JMIR Bioinformatics and Biotechnology

Methods, devices, web-based platforms, open data and open software tools for big data analytics, understanding biological/medical data, and information retrieval in biology and medicine.
Volume 3 (2022), Issue 1 ISSN 2563-3570 Editor in Chief: Ece D. Uzun, MS, PhD, FAMIA

Contents

Reviews

Original Papers

Review

Digital Phenotyping in Health Using Machine Learning Approaches: Scoping Review

Schenelle Dayna Dlima[1], MSc; Santosh Shevade[1], MPharm; Sonia Rebecca Menezes[1], MSc; Aakash Ganju[1], MD

Saathealth, Mumbai, India

Corresponding Author:
Schenelle Dayna Dlima, MSc
Saathealth
1103, Glen Croft, Hiranandani Gardens, Powai
Mumbai, 400076
India
Phone: 971 559558006
Email: schenelle@saathealth.com

Abstract

Background: Digital phenotyping is the real-time collection of individual-level active and passive data from users in naturalistic and free-living settings via personal digital devices, such as mobile phones and wearable devices. Given the novelty of research in this field, there is heterogeneity in the clinical use cases, types of data collected, modes of data collection, data analysis methods, and outcomes measured.

Objective: The primary aim of this scoping review was to map the published research on digital phenotyping and to outline study characteristics, data collection and analysis methods, machine learning approaches, and future implications.

Methods: We utilized an a priori approach for the literature search and data extraction and charting process, guided by the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-analyses Extension for Scoping Reviews). We identified relevant studies published in 2020, 2021, and 2022 on PubMed and Google Scholar using search terms related to digital phenotyping. The titles, abstracts, and keywords were screened during the first stage of the screening process, and the second stage involved screening the full texts of the shortlisted articles. We extracted and charted the descriptive characteristics of the final studies, which were countries of origin, study design, clinical areas, active and/or passive data collected, modes of data collection, data analysis approaches, and limitations.

Results: A total of 454 articles on PubMed and Google Scholar were identified through search terms associated with digital phenotyping, and 46 articles were deemed eligible for inclusion in this scoping review. Most studies evaluated wearable data and originated from North America. The most dominant study design was observational, followed by randomized trials, and most studies focused on psychiatric disorders, mental health disorders, and neurological diseases. A total of 7 studies used machine learning approaches for data analysis, with random forest, logistic regression, and support vector machines being the most common.

Conclusions: Our review provides foundational as well as application-oriented approaches toward digital phenotyping in health. Future work should focus on more prospective, longitudinal studies that include larger data sets from diverse populations, address privacy and ethical concerns around data collection from consumer technologies, and build “digital phenotypes” to personalize digital health interventions and treatment plans.

(JMIR Bioinform Biotech 2022;3(1):e39618) doi:10.2196/39618

KEYWORDS

digital phenotyping; machine learning; personal device data; passive data; active data; wearable device; wearable sensor; mobile application; digital health

Introduction

Patient engagement is a significant challenge that health care organizations face, as consumers expect and demand a more personalized approach when they seek health care services [1]. Artificial intelligence (AI)–led smart health care services are emerging as promising tools to improve the efficiency and effectiveness of health care service delivery [2]. Among these is digital phenotyping, which is the real-time collection of

individual-level active and passive data from users in naturalistic and free-living settings via personal digital devices, such as mobile phones and wearable devices [3]. Personal digital devices and platforms, such as smartphones, wearable devices, and social media, offer a wealth of information about an individual's behavior and health status. These are valuable sources of several active and passive data points, such as phone utilization metrics, GPS information, search histories, linguistic nuances in text messages, duration of sleep, step counts, calories burned, and heart rate variability. These data points can be leveraged to gain a nuanced understanding of individual behaviors to predict disease exacerbation or relapse, design a more targeted intervention, and improve decision making in clinical settings [2,3].

Digital phenotyping is an emerging field that intersects data analysis, engineering, and clinical practice, bringing about unique challenges in reporting and reproducibility. Although the advantages of a multidisciplinary approach are evident, these multidisciplinary domains have yet to be brought together efficiently to ensure standardized reporting and easier replicability [4].

The techniques and methodologies used to collect, process, and classify active and passive data in digital phenotyping vary across the literature. AI and machine learning have already driven developments in wearable sensing and mobile health; they have helped enhance human activity recognition models, improve the accuracy of predicting human behaviors, and deliver more personalized lifestyle recommendations [5]. Research points to trust, perceived usefulness, and personalization directly influencing the frequency of use of digital health care services [2].

Given the plethora of data points that smartphones and wearable sensors and devices yield, AI and machine learning can be used to process and analyze these large data sets [6]. The purpose of passive data is to improve patient monitoring and outcomes across a variety of clinical applications [7]. In a systematic review of machine learning studies on digital phenotyping across psychosis spectrum illnesses, the machine learning approaches used included random forests, support vector machines, neural nets, k-nearest neighbors, and naive Bayes classifiers [8]. Machine learning algorithms used to analyze these multidimensional data can also be used to predict risks and probabilities and make binary decisions, such as discharge versus no discharge [9]. Other computational tools that have been used for digital phenotyping include data mining and statistical methods [10].

The immense potential of digital phenotyping in the clinical landscape is gaining increasing attention, leading to a measurable increase in related published research in the past 5 years. This trend has also been observed for health and clinical research related to analyzing active and passive data from smartphones and wearable devices. Digital phenotyping perhaps demonstrates the greatest potential for precision digital health interventions. Assigning a digital phenotype can help build predictive models around user behavior, providing insights into their engagement levels and the means to optimize the efficacy of digital health interventions. This method of segmentation offers further opportunities to enhance diagnosis, risk prediction, treatment effectiveness, and patient monitoring [11]. Given the nascency of research in the digital phenotyping field, there is heterogeneity in the clinical use cases, types of data collected, modes of data collection, data analysis methods, and outcomes measured.

Thus, the primary aim of this scoping review was to map the published research on digital phenotyping and to outline study characteristics, methods of active and passive data collection, data analysis approaches used (specifically machine learning techniques, if any), and future implications. The desired outcomes of this review are to provide a broad overview of ongoing research on digital phenotyping and identify gaps and opportunities in future research and practice, especially regarding leveraging machine learning techniques for digital phenotyping.

Methods

Overview

We conducted this scoping review to examine the breadth of published evidence related to digital phenotyping in health care. We utilized an a priori approach for the literature search and data extraction process to ensure the search protocol was replicable. The PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-analyses Extension for Scoping Reviews) checklist guided the methodology and reporting of this scoping review (Multimedia Appendix 1) [12].

Search Terms

As the term "digital phenotype" is relatively nascent in the research landscape, we conducted a preliminary scoping of literature on PubMed and Google Scholar to identify different search terms associated with digital phenotyping. This ensured that our literature search would capture all published research related to digital phenotyping, even if the term was not explicitly mentioned anywhere in the text. These were the search terms finally used to conduct the literature search: "digital phenotyp*" OR "active data" OR "passive data" OR "digital biomarker*" OR "digital footprint" OR "mobile data" OR "mobile phone data" OR "digital sensing" OR "digital fingerprint*" OR "smartphone data" OR "wearable*" OR "wearable device*" OR "wearable data" OR "precision data."

Eligibility Criteria

We included peer-reviewed original research articles in English, as our aim was to explore the gaps and opportunities in scientific research on digital phenotyping. Furthermore, in line with the breakdown of the definition of digital phenotyping by Onnela [3], studies were deemed eligible if they included the following characteristics: (1) if any types of active or passive data were collected. For this review, active data referred to data that required direct input from users in response to prompts, and passive data referred to data generated and collected without inputs from the user [13]; (2) if a wearable device or mobile phone was used to collect the active and/or passive data; (3) if the terms "digital phenotype" or "digital phenotyping" were in the title, abstract, or keywords; and (4) if the active and/or passive data were classified in some ways (ie, if any

"phenotypes" were established or if the data were used to make predictions regarding diagnosis, symptom exacerbation, or relapse).

We limited the years of publication to 2020, 2021, and 2022 because from our preliminary search, we conjectured that these years witnessed a sharp increase in the number of publications related to digital health, active and passive data collection, and wearable devices. Moreover, focusing on these years would provide the most recent snapshot of digital phenotyping research, as the field is rapidly and continually evolving. Table 1 shows the uptick in digital phenotyping research published in the last 5 years. This timeline was the result of using the search terms and article type filters that were part of our eligibility criteria.

