However, sometimes the relevance of search results needs to be evaluated from multiple aspects in specific retrieval tasks, thereby increasing the difficulty of biomedical information retrieval.

Objective: This study aimed to find a more systematic method for retrieving relevant scientific literature for a given patient.

Methods: In the initial retrieval stage, we supplemented query terms through query expansion strategies and applied query boosting to obtain an initial ranking list of relevant documents. In the re-ranking phase, we employed a text classification model and relevance matching model to evaluate documents from different dimensions and then combined the outputs through logistic regression to re-rank all the documents from the initial ranking list.

Results: The proposed ensemble method contributed to the improvement of biomedical retrieval performance. Compared with the existing deep learning–based methods, experimental results showed that our method achieved state-of-the-art performance on the data collection provided by the Text Retrieval Conference 2019 Precision Medicine Track.

Conclusions: In this paper, we proposed a novel ensemble method based on deep learning. As shown in the experiments, the strategies we used in the initial retrieval phase such as query expansion and query boosting are effective. The application of the text classification model and relevance matching model better captured semantic context information and improved retrieval performance.

(JMIR Med Inform 2021;9(6):e28272) doi:10.2196/28272

KEYWORDS

biomedical information retrieval; document ranking; precision medicine; deep learning

Introduction

In recent years, biomedical research has developed rapidly leading to a great increase in the number of biomedical publications. Biomedical development promotes the treatment of intractable diseases; however, the huge number of biomedical documents brings a great challenge for researchers in obtaining the documents related to one topic. Biomedical information retrieval (IR) is thus a hot research topic in the biomedical domain.

Given a query, biomedical IR systems are designed to provide users with all relevant documents in a ranked list, sorted according to their relevance to the query. The relevance can be evaluated by applying different IR models [1-4] based on either the occurrence of query terms in the documents or probabilistic measures. However, it is difficult to achieve an ideal retrieval

performance when directly applying these IR models to biomedical IR. One possible reason is that the IR models cannot interpret the semantic information of the query and can only use frequencies and other features of query terms appearing in documents to determine the relevance. For example, when given a query "How is melanoma treated?" the goal of the query is to find relevant documents focusing on the treatment of melanoma. Since some documents focusing on other aspects such as clinical trials and pathology also contain many instances of the query term melanoma, the model considers these documents related, thus leading a poor retrieval performance. Moreover, biomedical documents usually contain diversified concept expressions and abundant professional vocabularies, and these vocabularies can usually be replaced by their synonyms or abbreviations, which increases the difficulty in relevance evaluation. In addition, in some specific biomedical IR tasks, the relevance between query and document needs to be evaluated from multiple aspects. For example, in the precision medicine (PM) retrieval task, for patients with certain diseases and genetic variants, researchers need to connect patients with experimental treatments if existing treatments have been ineffective; the retrieval goal is to find the experimental treatments for which the patients are eligible. The retrieval system must determine whether the patient meets the experiment requirements from multiple aspects such as disease, genetic variants, age, and so on, thereby increasing the difficulty and cost of system design. All the above situations bring domain-specific challenges for biomedical IR. Therefore, it is necessary to explore an effective biomedical IR method.

To alleviate the above problems, in this paper we propose a novel ensemble method based on deep learning for biomedical IR. Given the patient's disease, genetic variants, and demographic information, our method aims to find documents that provide information relevant to the treatment of the patient's disease. Therefore, our method needs to evaluate documents from treatment, disease, and gene dimensions. In particular, existing studies have proved that the IR task can be treated as a relevance matching problem between query and document [5,6]. Based on this, researchers have proposed a variety of matching models from different perspectives. To refine the retrieval performance in these approaches, after obtaining an initial ranked list of relevant documents retrieved through a search engine, a relevance matching model is deployed as a re-ranker over the ranked list to re-rank all relevant documents. Following other researchers, we also consider the relevance matching model as a component of the re-ranker to re-rank relevant documents. Specifically, our method can be divided into two phases: initial retrieval and re-ranking. During the initial retrieval phase, to alleviate the problem of diverse concept expressions and abundant professional vocabularies with synonyms in biomedical IR, we introduce external biomedical resources to create a local database, and based on this, we design effective query expansion strategies to reformulate the original query by supplementing relevant terms to better describe the retrieval need. We also design query boosting strategies to adjust the weights of query terms. During the re-ranking phase, to alleviate the problem of a retrieval model that cannot interpret the query semantics, we employ a relevance matching model based on deep learning to capture semantic signals between query and document from the disease dimension. To make our re-ranker evaluate articles from multiple dimensions, we built an effective text classification model to determine whether a document is treatment-focused. In particular, to combine the two models effectively, we apply a logistic regression (LR) model to output the final score for each document and reorder our initial ranking list according to their scores. Experimental results on the collections from the Text Retrieval Conference (TREC) 2019 PM track demonstrate that the proposed method can effectively improve the retrieval performance in biomedical IR.

We summarize the contributions of our work as follows:

- We propose an ensemble method based on deep learning to evaluate relevance between query and document from multiple dimensions in biomedical IR.
- We introduce an effective relevance matching model and text classification model to fully capture semantic information from a query and refine the retrieval performance in biomedical IR.
- We apply the LR method to combine the relevance matching model and text classification model, and experimental results show that it is more effective than the voting method.