We excluded reviews, meta-analyses, opinion pieces, grey literature, letters to the editor, commentaries, study protocols, articles describing phenotyping in the context of genetics, and articles not in English. We also excluded studies that solely focused on the feasibility and acceptability of interventions using digital phenotyping.

Table 1. PubMed timeline of digital phenotyping research published from 2017 to 2022. The timeline indicates a sharp increase in published literature from 2019 onward.

Year	Research articles published, n
2017	129
2018	173
2019	257
2020	246
2021	232
2022	114

Sources of Evidence

We used PubMed and Google Scholar to identify relevant literature. We chose PubMed due to its focus on clinical and health-related research and Google Scholar to surface literature that intersected multiple disciplines.

We utilized additional filters on PubMed to exclude the following articles that did not meet our study type and year of publication criteria: (1) study type: clinical study, clinical trial, comparative study, controlled clinical trial, multicenter study, observational study, randomized controlled trial (RCT); and (2) results by year: from January 1, 2020, to January 18, 2022.

In Google Scholar, we filtered the results according to the date of publication. We used the custom range of 2020-2022.

Screening Process

After applying the search terms and filters on PubMed and Google Scholar to identify relevant articles, the citations were imported into the Rayyan.ai system (Rayyan Systems Inc), a free online tool to create and manage systematic reviews. Author SDD conducted the final search and imported the citations on January 18, 2022. Then, authors SDD and SS independently screened the titles, abstracts, and keywords using the predetermined eligibility criteria. Any discrepancies regarding which articles should be shortlisted were resolved by discussions between SDD and SS. The next step of the screening process involved screening the full texts of these shortlisted articles; all reviewers were randomly assigned articles to screen for concordance with the eligibility criteria. The reviewers had regular discussions to resolve any disagreements on studies to include in the final analysis.

Data Extraction and Charting

After the authors screened the full-text articles for inclusion in the scoping review, a Google Sheet was created to extract descriptive characteristics of the final articles. Details recorded in the Google Sheet included study title, author(s), year of publication, country of origin, study design, clinical area, active and/or passive data collected, mode of data collection, data analysis approaches, and limitations of the study.

The reviewers independently conducted the data extraction and charting of the final articles. SDD and SS were consulted for any queries regarding the data extraction and charting process that the other reviewers had. The results of the data extraction and charting process are presented in Multimedia Appendix 2.

We did not conduct a formal critical appraisal of the final articles because the primary aim of our scoping review was to describe the breadth of evidence and map the characteristics of the literature on digital phenotyping.

Synthesis of Results

We summarized the studies for the following characteristics: countries of origin, study designs, clinical areas, active and/or passive data collected, modes of data collection, data analysis approaches, and limitations. The World Health Organization's region classification was used to group the countries of origin [14]. The study designs were grouped as follows: observational studies, randomized trials, post hoc analyses of observational studies, and post hoc analyses of RCTs.

In this scoping review, we mapped the types of data collected in the studies into the following categories: wearable/activity (passive data), mobile phone (passive data), clinical/biometric (passive data), and active. The passive data categories were based on the Activity-Biometrics-Communication framework by Jayakumar and colleagues [15]. Wearable/activity data included those generated by and collected from wearable devices, mobile phone data included those passively collected from a mobile app or from the mobile device itself (such as the microphone), and clinical/biometric data included passively

collected biological data such as blood pressure, body temperature, heart rate, and so on. Active data included patient-reported outcome measurements on a mobile app, as well as responses to survey questions on a mobile app. We tabulated all the passive and active data points collected in the included studies.

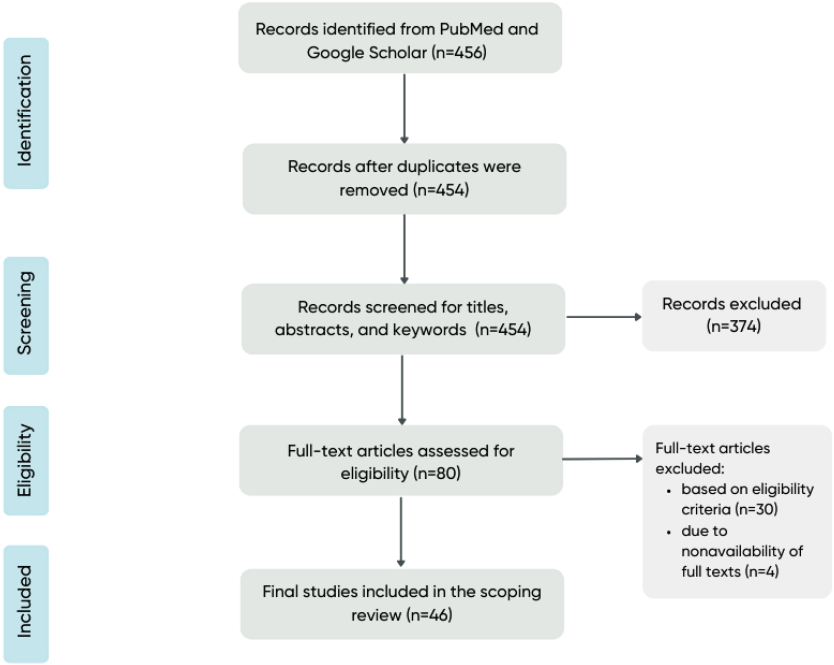
The following categories were used to map how active and passive data were collected in the included studies: wearable device, mobile app, wearable device + mobile app, wearable device + other, and other. We tabulated the wearable devices and mobile apps used in the studies. We used the following broad categories to map the data analysis approaches: regression, statistical methods, machine learning techniques, and latent growth analysis.

Results

Search Results

Figure 1 depicts the PRISMA flowchart of the study selection process. A total of 454 articles were identified from PubMed and Google Scholar after removal of duplicates. Following the screening of the titles, abstracts, and keywords, 80 articles were eligible for full-text review. After reviewing the full-text articles, we excluded 30 that did not meet our eligibility criteria and 4 whose full texts were unavailable. Thus, 46 articles were deemed eligible for inclusion in this scoping review. Detailed characteristics of these 46 articles are presented in Multimedia Appendix 2.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analyses) flowchart of the process of study identification, screening for eligibility, and final inclusion in this scoping review.



Countries of Origin

Most studies (n=26, 56.5%) originated from North America, including the United States (n=24) [16-39] and Canada (n=2) [40,41]. Twelve studies (26.1%) were conducted in European countries, such as France [42,43], Germany [44,45], Italy [46,47], Luxembourg [43], Spain [48,49], Switzerland [50], the Netherlands [48,49], and the United Kingdom [47-49,51-53]. Six studies (13%) originated from countries in the Western Pacific region, including Australia [54,55], Japan [56,57], and South Korea [58,59]. Only 1 study (2.2%) came from the Southeast Asian (China) [60] and Eastern Mediterranean (Qatar) [61] regions. Table 2 summarizes the studies' regions of origin.

Table 2. Summary of the number of studies by the World Health Organization's region classification.

World Health Organization's region classification	Countries of origin	Studies, n (%)
Eastern Mediterranean	Qatar	1 (2.2)
Europe	France, Germany, Italy, Luxembourg, Spain, Switzerland, the Netherlands, and the United Kingdom	12 (26.1)
Southeast Asia	China	1 (2.2)
North America	Canada, the United States	26 (56.5)
Western Pacific	Australia, Japan, South Korea	6 (13)

Study Designs

The most dominant study design was observational (n=28, 60.9%) [17, 20, 21, 23-25, 27, 28, 31, 32, 34, 36-40, 42-47, 49-51, 57, 58, 60], followed by randomized trials (n=10, 21.7%) [19,22,26,30,35,41,52-55], post hoc analyses of RCTs (n=5, 10.9%) [18,29,56,59,61], and post hoc analyses of observational studies (n=3, 6.5%) [16,33,48].

Clinical Areas

The clinical areas investigated in the included studies were heterogeneous. Most (n=15, 32.6%) studies focused on psychiatric disorders, mental health disorders, and neurological diseases, including Parkinson disease [44,51]. Psychiatric and mental health disorders included body dysmorphic disorder [37], disordered eating [54], cognitive impairment [61], substance use disorder [17,31], depression [40,46,48,49,53,60], anxiety disorders [40,53], schizophrenia [23], and stress [26].

A total of 7 (15.2%) studies focused on cardiovascular diseases, which included hypertension [19,21,45], hypercholesterolemia [56], heart failure [24], and general cardiovascular health [32,47]. Five studies (10.9%) focused on cancer, including skin cancer [28], melanoma [34,35], breast cancer [55], and monitoring patients undergoing chemotherapy [27]. Moreover, 3 (6.5%) focused on diabetes [30,38,52], and 7 (15.2%) focused on participants who were overweight or obese [16,18,30,33,36,43,59]. Meanwhile, 4 (8.7%) studies assessed hospital-related outcomes, including postoperative recovery [20], posthospital discharge [22,29], and in-hospital admission of geriatric patients [50]. Three studies (6.5%) included patients undergoing hemodialysis [25,46,61]. Other clinical areas investigated included circadian rhythms [42], cough [57], sarcopenia [58], physical training [39], and rheumatoid arthritis and lupus erythematosus [41].

Types of Active and Passive Data Collected

We categorized the types of data collected in the studies as follows: wearable/activity (passive data), mobile phone (passive data), clinical/biometric (passive data), and active.

Regarding passively collected data, 37 (80.4%) studies evaluated wearable/activity data, 3 (6.5%) studies evaluated mobile phone data, and 13 (28.3%) studies evaluated clinical/biometric data. Nine (19.6%) studies assessed active data. Table 3 summarizes the wearable/activity, mobile phone, clinical/biometric, and active data points collected in the studies.

Table 3. List of the active and passive data points collected in the studies included in this scoping review.

Passive data			Active data
Wearable/activity	Mobile phone	Clinical/biometric	
Mobility pattern [37]	Frequency of app use [37]	Heart rate [17, 19-21, 32, 39, 43, 45, 48, 53, 60]	Exercise amount [54,59]
Ultraviolet radiation exposure [28,34,35]	Quantity of app use [36]	Skin conductance [17]	Body satisfaction [54]
Step count [18-22, 26, 27, 29, 30, 39, 43, 46, 56, 59-61]	Number of days activity monitor data were uploaded to the web-based app [52]	Skin temperature [17]	Fitness/health motives for exercise [54]
Gait parameters [44,51,58]	Call logs [60]	Blood pressure [19,21,43]	Engagement in binge eating [54]
Anticipatory postural adjustments [51]	Text message logs [60]	Movements in epigastric region [57]	Engagement in dietary restraint [54]
Sit-to-stand duration [51]	App usage logs [60]	Expansion of throat skin [57]	Immediate mood [60]
Energy expenditure [39,52]	GPS location [40,60]	Weight [38,43]	Patient Health Questionnaire-9 in an app [60]
Sleep duration [19, 26, 39, 48, 49, 53, 56, 60]	Screen on-and-off status [40,60]	Blood glucose levels [38]	Liebowitz Social Anxiety Scale [40]
Sleep efficiency [19,48,49,53,56]	Ambient audio [40]	N/A[a]	Generalized Anxiety Disorder 7-Item Scale [40]
Sleep stage [56]	Light sensor data [40]	N/A	Patient Health Questionnaire 8-item scale [40,48,49]
Distance walked [45,56]	Telephone call recipient [42]	N/A	Sheehan Disability Scale [40]
Daytime nap duration [24]	Moment in time of telephone call [42]	N/A	Responses to daily assessment [59]
Daytime nap frequency [24]	Telephone call duration [42]	N/A	Meals logged [59]
Repositioning events [36]	Articles read [59]	N/A	Intake of green foods logged [59]
Three-dimensional acceleration [17]	Comments posted [59]	N/A	Rosenberg Self-Esteem Scale [48]
Number of activity monitor wear days across the intervention [52]	Number of posts [59]	N/A	Weigh-ins logged [59]
Number of interactions with wearable sensor [17]	Messages sent to coaches [59]	N/A	Self-reported location [31]
Physical activity [16, 33, 38, 41, 45, 47, 48, 50, 52]	Number of likes [59]	N/A	Self-reported social context [31]
Number of postural transitions [61]	Screen time metrics [24]	N/A	Self-reported cannabis use [31]
Exercise time [59]	N/A	N/A	Mental and physical 5-point scale [39]
Step speed [19]	N/A	N/A	Self-reported sleep, hydration, and nutrition [39]
Time spent walking [16]	N/A	N/A	Confidence in instructors and graduation [39]
Durations of postural transitions [61]	N/A	N/A	Speech patterns [48]
N/A	N/A	N/A	Cognitive function [23,48]

[a]N/A: not applicable.

Modes of Data Collection

The categories used to map how active and passive data were collected in the included studies were as follows: wearable device, mobile app, wearable device + mobile app, wearable device + other, and other. Most (n=25, 54.3%) studies fell under the wearable device category [16-20, 22, 24, 25, 32-34, 36, 38, 43, 44, 46, 47, 49-51, 55-58, 61]. Many (n=14, 30.4%) studies also collected data using a combination of wearable devices and a mobile app and thus fell under the wearable device + mobile app category [21,23,26-30,35,39,45,48,53,54,60]. Of the studies, 8.7% (n=4) fell under the mobile app category [31,37,40,59], 4.4% (n=2) under the wearable device + other category [41,52], and 2.2% (n=1) under the other category [42], which included data collection through web-based applications. Textbox 1 lists the types of wearable devices and mobile apps used in the studies.

Textbox 1. List of wearable devices and mobile apps used to collect active and passive data in the studies included in this scoping review.

Wearable devices:

- Activity monitor (Actical, Philips Respironics) [24]
- activPAL (PAL Technologies Limited) [55]
- Apple Watch Series 2, 3, or 4 smartwatches [21,39,45]
- Biobeam wearable device [53]
- Body weighing scale (Withings) [43]
- BP-800 blood pressure monitor (Withings) [43]
- Cellular-enabled scale [38]
- E4 wearable sensor (Empatica) [17]
- FitBit [16,20,25,26,32,33,38,41,48,49,54,56]
- Garmin Vivofit2 activity monitor [55]
- Inertial SHIMMER sensors (Shimmer Research Limited) [44]
- Mi Band 2 (Xiaomi Corporation) [60]
- Microsoft Band 2 [27]
- Omron Evolv Wireless Blood Pressure Monitor [19,21]
- Phone-tethered glucometer [38]
- Withings pulse activity tracker [43]
- Samsung Galaxy Watch [19]
- SenseWear Mini (BodyMedia) multisensory monitor [41]
- SenseWear Armband [46]
- Shade wearable ultraviolet radiation sensor [28]
- Smartwatch (unspecified) [23]
- Ultraviolet radiation exposure sensor [28,34]
- Validated pendant sensor (PAMSysTM, BioSensics LLC) [61]
- Waist-worn activity tracker (ActiGraph wGT3X-BT) [34]
- Wearable smart belt (WELT) [58]
- Wearable triaxial accelerometer sensor [36]
- Wrist-worn ActiGraph GT3X+ [55]
- Wrist-worn ultraviolet dosimeter [35]
- Wrist-worn wearable device (Withings Activite Steel) [18,22,29,30]

Mobile apps:

- Apple Health app [21]
- Beiwe app [23]
- BreeConnect App [45]
- InstantSurvey smartphone app [54]
- iOS Biobase app [53]
- MApp [31]
- mindLAMP app [23]
- Mood Mirror app [60]
- Noom app (for food diaries) [59]
- Patient-reported outcomes app [27]
- Perspectives app on iOS [37]

- Withings HealthMate app [29]

Data Analysis Approaches

Regarding the data analysis techniques, 22 (47.8%) studies used regression-based statistical methods [16,20,22,23,28,30,33,35,37,40,41,43,45,48-50,53,54,56,58,61], 2 (4.3%) used latent growth analysis [18,38], and 14 (30.4%) used other statistical analysis methods [21,24-26,29,31,32,34,42,44,46,47,52,55]. One (2.2%) study did not perform any statistical analyses because it was a case report [36]. Only 7 (15.2%) studies used machine learning approaches to build predictive models [17,19,39,51,57,59,60], while 1 study used logistic regression and random forest classifiers [51]. Another study tested 25 classification models from the following categories: decision trees, discriminant analysis, logistic regression, naive Bayes classifiers, support vector machines, nearest neighbor classifiers, and ensemble classifiers [17]. One study used 6 different machine learning models: support vector machines, k-nearest neighbors, decision trees, naive Bayes, random forest, and logistic regression [60]. A study conducted in Japan used a deep learning–based machine learning algorithm called variational autoencoder for feature extraction and k-means clustering algorithm for classification [57]. Another study used random forest, support vector machine, gradient boosting decision trees, long short-term memory, and autoregressive integrated moving average techniques [19]. A study from South Korea used an elastic net machine learning approach [59], and 1 from the United States used a random forest approach [39].