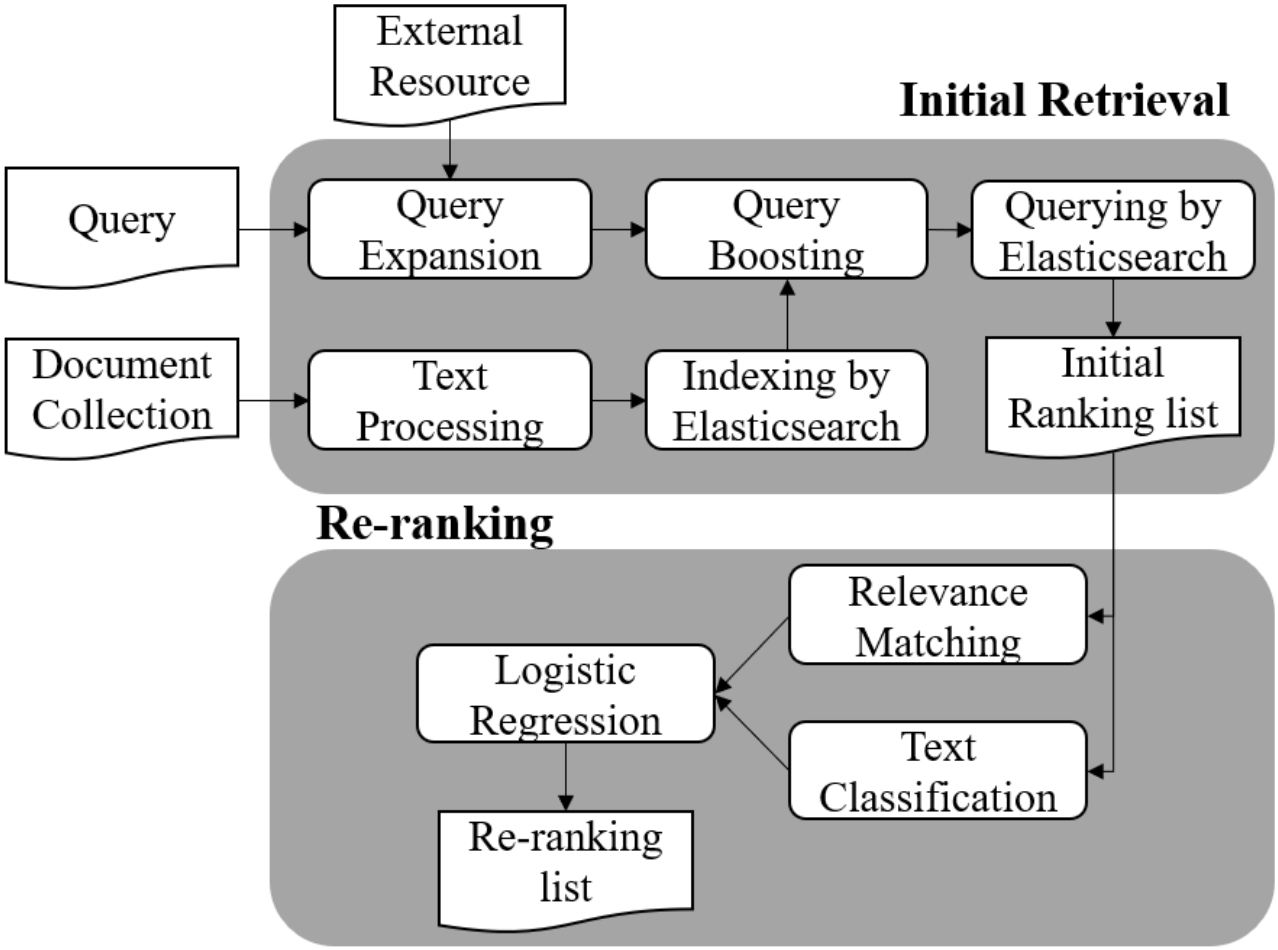
The remainder of this paper is organized as follows. In section 2, we discuss some related work. In section 3, we describe our method in detail. In section 4, we discuss the experiments conducted to evaluate the effectiveness of the proposed method. In section 5, we conclude the paper and provide suggestions for future work.

Methods

Model Architecture

In this section, we illustrated the framework of our ensemble method and provided detailed descriptions. Figure 1 describes the overview of the architecture of our method, divided into two phases: initial retrieval phase and re-ranking phase. In the initial retrieval phase, we first supplemented query terms through query expansion strategies we designed, then we used a search engine named Elasticsearch to index documents, and finally, we applied query boosting to obtain an initial ranked list of relevant documents. In the re-ranking phase, we first employed a text classification model and relevance matching model to evaluate documents respectively from different dimensions, then we combined their outputs through LR, and finally we re-ranked all the documents from the initial ranking list according to their relevance scores.

Figure 1. Overview of the architecture of our method.