Limitations of the Included Studies

The limitations put forward by the authors of the studies in this review were heterogenous. Most studies reported low generalizability of their findings due to small sample size, single-center study designs, short study durations, and narrow population segments included in the studies. Due to the observational nature of the studies, causal relationships between the passive and active data collected and outcome measures could not be confirmed. Some studies also reported device- and app-related limitations, including short battery life of smartwatches (leading to underestimation of physical activity) [21], challenges in keeping the app running 24/7 [60], no measurements of users' interactions with mobile phone notifications [26], missing data [23,30,48,49], and drawbacks in the algorithms tested [16,32,45,57,58]. Another limitation reported was reliance on self-reported data, which included active data collected and those collected for outcome measurements.

Discussion

Principal Findings

Our scoping review provides an insight into the breadth of research on digital phenotyping published in the last 3 years. Most studies originated from North America, had observational study designs, and used wearable devices to collect passive and/or active data. The studies spanned various clinical indications, but psychiatric disorders, mental health disorders, and neurological diseases were the most common areas. Only 7 (15.2%) studies used machine learning–based approaches for data analysis, while the rest predominantly used statistical methods. Most studies had low sample sizes, limiting their generalizability to other populations and clinical settings.

Digital maturity and uptake of wearables vary significantly across regions; however, the onset of the COVID-19 pandemic has generally led to an increase in the use of digital health tools for remote monitoring [62]. In our scoping review, 56.5% (n=26) of the studies were conducted in North America. Market research trends from 2021 indicated that North America is currently leading the global digital health market, and this market is poised to accelerate even faster than the global average between 2021 and 2025 [63]. There is also a significant impact on the pace of transformation from the aftereffects of large-scale enterprise systems implementations. Consumers from this region reported an increase in wearable use from 9% to 33% over the last 4 years, while the number of smartwatch users grew from 42 million to 45.2 million users from 2020 to 2021 and is expected to reach 51.9 million by 2024 [64]. These trends point to greater personalization and innovation in the use of health monitoring tools and wearables in North America. In Europe, the adoption of digital health tools among patients increased from 85% in 2015 to 87% in 2017, with patients increasingly adopting technologies such as wearables and remote patient monitoring tools [65]. The increase in the uptake of digital tools in Europe is attributed to the growing geriatric population coupled with the rising preference for remote patient monitoring. Increasing government initiatives for the development of digital health in the region and growing digital infrastructure will drive market growth [66].

The types of studies in this review were primarily observational (n=28, 60.9%), most of which were cohort-based prospective observational studies. Since wearable device–related studies are relatively new, the rigor and complexity of the study protocols varied significantly, from randomized trials to simple observational studies. We found that digital phenotyping research has been primarily explored in clinical indications related to mental illnesses and psychiatric disorders, but several studies also focused on chronic conditions such as cardiovascular diseases, obesity, and cancer. This points toward growing attention on the real-time monitoring of chronic, long-term conditions, as the patient journeys of these conditions largely occur outside clinical settings.

We observed that the most common data collection tool used across the studies was commercial wearable devices, in line with other reviews conducted in this area [15,67]. Wearable devices have immense potential in both research and disease management due to their ability to collect vast amounts of lifestyle data with high granularity and continuity [19]. While such devices provide a lower barrier to entry, some challenges regarding commercial wearable device use were reported in the studies. For example, one study in our scoping review reported that the short battery lives of smartwatches may have

underestimated physical activity levels [21], and another shortlisted study reported that the Apple Watch could only collect a limited range of heart rate data [39]. Moreover, these devices are associated with data privacy concerns [39]. The "black box" algorithms typically used by most of these devices do not provide clarity on their data collection and analysis practices, leading to inherent biases and subsequent ethical drawbacks when collecting passive data [68].

Although less commonly used in the included studies, smartphone apps are useful in ecological momentary assessments through user-reported, real-time active data. This can help in self-monitoring of behaviors, symptoms, and treatment compliance, as well as in providing information/education and feedback [31]. In their review, Coghlan and D'Alfonso [13] describe a third type of data for digital phenotyping, called interactive data. These can be content-free interactions (such as swiping, tapping, and web searching) or content-rich interactions (such as social media use) [9]. For example, one of the shortlisted studies used interactive data, such as articles read per week, group posts per week, and likes per week, on an app to identify digital behavioral phenotypes of patients with obesity [59]. Such data from a smartphone can provide valuable insights into a user's health status and behaviors, but they are also prone to data privacy concerns and inherent biases.

The use and adoption of newer analytical and machine learning methods for longitudinal data typically collected using wearables are gaining traction in digital health. We found 2 (4.3%) studies using latent class analysis [18,38], which is a statistical procedure used to identify qualitatively different subgroups within populations that share certain outward characteristics. Random forest was most common machine learning technique used [19,39,51,60], followed by logistic regression [17,51,60] and support vector machines [17,19,60]. Random forests work by combining many small, weak decisions for a single strong prediction [6]. This machine learning approach is gaining traction in noncomputational fields and is becoming a standard classification approach in many scientific fields [69]. Random forest algorithms are robust to overfitting, can deal with highly nonlinear data, and remain stable when outliers are present [70]. As 1 of our shortlisted studies reported, although neural network–based approaches outperform in unstructured data such as image and language, tree-based ensemble machine learning models such as random forests have the best performance in structured data that are essentially in tabular form [19]. One study included in our scoping review used and compared a variety of machine learning approaches, including support vector machines, k-nearest neighbors, decision trees, naive Bayes, random forest, and logistic regression; in most cases, the authors found that the random forest method worked the best [60].

Using novel machine learning approaches, passive and active data collected from wearable devices and mobile phones can be used to build "digital phenotypes," enabling the personalization of digital health interventions and treatment plans. These digital phenotypes can be likened to customer segmentation models used by other industries. Better segmentation of health consumer behaviors can play a critical role in our ability to deliver precision digital health interventions. Some studies included in this scoping review established digital phenotypes using the digital data they collected, but these categories were not explicitly called digital phenotypes. For example, 1 study used FitBit data to classify participants into the following physical activity groups: stable active (ie, meeting physical activity recommendations for 2 weeks), stable insufficiently active, stable nonvalid wear, favorable transition (ie, improvements in the physical activity category), and unfavorable transition [33]. Another study used clinical/biometric data from a wearable sensor to develop a cough monitoring system that employed machine learning to distinguish cough and noncough units [57]. Such digital phenotypes can help "close the loop" between monitoring and taking action, helping create adaptive, tailored preventive and treatment journeys [71].

Regular use of wearable technology or behavior-tracking digital health technologies is a valuable intervention in managing health; however, personalized solutions are crucial to users' engagement, as shown by research on the use of wearables in health care [72]. Myneni and colleagues [73] analyzed the behavior change content of a community-based wearable that supports smoking cessation and found evidence from various behavior change theories, including the self-efficacy theory. Other studies examining behavior change technologies that addressed the role of self-efficacy in changing one's behavior proposed the theory of self-efficacy as a key foundation for wearables, suggesting that perceived self-efficacy facilitates the link between intervention and behavior change [72]. Thus, integrating digital phenotyping and wearable device use can improve self-efficacy behaviors, enabling patients and health consumers to take ownership of their health and wellness.

Future Implications

Digital phenotyping shows promise in improving person-centered care. Such precision care can help drive a proactive, predictive approach to health interventions and improved outcomes. Our scoping review highlights the increasing application of statistical and machine learning models on health consumer data from wearable devices. The opportunity to refine digital phenotypes with personal, self-reported data points and real-world passive health information is likely to add value to multiple medical research disciplines and accelerate behavioral health. The success of digital phenotyping is dependent on the willingness of hospitals, physicians, and health care organizations to participate in its development for the benefit of patients and health consumers. Hence, prospective, longitudinal studies that include larger data sets from diverse populations will be important to instill greater confidence in digital phenotyping approaches. Digital phenotyping research has been primarily explored in clinical indications related to mental illnesses and psychiatric disorders. Future work should focus on multivariate, replicable models that link to health outcomes across various indications as well as combine and analyze multiple data sources to provide a more holistic picture of an individual's behaviors and disease state.

Furthermore, given the rapid evolution of privacy concerns affecting consumer technologies, finding ways to ensure data privacy and ethical use of health information should be seen as

a strategic priority not only to understand the boundaries of the type of information that can be used for digital phenotyping but to prioritize systems and checks for health consumer consent and participation. AI and machine learning approaches need to use more transparent, replicable, bias-free algorithms to aid in robust decision making. This is especially important in low- and middle-income contexts, where legal and regulatory frameworks around machine learning deployment in health care may be inadequately defined [74].

Building digital phenotypes has tremendous opportunities in improving the user experience of mobile app–based digital health solutions, helping drive positive health outcomes. Interactive data from a smartphone can be used to generate "engagement phenotypes," and digital journeys can be tailored to each phenotype [71]. Our previous work in machine learning suggests that metrics such as user churn combined with digital phenotyping can help improve user engagement with digital health interventions, thereby potentially leading to better outcomes [75]. Further work needs to be done on the real-world application of machine learning–based models for digital phenotyping in health care settings.

Scoping Review Limitations

Our scoping review may have missed relevant articles because we only used 2 evidence sources (Google Scholar and PubMed) to find articles due to their open-source nature. Because we wanted to capture the breadth of digital phenotyping literature published more recently, we only considered articles published from 2020 onward. However, evidence on digital phenotyping has rapidly grown in the past couple of years. Hence, our scoping review most likely provided an apt snapshot of emerging research on digital phenotyping. For speed, multiple reviewers were involved in screening the full-text articles, which may have led to different interpretations of the results and implications. To help counteract this, we organized frequent discussions among the reviewers to address any concerns about whether a study should be included and reach a consensus. We did not conduct an in-depth citation search of the final articles. Thus, we may have missed relevant articles. Finally, we did not evaluate the quality of the included articles using validated quality assessment checklists. This was mainly due to the heterogeneity of the study characteristics.

Conclusions

Our scoping review provides insightful foundational and application-oriented approaches toward digital phenotyping, including the use of active and passive data, differences in study design, and perhaps most importantly, the growing use of newer data analytics and machine learning algorithms to define and implement digital phenotypes in health care. Future work should focus on conducting longitudinal studies with diverse populations and larger data sets from multiple sources, leveraging newer machine learning approaches for digital phenotyping, addressing privacy and ethical concerns around passive data collection from commercial wearable devices and smartphones, and building digital phenotypes to tailor treatment plans and digital health interventions.

Acknowledgments

We thank our colleagues Anjali Dhingra and Cheryl Gonsalves at Saathealth for their contribution in the data extraction and charting process in this scoping review.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-analyses Extension for Scoping Reviews) checklist.
[DOCX File , 21 KB - bioinform_v3i1e39618_app1.docx]

Multimedia Appendix 2
Results of the data extraction and charting process of the final studies included in the scoping review.
[XLSX File (Microsoft Excel File), 21 KB - bioinform_v3i1e39618_app2.xlsx]

References

1. Snowdon AW, Alessi C, Bassi H, DeForge RT, Schnarr K. Enhancing patient experience through personalization of health services. Healthc Manage Forum 2015 Sep 01;28(5):182-185. [doi: 10.1177/0840470415588656] [Medline: 26135292]
2. Liu K, Tao D. The roles of trust, personalization, loss of privacy, and anthropomorphism in public acceptance of smart healthcare services. Comput Hum Behav 2022 Feb;127:107026. [doi: 10.1016/j.chb.2021.107026]
3. Onnela J. Opportunities and challenges in the collection and analysis of digital phenotyping data. Neuropsychopharmacology 2021 Jan;46(1):45-54 [FREE Full text] [doi: 10.1038/s41386-020-0771-3] [Medline: 32679583]
4. de Angel V, Lewis S, White K, Oetzmann C, Leightley D, Oprea E, et al. Digital health tools for the passive monitoring of depression: a systematic review of methods. NPJ Digit Med 2022 Jan 11;5(1):3 [FREE Full text] [doi: 10.1038/s41746-021-00548-8] [Medline: 35017634]

5. Perez-Pozuelo I, Spathis D, Clifton E, Mascolo C. Wearables, smartphones, and artificial intelligence for digital phenotyping and health. In: Syed-Abdul S, Zhu X, Fernandez-Luque L, editors. Digital Health: Mobile and Wearable Devices for Participatory Health Applications. Amsterdam, the Netherlands: Elsevier; 2020:33-54.
6. Carmel S. Data talking to machines: the intersection of deep phenotyping and artificial intelligence internet. Ethical, Legal, and Social Implications of Deep Phenotyping Symposium.: Harvard Law; 2021 Jul 27. URL: https://blog.petrieflom.law.harvard.edu/2021/01/27/deep-phenotyping-artificial-intelligence/ [accessed 2022-02-25]
7. Maher NA, Senders JT, Hulsbergen AF, Lamba N, Parker M, Onnela J, et al. Passive data collection and use in healthcare: A systematic review of ethical issues. Int J Med Inform 2019 Sep;129:242-247. [doi: 10.1016/j.ijmedinf.2019.06.015] [Medline: 31445262]
8. Benoit J, Onyeaka H, Keshavan M, Torous J. Systematic review of digital phenotyping and machine learning in psychosis spectrum illnesses. Harv Rev Psychiatry 2020;28(5):296-304. [doi: 10.1097/HRP.0000000000000268] [Medline: 32796192]
9. Martinez-Martin N, Insel TR, Dagum P, Greely HT, Cho MK. Data mining for health: staking out the ethical territory of digital phenotyping. NPJ Digit Med 2018 Dec 19;1(1):1-10 [FREE Full text] [doi: 10.1038/s41746-018-0075-8] [Medline: 31211249]
10. Mendes JPM, Moura IR, Van de Ven P, Viana D, Silva FJS, Coutinho LR, et al. Sensing apps and public data sets for digital phenotyping of mental health: systematic review. J Med Internet Res 2022 Feb 17;24(2):e28735 [FREE Full text] [doi: 10.2196/28735] [Medline: 35175202]
11. Spinazze P, Rykov Y, Bottle A, Car J. Digital phenotyping for assessment and prediction of mental health outcomes: a scoping review protocol. BMJ Open 2019 Dec 30;9(12):e032255 [FREE Full text] [doi: 10.1136/bmjopen-2019-032255] [Medline: 31892655]
12. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med 2018 Sep 04;169(7):467. [doi: 10.7326/M18-0850]
13. Coghlan S, D'Alfonso S. Digital phenotyping: an epistemic and methodological analysis. Philos Technol 2021 Nov 11;34(4):1905-1928 [FREE Full text] [doi: 10.1007/s13347-021-00492-1] [Medline: 34786325]
14. Countries. World Health Organization. 2022. URL: https://www.who.int/countries [accessed 2022-02-25]
15. Jayakumar P, Lin E, Galea V, Mathew AJ, Panda N, Vetter I, et al. Digital phenotyping and patient-generated health data for outcome measurement in surgical care: a scoping review. J Pers Med 2020 Dec 15;10(4):282 [FREE Full text] [doi: 10.3390/jpm10040282] [Medline: 33333915]
16. Beagle AJ, Tison GH, Aschbacher K, Olgin JE, Marcus GM, Pletcher MJ. Comparison of the physical activity measured by a consumer wearable activity tracker and that measured by self-report: cross-sectional analysis of the health eheart study. JMIR Mhealth Uhealth 2020 Dec 29;8(12):e22090 [FREE Full text] [doi: 10.2196/22090] [Medline: 33372896]
17. Carreiro S, Chintha KK, Shrestha S, Chapman B, Smelson D, Indic P. Wearable sensor-based detection of stress and craving in patients during treatment for substance use disorder: A mixed methods pilot study. Drug Alcohol Depend 2020 Apr 01;209:107929 [FREE Full text] [doi: 10.1016/j.drugalcdep.2020.107929] [Medline: 32193048]
18. Chen XS, Changolkar S, Navathe AS, Linn KA, Reh G, Szwartz G, et al. Association between behavioral phenotypes and response to a physical activity intervention using gamification and social incentives: Secondary analysis of the STEP UP randomized clinical trial. PLoS One 2020;15(10):e0239288 [FREE Full text] [doi: 10.1371/journal.pone.0239288] [Medline: 33052906]
19. Chiang P, Wong M, Dey S. Using wearables and machine learning to enable personalized lifestyle recommendations to improve blood pressure. IEEE J Transl Eng Health Med 2021;9:1-13. [doi: 10.1109/jtehm.2021.3098173]
20. de Boer C, Ghomrawi H, Many B, Bouchard ME, Linton S, Figueroa A, et al. Utility of wearable sensors to assess postoperative recovery in pediatric patients after appendectomy. J Surg Res 2021 Jul;263:160-166. [doi: 10.1016/j.jss.2021.01.030] [Medline: 33667871]
21. Golbus JR, Pescatore NA, Nallamothu BK, Shah N, Kheterpal S. Lancet Digit 2021 Nov;3(11):e707-e715 [FREE Full text] [doi: 10.1016/S2589-7500(21)00138-2] [Medline: 34711377]
22. Greysen SR, Changolkar S, Small DS, Reale C, Rareshide CAL, Mercede A, et al. Effect of behaviorally designed gamification with a social support partner to increase mobility after hospital discharge: a randomized clinical trial. JAMA Netw Open 2021 Mar 01;4(3):e210952 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.0952] [Medline: 33760089]
23. Henson P, Rodriguez-Villa E, Torous J. Investigating associations between screen time and symptomatology in individuals with serious mental illness: longitudinal observational study. J Med Internet Res 2021 Mar 10;23(3):e23144 [FREE Full text] [doi: 10.2196/23144] [Medline: 33688835]
24. Li P, Gaba A, Wong PM, Cui L, Yu L, Bennett DA, et al. Objective assessment of daytime napping and incident heart failure in 1140 community-dwelling older adults: a prospective, observational cohort study. J Am Heart Assoc 2021 Jun 15;10(12):e019037 [FREE Full text] [doi: 10.1161/JAHA.120.019037] [Medline: 34075783]
25. Malhotra R, Kumar U, Virgen P, Magallon B, Garimella PS, Chopra T, et al. Physical activity in hemodialysis patients on nondialysis and dialysis days: Prospective observational study. Hemodial Int 2021 Apr;25(2):240-248. [doi: 10.1111/hdi.12913] [Medline: 33650200]