Initial Retrieval

Given a query Q = {q_1, q_2, ..., q_M}, the goal of this phase is to obtain an initial ranking list D = {d_1, d_2, …, d_N}, where q_i represents the i-th term in the query, d_i represents a candidate document related to query, M stands for the number of query terms, and N stands for the number of candidate documents. Specifically, we chose BM25 [7], a probabilistic retrieval model commonly used in search engines, to calculate the relevance score between query and document. To make our retrieval process more efficient and convenient, after preprocessing the whole document collection, we used Elasticsearch, an open-source Lucene-based full-text search engine, to index all documents and search relevant candidate documents.

Text Processing

The document collection is a snapshot of PubMed abstracts, and XML and TXT versions are available. The XML versions have the complete information for each abstract. We extracted text information from fields that might be useful like ArticleTitle, Abstract, ChemicalList, MeshHeadingList, and OtherAbstract fields. The information was saved in JSON format, which is convenient for index building.

Query Expansion

Considering that biomedical documents usually contain abundant specialized words with synonyms and abbreviations, we first built a local database to introduce external biomedical resources to improve the recall rate of retrieval results. In the database, we stored biomedical disease and gene entities, as well as their entity IDs, synonyms, hypernyms, and acronyms. In particular, the disease information is derived from the Comparative Toxicogenomics Database [8], while the genetic information is derived from the National Center for Biotechnology Information gene database. Next, we supplemented query terms with their synonyms and acronyms to better describe the retrieval need. Since our method aims to retrieve documents that focus on disease treatment, we additionally introduced some treatment-related keywords into the queries such as surgery, therapy, patient, resistance, recurrence, therapeutic, prevent, prophylaxis, prophylactic, prognosis, outcome, survival, treatment, and efficacy.

Query Boosting

To improve the retrieval performance in the initial retrieval phase, we used query boosting to define different weights for different query fields during the retrieval process. Specifically, we compiled the query template provided by Elasticsearch to boost some query fields. In our custom query template, there were 4 query fields: disease, genetic variant, treatment keyword, and demographic. Among them, disease and gene fields were considered as the most important query fields, hence they had higher weight values than other fields.

Querying

For each original query Q, we reformulated their query terms through query expansion and defined their weights by query boosting. Then we used Elasticsearch based on BM25 to retrieve candidate documents related to Q and rank them in descending order according to their BM25 scores. Finally, we obtained our initial ranking list D.

In this phase, the relevance scores of documents were calculated based on the prominence of query terms appearing in documents, and the semantic information was not considered. In the next phase, all candidate documents in D were reevaluated from multiple dimensions.

Re-Ranking

The goal of the re-ranking phase was to refine the retrieval performance by re-ranking all documents obtained from the previous phase. Given the initial ranking list D = $\{d_1, d_2, \ldots, d_N\}$, our re-ranker reevaluated relevance between queries and documents from multiple dimensions and reordered them in descending order according to their new relevance scores.

During the re-ranking phase, for all documents in the initial ranking list, a text classification model and relevance matching model were employed to reevaluate these documents from different dimensions. Then, an LR model was applied to combine the two models and output final relevance scores. Finally, according to the final scores, a heuristic rule was applied to re-rank these documents.