26. NeCamp T, Sen S, Frank E, Walton MA, Ionides EL, Fang Y, et al. Assessing real-time moderation for developing adaptive mobile health interventions for medical interns: micro-randomized trial. J Med Internet Res 2020 Mar 31;22(3):e15033 [FREE Full text] [doi: 10.2196/15033] [Medline: 32229469]
27. Nilanon T, Nocera LP, Martin AS, Kolatkar A, May M, Hasnain Z, et al. Use of wearable activity tracker in patients with cancer undergoing chemotherapy: toward evaluating risk of unplanned health care encounters. JCO Clin Cancer Inform 2020 Nov(4):839-853. [doi: 10.1200/cci.20.00023]
28. Parsons BG, Nagelhout ES, Wankier AP, Hu N, Lensink R, Zhu A, et al. Reactivity to UV radiation exposure monitoring using personal exposure devices for skin cancer prevention: longitudinal observational study. JMIR Mhealth Uhealth 2021 Sep 28;9(9):e29694 [FREE Full text] [doi: 10.2196/29694] [Medline: 34581683]
29. Patel MS, Polsky D, Kennedy EH, Small DS, Evans CN, Rareshide CAL, et al. Smartphones vs wearable devices for remotely monitoring physical activity after hospital discharge: a secondary analysis of a randomized clinical trial. JAMA Netw Open 2020 Feb 05;3(2):e1920677 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.20677] [Medline: 32031643]
30. Patel MS, Small DS, Harrison JD, Hilbert V, Fortunato MP, Oon AL, et al. Effect of behaviorally designed gamification with social incentives on lifestyle modification among adults with uncontrolled diabetes: a randomized clinical trial. JAMA Netw Open 2021 May 03;4(5):e2110255 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.10255] [Medline: 34028550]
31. Prince MA, Collins RL, Wilson SD, Vincent PC. A preliminary test of a brief intervention to lessen young adults' cannabis use: Episode-level smartphone data highlights the role of protective behavioral strategies and exercise. Exp Clin Psychopharmacol 2020 Apr;28(2):150-156 [FREE Full text] [doi: 10.1037/pha0000301] [Medline: 31144836]
32. Quer G, Gouda P, Galarnyk M, Topol EJ, Steinhubl SR. Inter- and intraindividual variability in daily resting heart rate and its associations with age, sex, sleep, BMI, and time of year: retrospective, longitudinal cohort study of 92,457 adults. PLoS One 2020 Feb 5;15(2):e0227709 [FREE Full text] [doi: 10.1371/journal.pone.0227709] [Medline: 32023264]
33. Robertson MC, Green CE, Liao Y, Durand CP, Basen-Engquist KM. Self-efficacy and physical activity in overweight and obese adults participating in a worksite weight loss intervention: multistate modeling of wearable device data. Cancer Epidemiol Biomarkers Prev 2020 Dec 23;29(4):769-776. [doi: 10.1158/1055-9965.epi-19-0907]
34. Robinson JK, Durst DA, Gray E, Kwasny M. Protection-adjusted UV dose estimated for body areas: Daily self-reported sun protection modification of wearable UV sensor dose. Photodermatol Photoimmunol Photomed 2020 Sep 28;36(5):357-364. [doi: 10.1111/phpp.12557] [Medline: 32189399]
35. Robinson JK, Durst DA, Gray E, Kwasny M, Heo SY, Banks A, et al. Sun exposure reduction by melanoma survivors with wearable sensor providing real-time UV exposure and daily text messages with structured goal setting. Arch Dermatol Res 2021 Oct 13;313(8):685-694. [doi: 10.1007/s00403-020-02163-1] [Medline: 33185716]
36. Sabol VK, Kennerly SM, Alderden J, Horn SD, Yap TL. Insight into the movement behaviors of nursing home residents living with obesity: a report of two cases. Wound Manag Prev 2020 May 6;66(5):18-29. [doi: 10.25270/wmp.2020.5.1829]
37. Weingarden H, Matic A, Calleja RG, Greenberg JL, Harrison O, Wilhelm S. Optimizing smartphone-delivered cognitive behavioral therapy for body dysmorphic disorder using passive smartphone data: initial insights from an open pilot trial. JMIR Mhealth Uhealth 2020 Jun 18;8(6):e16350 [FREE Full text] [doi: 10.2196/16350] [Medline: 32554382]
38. Yang Q, Hatch D, Crowley MJ, Lewinski AA, Vaughn J, Steinberg D, et al. Digital phenotyping self-monitoring behaviors for individuals with type 2 diabetes mellitus: observational study using latent class growth analysis. JMIR Mhealth Uhealth 2020 Jun 11;8(6):e17730 [FREE Full text] [doi: 10.2196/17730] [Medline: 32525492]
39. Saxon L, DiPaula B, Fox GR, Ebert R, Duhaime J, Nocera L, et al. Continuous measurement of reconnaissance marines in training with custom smartphone app and watch: observational cohort study. JMIR Mhealth Uhealth 2020 Jun 15;8(6):e14116 [FREE Full text] [doi: 10.2196/14116] [Medline: 32348252]
40. di Matteo D, Fotinos K, Lokuge S, Mason G, Sternat T, Katzman MA, et al. Automated screening for social anxiety, generalized anxiety, and depression from objective smartphone-collected data: cross-sectional study. J Med Internet Res 2021 Aug 13;23(8):e28918 [FREE Full text] [doi: 10.2196/28918] [Medline: 34397386]
41. Li LC, Feehan LM, Xie H, Lu N, Shaw C, Gromala D, et al. Efficacy of a physical activity counseling program with use of a wearable tracker in people with inflammatory arthritis: a randomized controlled trial. Arthritis Care Res (Hoboken) 2020 Dec 27;72(12):1755-1765. [doi: 10.1002/acr.24199] [Medline: 32248626]
42. Aubourg T, Demongeot J, Provost H, Vuillerme N. Circadian rhythms in the telephone calls of older adults: observational descriptive study. JMIR Mhealth Uhealth 2020 Feb 25;8(2):e12452 [FREE Full text] [doi: 10.2196/12452] [Medline: 32130156]
43. el Fatouhi D, Delrieu L, Goetzinger C, Malisoux L, Affret A, Campo D, et al. Associations of physical activity level and variability with 6-month weight change among 26,935 users of connected devices: observational real-life study. JMIR Mhealth Uhealth 2021 Apr 15;9(4):e25385 [FREE Full text] [doi: 10.2196/25385] [Medline: 33856352]
44. Gaßner H, Sanders P, Dietrich A, Marxreiter F, Eskofier BM, Winkler J, et al. Clinical relevance of standardized mobile gait tests. Reliability analysis between gait recordings at hospital and home in Parkinson's disease: a pilot study. JPD 2020 Oct 27;10(4):1763-1773. [doi: 10.3233/jpd-202129]
45. Stollfuss B, Richter M, Drömann D, Klose H, Schwaiblmair M, Gruenig E, et al. Digital tracking of physical activity, heart rate, and inhalation behavior in patients with pulmonary arterial hypertension treated with inhaled iloprost: observational