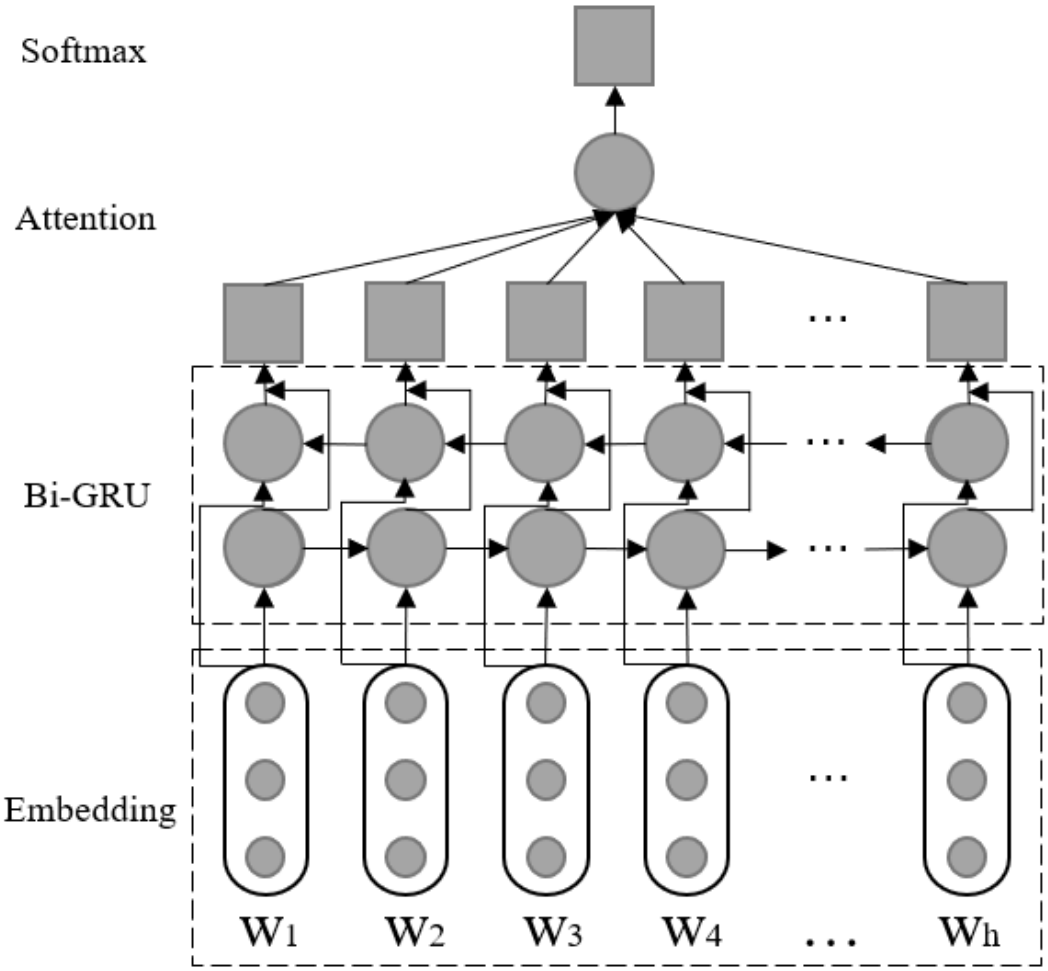
Text Classification

In our method, a text classification model was used to determine whether a document was treatment-focused. As a component of the re-ranker, our text classification model is a binary classification model.

When preprocessing the documents in initial ranking list D, for each document, we connected its title and abstract with delimiter SEP, and converted all letters to lowercase. Next, we replaced all numbers that appeared in the document with token NUM. Finally, we normalized it by defining the maximum document length h: if the document is shorter than h, we used token PAD filling it to h; otherwise, we truncated it to h directly.

When building the text classification model (bidirectional gated recurrent unit– attention [BiGRU-Att]), we adopted a bidirectional gated recurrent unit (GRU) [9] layer to encode the input word sequence and capture the context information. Also, an attention mechanism [10] was used to focus on relevant words to each category so that our method accurately picked out the corresponding documents to a query. We illustrate the structure of our text classification model in Figure 2.

Figure 2. Structure of the text classification model.



Given a document $d = \{w_1, w_2, \ldots, w_h\}$, where w_i represents the i-th word in it, the Embedding layer represented it to the word embedding matrix $M = \{v_1, v_2, \ldots, v_h\}$, where v_i represented the embedding vector of the i-th word.

In the bidirectional GRU layer, given the input v_t, the hidden state h_t was computed as follows:

where, r_t is the reset gate, z_t is the update gate, h_{t-1} is the previous state, $\tilde{h}_t$ is the candidate state at time t, v_t is the sequence vector at time t, $\sigma(\cdot)$ and $\tanh(\cdot)$ is sigmoid and hyperbolic tangent functions. b_z, b_r, and b_h are bias terms. The operator  denotes element-wise multiplication.

To generate the context feature matrix H, we concatenated the matrixes $H_{forward}$ and H_{back}, which are the output of the forward and back GRU's hidden layers respectively, namely, $H = [H_{forward}; H_{back}]$.

In the Attention layer, for the input xt, the attention weight αt was calculated as follows:

where, d is the dimension of input vector, and q is the query vector. Finally, the probability that the input document was classified into each category was calculated through a softmax layer.

Relevance Matching

In our method, a relevance matching model was applied to reevaluate the relevance between query and document from disease dimension.

When preprocessing the documents in initial ranking list D, for each document, based on the text classification preprocessing, we removed all stop words in the document and applied the Porter Stemmer [11] to stem the remain words. Since the biomedical documents usually contain abundant professional words with synonyms, for all disease entities in the document, if their synonyms appear in the same document, we replaced them with the same entity ID. Simultaneously, we also used the same entity ID to replace the disease entities that belonged to the same concept in the query.

To consider the semantic information between the query and the document when reevaluating the relevance, we adopted the MatchPyramid [12] model as our relevance matching model. Specifically, given the query q and the document d, MatchPyramid takes the query-document pair (q, d) as input. Then, a dot product operation as follows was employed to generate a matching matrix S between query and document.

Where α_i and β_j are the i-th and the j-th word embedding vectors from q and d, respectively. Finally, hierarchical convolutional neural networks and multilayer perceptron were applied to output the relevance score.

Logistic Regression

In our method, given the query q, each document d in the initial ranking list was judged with a simple measure: definitely relevant, partially relevant, and not relevant. Therefore, the goal of employing the LR model was to determine the relevance levels of documents by comprehensively considering the evaluation results of all models.

When building our LR model, we transferred the problem of relevance evaluation into a classification problem. Namely, given the labels of definitely relevant, partially relevant, and not relevant for each query, we used the LR model to determine the label of each candidate document. Since ordinary LR can only handle binary classification problems, we built 3 binary classification models (LR1, LR2, and LR3) based on the LR (ie, our LR model [LR] is composed of the 3 binary classifiers). In LR1, definitely relevant documents were treated as positive samples while other documents were treated as negative samples. In LR2, partially relevant documents were treated as positive samples while other documents were treated as negative samples. In LR3, documents that were not relevant were treated as positive samples while other documents were treated as negative samples.

Specifically, for a document d from the initial ranking list, we built 3 LR models with $[v_1, v_2, v_1^2, v_2^2]$ as input and the sigmoid function as the activation function. Cross entropy was used as the loss function. v_1 was the probability that measure d is treatment-focused or not, and v_2 was the relevance score. We obtained the 2 values from the text classification model and the relevance matching model. Then we computed their outputs that represented probabilities on the 3 labels, respectively. Finally, we chose the label that indicated the maximum probability as the final output of the LR model.

Re-Ranking Rule

After determining the relevance level of each document through the LR model, we used a heuristic rule to reorder the initial ranking list. Specifically, we ranked the definitely relevant documents over partially relevant documents, and the documents that were not relevant were ranked last. For the documents belonging to the same relevance level, we ranked them in descending order according to their BM25 scores obtained through Elasticsearch.

Results

Experiment Settings

TREC 2019 PM Track focused on an important use case in PM for clinical decision support: providing useful PM-related information to clinicians treating cancer patients. Participants of the track were challenged with retrieving (1) biomedical articles in the form of article abstracts (largely from MEDLINE/PubMed) addressing relevant treatments for the given patient and (2) clinical trials (from ClinicalTrials.gov) addressing relevant clinical trials for which the patient was eligible. In particular, we mainly focused on the first task. The first task aimed to retrieve relevant treatment information for the given diseases from the scientific literature. This task

provided 40 topics consisting of the disease, genetic variants, and demographic information about the patients. And for each topic, the participating system needed to return up to 1000 related documents retrieved from the scientific literature. For judging the relevance of documents, the organizer mainly evaluated from 3 dimensions: treatment, disease, and gene. To evaluate each retrieval result, precision at rank 10 (P@10), R-precision (R-prec), and inferred normalized discounted cumulative gain (infNDCG) are used as the evaluation metrics [13].

Specifically, we used TREC-Eval, a tool provided by the TREC organizer, to implement the evaluation of our experimental results on the 3 metrics. To train the BiGRU-Att, we used the gold standard of the TREC 2017 PM Track [14] as the data source of the model. According to the annotation of the gold standard, we first labeled all documents with 2 categories: treatment-focused or not. Then, we randomly divided the whole dataset into a training set, a development set, and a test set at a ratio of 8:1:1. When training the model, we applied the Adam algorithm [15] for parameter optimization and used the development set to optimize the hyperparameters. Finally, we applied the early stop mechanism to select the number of training iterations. Table 1 lists the hyperparameters of the BiGRU-Att. When the training was complete for each document in the initial ranking list, we used BiGRU-Att to determine whether it was treatment-focused.

To train the MatchPyramid, we also used the gold standard of the TREC 2017 PM Track as the data source of the model. For each document in the gold standard, according to its annotation on the disease dimension, we adopted the following labeling strategies: if the disease was exact, we labeled it with 2; if the disease was more specific or more general, we labeled it with 1; and if the disease is not disease, we labeled it with 0. Then, we randomly divided all data into a training set, development set, and test set at a ratio of 8:1:1. To implement the model, we used MatchZoo [16], an open source text matching tool, to build the MatchPyramid. When training the model, we applied the Adagrad algorithm [17] for parameter optimization and used the development set to optimize the hyperparameters. Finally, we applied the early stop mechanism to select the number of training iterations. Table 1 lists the hyperparameters of the MatchPyramid. When the training was complete for each document in the initial ranking list, we used MatchPyramid to determine its relevance level on the disease dimension.

To train the LR1, LR2, and LR3, the gold standard of the TREC 2017 PM was used as the data source of the models. According to the relevance level annotated in the gold standard, we adopted the following labeling strategies to construct a dataset for each LR model: given a topic in LR1, we labeled definitely relevant documents with 1 and others with 0; in LR2, we labeled partially relevant documents with 1 and others with 0; and in LR3, we labeled unrelated documents with 1 and others with 0. We then randomly divided each dataset into a training set, development set, and test set at a ratio of 8:1:1. When training these models, we used scikit-learn [18], a machine learning library, to build and select the 3 LR models, and we applied the early stop mechanism to select the number of training iterations. When the training was complete for each document in the initial ranking list, we used the 3 LR models in turn to predict its relevance level and re-ranked the initial ranked list according to our heuristic rule.

Table 1. The hyperparameters of BiGRU-Att and MatchPyramid.

Model and parameter	Value
BiGRU-Att[a]	
Document max length	256
Word embedding dimension	200
Hidden layer dimension	200
Learning rate	0.001
Batch size	128
MatchPyramid	
Query max length	30
Document max length	200
Word embedding dimension	200
Number of convolution kernel	400
Convolution kernel size	5
Learning rate	0.001
Batch size	64

[a]BiGRU-Att: bidirectional gated recurrent unit– attention.