study (VENTASTEP). J Med Internet Res 2021 Oct 08;23(10):e25163 [FREE Full text] [doi: 10.2196/25163] [Medline: 34623313]
46. Brys ADH, Bossola M, Lenaert B, Biamonte F, Gambaro G, Di Stasio E. Daily physical activity in patients on chronic haemodialysis and its relation with fatigue and depressive symptoms. Int Urol Nephrol 2020 Jul 28;52(10):1959-1967. [doi: 10.1007/s11255-020-02578-9]
47. Moscarelli M, Lorusso R, Abdullahi Y, Varone E, Marotta M, Solinas M, et al. The effect of minimally invasive surgery and sternotomy on physical activity and quality of life. Heart Lung Circ 2021 Jun;30(6):882-887. [doi: 10.1016/j.hlc.2020.09.936] [Medline: 33191139]
48. Leightley D, Lavelle G, White KM, Sun S, Matcham F, Ivan A, RADAR-CNS Consortium. Investigating the impact of COVID-19 lockdown on adults with a recent history of recurrent major depressive disorder: a multi-Centre study using remote measurement technology. BMC Psychiatry 2021 Sep 06;21(1):435 [FREE Full text] [doi: 10.1186/s12888-021-03434-5] [Medline: 34488697]
49. Zhang Y, Folarin AA, Sun S, Cummins N, Bendayan R, Ranjan Y, RADAR-CNS Consortium. Relationship between major depression symptom severity and sleep collected using a wristband wearable device: multicenter longitudinal observational study. JMIR Mhealth Uhealth 2021 Apr 12;9(4):e24604 [FREE Full text] [doi: 10.2196/24604] [Medline: 33843591]
50. Tasheva P, Kraege V, Vollenweider P, Roulet G, Méan M, Marques-Vidal P. Accelerometry assessed physical activity of older adults hospitalized with acute medical illness - an observational study. BMC Geriatr 2020 Oct 02;20(1):382 [FREE Full text] [doi: 10.1186/s12877-020-01763-w] [Medline: 33008378]
51. de Vos M, Prince J, Buchanan T, FitzGerald JJ, Antoniades CA. Discriminating progressive supranuclear palsy from Parkinson's disease using wearable technology and machine learning. Gait Posture 2020 Mar;77:257-263. [doi: 10.1016/j.gaitpost.2020.02.007] [Medline: 32078894]
52. Peacock OJ, Western MJ, Batterham AM, Chowdhury EA, Stathi A, Standage M, et al. Effect of novel technology-enabled multidimensional physical activity feedback in primary care patients at risk of chronic disease - the MIPACT study: a randomised controlled trial. Int J Behav Nutr Phys Act 2020 Aug 08;17(1):99 [FREE Full text] [doi: 10.1186/s12966-020-00998-5] [Medline: 32771018]
53. Ponzo S, Morelli D, Kawadler JM, Hemmings NR, Bird G, Plans D. Efficacy of the digital therapeutic mobile app Biobase to reduce stress and improve mental well-being among university students: randomized controlled trial. JMIR Mhealth Uhealth 2020 Apr 06;8(4):e17767 [FREE Full text] [doi: 10.2196/17767] [Medline: 31926063]
54. Gittus M, Fuller-Tyszkiewicz M, Brown HE, Richardson B, Fassnacht DB, Lennard GR, et al. Are Fitbits implicated in body image concerns and disordered eating in women? Health Psychol 2020 Oct;39(10):900-904. [doi: 10.1037/hea0000881] [Medline: 32406725]
55. Nguyen NH, Vallance JK, Buman MP, Moore MM, Reeves MM, Rosenberg DE, et al. Effects of a wearable technology-based physical activity intervention on sleep quality in breast cancer survivors: the ACTIVATE Trial. J Cancer Surviv 2021 Apr 01;15(2):273-280. [doi: 10.1007/s11764-020-00930-7] [Medline: 32875536]
56. Meguro K, Svensson T, Chung U, Svensson AK. Associations of work-related stress and total sleep time with cholesterol levels in an occupational cohort of Japanese office workers. J Occup Health 2021 Jan;63(1):e12275 [FREE Full text] [doi: 10.1002/1348-9585.12275] [Medline: 34679211]
57. Otoshi T, Nagano T, Izumi S, Hazama D, Katsurada N, Yamamoto M, et al. A novel automatic cough frequency monitoring system combining a triaxial accelerometer and a stretchable strain sensor. Sci Rep 2021 May 11;11(1):9973 [FREE Full text] [doi: 10.1038/s41598-021-89457-0] [Medline: 33976286]
58. Kang M, Kang S, Roh H, Jung H, Kim S, Choi J, et al. Accuracy and diversity of wearable device-based gait speed measurement among older men: observational study. J Med Internet Res 2021 Oct 11;23(10):e29884 [FREE Full text] [doi: 10.2196/29884] [Medline: 34633293]
59. Kim M, Yang J, Ahn W, Choi HJ. Machine learning analysis to identify digital behavioral phenotypes for engagement and health outcome efficacy of an mHealth intervention for obesity: randomized controlled trial. J Med Internet Res 2021 Jun 24;23(6):e27218 [FREE Full text] [doi: 10.2196/27218] [Medline: 34184991]
60. Bai R, Xiao L, Guo Y, Zhu X, Li N, Wang Y, et al. Tracking and monitoring mood stability of patients with major depressive disorder by machine learning models using passive digital data: prospective naturalistic multicenter study. JMIR Mhealth Uhealth 2021 Mar 08;9(3):e24365 [FREE Full text] [doi: 10.2196/24365] [Medline: 33683207]
61. Zhou H, Al-Ali F, Wang C, Hamad A, Ibrahim R, Talal T, et al. Harnessing digital health to objectively assess cognitive impairment in people undergoing hemodialysis process: The Impact of cognitive impairment on mobility performance measured by wearables. PLoS One 2020 Apr 20;15(4):e0225358 [FREE Full text] [doi: 10.1371/journal.pone.0225358] [Medline: 32310944]
62. Negreiro M. The rise of digital health technologies during the pandemic. European Parliament. 2021 Apr. URL: https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/690548/EPRS_BRI(2021)690548_EN.pdf [accessed 2022-02-25]
63. Global Digital Health Market and Trends Report 2021: prompted by COVID-19, online health care is gaining momentum. Research and Markets. 2021 Dec 21. URL: https://tinyurl.com/4ptb6mkb [accessed 2022-02-25]
64. Samet A. The top medical monitoring and health care wearable device trends of 2022. Insider Intelligence. 2022 Feb 03. URL: https://www.insiderintelligence.com/insights/top-healthcare-wearable-technology-trends [accessed 2022-02-25]

65. Europe mHealth market: top 4 trends boosting the industry demand through 2026. BioSpace. 2021 Feb 16. URL: https://www.biospace.com/article/europe-mhealth-market-top-4-trends-boosting-the-industry-demand-through-2026-/ [accessed 2022-02-25]
66. Europe Digital Health Market Forecast 2027.: Graphical Research; 2021 Aug. URL: https://www.graphicalresearch.com/industry-insights/1163/europe-digital-health-market [accessed 2022-02-25]
67. Perez-Pozuelo I, Spathis D, Gifford-Moore J, Morley J, Cowls J. Digital phenotyping and sensitive health data: Implications for data governance. J Am Med Inform Assoc 2021 Aug 13;28(9):2002-2008 [FREE Full text] [doi: 10.1093/jamia/ocab012] [Medline: 33647989]
68. Kilgallon JL, Tewarie IA, Broekman MLD, Rana A, Smith TR. Passive data use for ethical digital public health surveillance in a postpandemic world. J Med Internet Res 2022 Feb 15;24(2):e30524 [FREE Full text] [doi: 10.2196/30524] [Medline: 35166676]
69. Couronné R, Probst P, Boulesteix A. Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics 2018 Jul 17;19(1):270 [FREE Full text] [doi: 10.1186/s12859-018-2264-5] [Medline: 30016950]
70. Sarica A, Cerasa A, Quattrone A. Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: a systematic review. Front Aging Neurosci 2017 Oct 06;9:329 [FREE Full text] [doi: 10.3389/fnagi.2017.00329] [Medline: 29056906]
71. Huckvale K, Venkatesh S, Christensen H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. NPJ Digit Med 2019 Sep 6;2(1):88 [FREE Full text] [doi: 10.1038/s41746-019-0166-1] [Medline: 31508498]
72. Rieder A, Eseryel UY, Lehrer C, Jung R. Why users comply with wearables: the role of contextual self-efficacy in behavioral change. Int J Hum-Comput Interact 2020 Sep 30;37(3):281-294. [doi: 10.1080/10447318.2020.1819669]
73. Myneni S, Cobb N, Cohen T. In pursuit of theoretical ground in behavior change support systems: analysis of peer-to-peer communication in a health-related online community. J Med Internet Res 2016 Feb 02;18(2):e28 [FREE Full text] [doi: 10.2196/jmir.4671] [Medline: 26839162]
74. Fletcher RR, Nakeshimana A, Olubeko O. Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health. Front Artif Intell 2021;3:561802 [FREE Full text] [doi: 10.3389/frai.2020.561802] [Medline: 33981989]
75. Ganju A, Satyan S, Tanna V, Menezes SR. AI for improving children's health: a community case study. Front Artif Intell 2020 Jan 6;3:544972 [FREE Full text] [doi: 10.3389/frai.2020.544972] [Medline: 33733204]

Abbreviations

AI: artificial intelligence
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-analyses Extension for Scoping Reviews
RCT: randomized controlled trial

Edited by A Mavragani; submitted 16.05.22; peer-reviewed by R Rastmanesh, I Mircheva; comments to author 16.06.22; revised version received 01.07.22; accepted 04.07.22; published 18.07.22.