Query Expansion Experiment

To explore the impact of query expansion on retrieval performance, we conducted corresponding experiments with different expansion strategies. As shown in Table 2, when expanding the disease field, using synonym expansion achieved better retrieval performance while using hypernym expansion

made the retrieval performance worse. The reason is that hypernyms represent a more general concept, but the TREC PM task requires retrieving the treatment information about a specific disease. Therefore, the disease name itself should be paid more attention during the retrieval process, which made synonym expansion outperform hypernym expansion. When expanding the gene field, synonym expansion greatly reduced the search performance, and compared with not using query expansion, acronym expansion had no obvious impact on retrieval performance. By analyzing retrieval results, we found that the treatment-focused articles usually did not mention related genes. After synonym expansion for gene, the proportion of gene keywords in query terms increased sharply, so when we searched based on exact matching models such as BM25, genetics-focused articles obtained a higher score, leading to a poor performance. In addition, the gene entities were usually expressed in abbreviation form, and their acronyms were rarely used. Therefore, acronym expansion had little effect on search results. Finally, it can be seen from Table 2 that, after treatment keywords were added, the treatment-focused documents achieved higher scores leading to improved retrieval performance.

In subsequent experiments, we adopted the following query expansion strategy: for the disease field, we used synonym expansion, and for the gene field, we did not. In addition, we added a treatment field to supplement treatment-related keywords.

Table 2. Experimental results of query expansion.

Disease		Gene		Treatment	P@10[a]	R-prec[b]	infNDCG[c]
Syn[d]	Hyper[e]	Syn	Acro[f]				
—[g]	—	—	—	—	0.5325	0.2934	0.4585
✓	—	—	—	—	0.5675	0.3113	0.4783
—	✓	—	—	—	0.5300	0.2942	0.4577
—	—	✓	—	—	0.4550	0.2801	0.4324
—	—	—	✓	—	0.5325	0.2933	0.4580
—	—	—	—	✓	0.5425	0.3104	0.4732
✓	—	—	—	✓	0.5700	0.3223	0.4882

[a]P@10: precision at rank 10.

[b]R-prec: R-precision.

[c]infNDCG: inferred normalized discounter cumulative gain.

[d]Syn: synonym.

[e]Hyper: hypernym.

[f]Acro: acronym.

[g]No expansion.

[h]Corresponding term is not applied for expansion.

Query Boosting Experiment

In our method, we used query boosting to optimize the weights of different query fields. Our query template included query clauses for disease, gene, treatment, and demographic information about the patients, respectively, and they are expressed as Q_d, Q_g, Q_t, and Q_p. To enhance the performance during the initial retrieval phase, we conducted the corresponding experiment by setting different weights for different fields. The experimental results are shown in Table 3. Among them, when we boost the weight of a field, the weights of other fields are set to 1.0.

It can be seen from Table 3 that when we boosted the weights of Q_d and Q_g, the retrieval performance improved, indicating that disease and genetic variants are more important than other clauses. When the weights of Q_t and Q_p are boosted, the retrieval performance was not improved, indicating that the treatment keywords and demographic of patients cannot provide more specific information for retrieval.

In subsequent experiments, we adopted the following query boosting strategy: $Q_d = 1.5$, $Q_g = 1.5$, $Q_t = 1.0$ and $Q_p = 1.0$.

Table 3. Experimental results of query boosting.

Strategy	P@10[a]	R-prec[b]	infNDCG[c]
No boosting	0.5700	0.3223	0.4882
$Q_d = 1.5$	0.5750	0.3246	0.4923
$Q_g = 1.5$	0.5750	0.3238	0.4911
$Q_t = 1.5$	0.5700	0.3231	0.4893
$Q_p = 1.5$	0.5700	0.3225	0.4884
$Q_d = 1.5$, $Q_g = 1.5$	0.5800	0.3250	0.4981

[a]P@10: precision at rank 10.

[b]R-prec: R-precision.

[c]infNDCG: inferred normalized discounter cumulative gain.

Ensemble Experiment

To explore the impact of models applied in the re-ranking phase on retrieval performance, we conducted the ensemble experiment. Based on the initial retrieval phase, we combined one model at a time to reorder the documents in the initial ranking list. To explore the effectiveness of the LR on integrating BiGRU-Att and MatchPyramid, we used a voting algorithm for comparison. That is, for one document, if it was predicted as treatment-focused and its relevance score was greater than zero, then it was definitely relevant; if it was predicted as not treatment-focused and its relevance score was less than zero, then it was not relevant; otherwise, it was partially relevant. The experimental results are shown in Table 4. Among them, the baseline is the performance of the initial retrieval phase, and Ensemble-TC and Ensemble-RM denote adding BiGRU-Att and MatchPyramid, respectively, ALL-Voting and ALL-LR denote integrating all models with voting algorithm and LR, respectively.

It can be seen from Table 4 that when integrating the BiGRU-Att, P@10 is increased from 0.58 to 0.63. This is because BiGRU-Att can determine whether a document is treatment-focused, and rank these documents in a higher position. When integrating the MatchPyramid, the retrieval performance is also improved due to the ability of our deep matching model on capturing semantic context information between query and document. Also, when integrating by the LR, the retrieval performance is better than the voting algorithm, and this suggests that the application of LR to ensemble models is more effective than the voting algorithm.

Table 4. Ensemble experiment results.

Model	P@10[a]	R-prec[b]	infNDCG[c]
Baseline	0.5800	0.3250	0.4981
Ensemble-TC[d]	0.6300	0.3324	0.5142
Ensemble-RM[e]	0.6075	0.3393	0.5049
ALL-Voting[f]	0.6375	0.3348	0.5143
ALL-LR[g]	0.6500	0.3391	0.5237

[a]P@10: precision at rank 10.

[b]R-prec: R-precision.

[c]infNDCG: inferred normalized discounter cumulative gain.

[d]Ensemble-TC: baseline + BiGRU-Att.

[e]Ensemble-RM: baseline + MatchPyramid.

[f]ALL-Voting: all models integrated with voting algorithm.

[g]ALL-LR: all models integrated with LR.

Performance Comparison Experiment

To explore the performance of the proposed method on the biomedical IR task, we compared our method with other deep learning–based systems participating in TREC 2019 PM task. The experimental results are shown in Table 5.

Table 5. Performance comparison of various methods.

Team	Method	P@10[a]	R-prec[b]	infNDCG[c]
Ours	baseline	0.5800	0.3250	0.4981
Ours	ALL-LR[d]	0.6500	0.3391	0.5237
CCNL[e]	SciBERT	0.6500	0.3066	0.5309
DUTIR[f]	Ensemble	0.5975	0.3273	0.5108
ECUN_ICA[g]	Doc2vec	0.5675	0.2718	0.4672

[a]P@10: precision at rank 10.

[b]R-prec: R-precision.

[c]infNDCG: inferred normalized discounter cumulative gain.

[d]ALL-LR: all models integrated with LR.

[e]CCNL: team name.

[f]DUTIR: team name.

[g]ECUN_ICA: team name.

Among these comparison systems based on deep learning, CCNL treated the document re-ranking problem as a two-category problem (ie, the documents definitely relevant and partially relevant to given topics are considered positive samples while unrelated documents are considered negative samples). Team CCNL trained a SciBERT (Scientific Bidirectional Encoder Representations from Transformers) [19] to classify all documents. DUTIR is the system we submitted in the task. In this system, we combined recurrent convolutional neural networks for text classification [20], deep relevance matching model for ad-hoc retrieval, and recurrent convolutional neural networks for text classification [21] to evaluate candidate documents from treatment, disease, and gene dimensions, respectively. Moreover, ECNU_ICA trained a doc2vec model to encode both documents and queries into fixed-length vectors and then used their cosine scores as similarity metrics.

As can be seen from Table 5, compared with other methods, the R-prec of our method (ALL-LR) was the best (0.3391). Additionally, our R-prec during the initial retrieval phase reached 0.3250, which was better than most of the other methods. This indicates that our query expansion and query boosting strategies worked well on biomedical IR. After integrating our re-ranker models, our R-prec further improved from 0.3250 to 0.3391. Meanwhile, our P@10 improved from 0.58 to 0.65, which was the same as the best result of other methods. Moreover, the result of infNDCG improved from 0.4981 to 0.5237. This shows that during the re-ranking phase, the semantic context features between queries and documents were better captured, thereby optimizing the initial retrieval performance. However, although the result of infNDCG improved, it was still lower than that of CCNL. One possible reason is that our ensemble method inevitably introduced the problem of error propagation because the accuracies of the employed models were not 100% and the deviation of these models led to some mistakes in determining the relevance level of some documents.

Discussion

Principal Findings

In this paper, we proposed a novel ensemble method based on deep learning for biomedical IR. The experimental results showed that (1) the query expansion and query boosting strategies we designed are effective, (2) the application of the text classification model and relevance matching model fully captured semantic context information and improved the retrieval performance, (3) using LR to combine models was more effective than the voting algorithm, and (4) our ensemble method evaluated relevance between query and document from multiple aspects in biomedical IR.

Limitations

However, there is still much room for performance improvement. The problem of error propagation limits the performance of the ensemble method, and using a joint model to address the problem may be an effective solution. In addition, domain feature engineering has been proven to effectively improve retrieval performance, and, therefore, constructing the domain features to enhance the retrieval performance is also our future work.

Conclusion

In this work, we introduced the ensemble method for the relevance evaluation from multiple aspects in biomedical IR. Our method annotated the usefulness of query expansion and query boosting by simultaneously applying them to obtain the large number of documents related to the query. To evaluate relevance from multiple dimensions and refine the retrieval performance, we integrated the text classification model and relevance matching model through LR modelling. Overall, we attributed the improvement of the proposed method in biomedical IR to two aspects: initial retrieval strategies and re-ranking models. For the initial retrieval strategies, we expanded the query terms with their synonyms and defined different weights for different query fields, which improved the accuracy and recall rate during the initial retrieval phase. For the re-ranking models, we introduced the text classification

model and relevance matching model, which evaluated the relevance of search results from multiple dimensions. These aspects jointly contributed to improvement in retrieval performance, and the proposed method showed the effectiveness of evaluating relevance from multiple aspects.

Acknowledgments

This work was supported by grant 2016YFC0901902 from the National Key Research and Development Program of China.

Conflicts of Interest

None declared.

References

1. Rocchio J. Relevance feedback in information retrieval. In: Smart Retrieval System. Englewood Cliffs: Prentice Hall; 1971:313-323.
2. Lavrenko W. Relevance based language models. Proc 24th Annu Int ACM SIGIR Conf Res Develop Inf Retrieval 2001:120-127. [doi: 10.1145/383952.383972]
3. Robertson E, Walker S, Beaulieu M, Gatford M. Okapi at TREC-4. Proc 4th Text Retrieval Conf NIST Special Publication 1996:73-97 [FREE Full text]
4. Zhai C, Lafferty J. Model-based feedback in the language modeling approach to information retrieval. Proc 10th Int Conf Inf Knowl Manag 2001:403-410. [doi: 10.1145/502585.502654]
5. Huang PS, He X, Gao J. Learning deep structured semantic models for web search using clickthrough data. Proc 22nd ACM Int Conf Inf Knowl Manag 2013:1. [doi: 10.1145/2505515.2505665]
6. Lu Z, Li H. A deep architecture for matching short texts. Adv Neural Inf Proc Syst 2013:1. [doi: 10.7551/mitpress/11474.003.0014]
7. Robertson S, Zaragoza H. The Probabilistic Relevance Framework: BM25 and Beyond. Found Trends Inf Retrieval 2009;3(4):333-389. [doi: 10.1561/1500000019]
8. Mattingly CJ, Colby GT, Forrest JN, Boyer JL. The Comparative Toxicogenomics Database. Enviro Health Perspect 2003 May;111(6):793-795. [doi: 10.1289/ehp.6028]
9. Dzmitry B, Kyunghyun C, Yoshua B. Neural Machine Translation by Jointly Learning to Align and Translate. 2015 Presented at: The International Conference on Learning Representations; 2015; San Diego URL: https://arxiv.org/pdf/1409.0473
10. Vaswani A, Shazeer N, Parmar N. Attention is all you need. 2017 Dec Presented at: Annual Conference on Neural Information Processing Systems; 2017; Long Beach.
11. Willett P. The Porter stemming algorithm: then and now. Prog Electr Library Inf Syst 2006 Jul;40(3):219-223. [doi: 10.1108/00330330610681295]
12. Pang L, Lan Y, Guo J. Text Matching as image recognition. 2016 Feb Presented at: Thirtieth AAAI Conference on Artificial Intelligence; 2016; Phoenix.
13. Roberts K, Demner-Fushman D. Overview of the TREC 2018 Precision Medicine Track. Proc 27th Text Retrieval Conf 2018:1 [FREE Full text]
14. Roberts K, Demner-Fushman D, Voorhees E, Hersh W, Bedrick S, Lazar A, et al. Overview of the TREC 2017 Precision Medicine Track. Text Retr Conf 2017 Dec;26:1 [FREE Full text] [Medline: 32776021]
15. Kingma DP, Ba J. Adam: a method for stochastic optimization. 2015 Presented at: The International Conference on Learning Representations; 2015; San Diego.
16. Fan Y, Pang L, Hou JP. MatchZoo: a toolkit for deep text matching. arXiv. URL: https://arxiv.org/abs/1707.07270 [accessed 2017-07-23]
17. Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. J Mach Learn Res 2011;12:2121-2159 [FREE Full text]
18. Kramer O. Scikit-Learn Machine Learning for Evolution Strategies. Berlin: Springer International Publishing; 2016.
19. Beltagy I, Cohan A, Lo K. SciBert: a pretrained language model for scientific text. 2019 Nov Presented at: Proc 2019 Conf Empirical Methods Nat Lang Proc; 2019; Hong Kong p. 3615-3620. [doi: 10.18653/v1/D19-1371]
20. Lai S, Xu L, Liu K. Recurrent convolutional neural networks for text classification. Proc 29th AAAI Conf Artificial Intellig 2015:2267-2273 URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9745 .
21. Hui K, Yates A, Berberich K. PACRR: a position-aware neural IR model for relevance matching. 2017 Sep Presented at: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing; 2017; Hong Kong p. 1049-1058 URL: https://www.aclweb.org/anthology/D17-1110.pdf [doi: 10.18653/v1/D17-1110]

Abbreviations

BiGRU-Att: bidirectional gated recurrent unit– attention
GRU: gated recurrent unit
infNDCG: inferred normalized discounter cumulative gain
IR: information retrieval
LR: logistic regression
P@10: precision at rank 10
PM: precision medicine
R-prec: R-precision
SciBERT: Scientific Bidirectional Encoder Representations from Transformers
TREC: Text Retrieval Conference

Edited by T Hao; submitted 27.02.21; peer-reviewed by B Dong, Z Ye; comments to author 30.03.21; revised version received 29.04.21; accepted 05.05.21; published 29.06.21.

Please cite as:
Liu Z, Feng J, Yang Z, Wang L
Document Retrieval for Precision Medicine Using a Deep Learning Ensemble Method
JMIR Med Inform 2021;9(6):e28272
URL: https://medinform.jmir.org/2021/6/e28272
doi:10.2196/28272
PMID:34185006

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