Please cite as:
Dlima SD, Shevade S, Menezes SR, Ganju A
Digital Phenotyping in Health Using Machine Learning Approaches: Scoping Review
JMIR Bioinform Biotech 2022;3(1):e39618
URL: https://bioinform.jmir.org/2022/1/e39618
doi:10.2196/39618
PMID:

Review

The Utilization of Heart Rate Variability for Autonomic Nervous System Assessment in Healthy Pregnant Women: Systematic Review

Zahra Sharifiheris[1*], BSc, MSc; Amir Rahmani[1*], BSc, MSc, PhD; Joseph Onwuka[1*], BSN; Miriam Bender[1*], BSc, PhD

University of California, Irvine, Irvine, CA, United States

[*]all authors contributed equally

Corresponding Author:
Zahra Sharifiheris, BSc, MSc
University of California, Irvine
8420 Palo Verde
Irvine, CA, 92697
United States
Phone: 1 6506805432
Email: sharifiz@uci.edu

Abstract

Background: The autonomic nervous system (ANS) plays a central role in pregnancy-induced adaptations, and failure in the required adaptations is associated with adverse neonatal and maternal outcomes. Mapping maternal ANS function in healthy pregnancy may help to understand ANS function.

Objective: This study aimed to systematically review studies on the use of heart rate variability (HRV) monitoring to measure ANS function during pregnancy and determine whether specific HRV patterns representing normal ANS function have been identified during pregnancy.

Methods: The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guideline was used to guide the systematic review. The CINAHL, PubMed, SCOPUS, and Web of Science databases were searched to comprehensively identify articles without a time span limitation. Studies were included if they assessed HRV in healthy pregnant individuals at least once during pregnancy or labor, with or without a comparison group (eg, complicated pregnancy). Quality assessment of the included literature was performed using the National Heart, Lung, and Blood Institute (NHLBI) tool. A narrative synthesis approach was used for data extraction and analysis, as the articles were heterogenous in scope, approaches, methods, and variables assessed, which precluded traditional meta-analysis approaches being used.

Results: After full screening, 8 studies met the inclusion criteria. In 88% (7/8) of the studies, HRV was measured using electrocardiogram and operationalized in 3 different ways: linear frequency domain (FD), linear time domain (TD), and nonlinear methods. FD was measured in all (8/8), TD in 75% (6/8), and nonlinear methods in 25% (2/8) of the studies. The assessment duration varied from 5 minutes to 24 hours. TD indexes and most of the FD indexes decreased from the first to the third trimesters in the majority (5/7, 71%) of the studies. Of the FD indexes, low frequency (LF [nu]) and the LF/high frequency (HF) ratio showed an ascending trend from early to late pregnancy, indicating an increase in sympathetic activity toward the end of the pregnancy.

Conclusions: We identified 3 HRV operationalization methods along with potentially indicative HRV patterns. However, we found no justification for the selection of measurement tools, measurement time frames, and operationalization methods, which threaten the generalizability and reliability of pattern findings. More research is needed to determine the criteria and methods for determining HRV patterns corresponding to ANS functioning in healthy pregnant persons.

(JMIR Bioinform Biotech 2022;3(1):e36791) doi:10.2196/36791

KEYWORDS

heart rate variability; pregnancy; systematic review; autonomic nervous system assessment

Introduction

The autonomic nervous system (ANS) is one of the central regulatory systems that responds to various internal and external stresses [1]. Pregnancy is one of the stimuli that requires various physiological changes in order to adapt to relevant demands regarding fetal development and thus needs ANS regulatory function [2,3]. Systemic vasodilation is the primary initial pregnancy-related event [4-6]. The outcome of systemic vasodilation is a series of systemic accommodations that involve almost all the body systems including respiratory, cardiovascular, digestion, and endocrine systems [7]. The combination of these systemic accommodations is known as "homeostasis," which is a dynamic and complex function [8]. Establishing homeostasis during pregnancy is necessary to promote embryotic and fetal growth. Due to the dynamic nature of pregnancy, it is critical to understand whether the corresponding dynamic ANS function results in a certain pattern of changes reflecting healthy accommodation in ways that can be observed and acted upon. However, the safeguard for the amount, severity, and pattern of safe changes in ANS function is not well known, in part because it is impractical to continuously observe the function in vivo using existing methods.

The traditional tests for ANS assessment are those that evaluate the cardiovascular reflexes in response to provocative maneuvers [9,10]. Although these ANS assessment maneuvers are widely applied in clinical settings for diagnostic purposes, the ability of these maneuvers to reflect ANS function in real life is not well-justified due to the fact that the tests are often performed in controlled situations and artificial settings such as laboratories and hospitals that intrinsically can affect ANS function during the assessment. Additionally, due to the dynamic nature of pregnancy-related accommodations, episodic-only assessments are insufficient to capture the dynamism in ANS function during pregnancy. Thus, to assess this dynamism, more in vivo assessment techniques in real time are needed.

Heart rate variability (HRV), defined as a variation in the beat-to-beat (RR or NN) interval, is a well-known, noninvasive assessment tool for ANS that has been recently applied widely for both clinical and nonclinical purposes [11]. A study performed with 8 million individuals indicated that HRV can vary by age, sex, and activity [12]. However, less is known regarding how pregnancy may affect HRV, to understand ANS regulations induced in pregnancy. This is especially important to investigate as various studies have indicated that ANS dysregulation assessed through HRV can be associated with common pregnancy complications including hypertensive disorders and gestational diabetes [13-16]. Thus, we aimed to review studies that have assessed HRV for ANS regulations during noncomplicated pregnancies to answer the following questions: (1) whether and how HRV has been used to measure ANS function during heathy pregnancy and (2) whether any specific HRV patterns have been identified during pregnancy.

Methods

Design

We conducted a systematic review using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) standards [17] to guide the study. The population, exposure, comparator, and outcome (PECO) framework was used to develop the research question and search terms. The research question was: Has HRV monitoring (exposure) been used to measure ANS function (outcome) during pregnancy (population), and if so, what specific HRV patterns (comparison) representing normal ANS function have been identified during pregnancy?

Information Sources

The PubMed, CINAHL, SCOPUS, and Web of Science databases were searched initially in August 2020 and updated in June 2021 (past June 2021). Although we applied no limitation to the time span for the search, the time span varied for each database depending on the publication history of each database. See Multimedia Appendix 1 for more details.

Search Strategy

To access further studies, additional sources were reviewed, including reference lists of the included articles and Google Scholar. Keywords including "Heart Rate Variability (HRV)" and "Pregnancy" were used for both simple and advanced searches of each database separately (see Multimedia Appendix 1 for all terms and search strategies used for each database).

Inclusion and Exclusion Criteria

The population included healthy pregnant individuals.

Studies that involved various interventions (eg, exercise) with healthy pregnant women were excluded. Being pregnant was considered as the exposure (E) component, which was required for all the studies. Studies with or without a comparison (C) group (eg, complicated pregnancy) were eligible for inclusion. HRV, assessed at least once during pregnancy or labor, was considered the expected outcome (O) for all the studies that were assessed. Studies were included if available in the English language. Exclusion criteria were systematic reviews, protocols, conference proceedings, letters to the editor, unpublished or under review papers, and dissertation proposals.

Selection and Data Collection Process

Selected articles were peer reviewed in Covidence online software by 2 independent reviewers. To assess the relevancy, all the studies were screened by both reviewers, ZS and JO, based on titles, abstracts, and full text in 2 steps. In the first step, the abstracts of all the articles that were gathered from the databases were screened in terms of their relevance to our study aim. Next, those articles with relevant titles or abstracts from the first step underwent a full-text assessment. To resolve disagreements, a third reviewer, MB, was involved.

Data Items

We collected the following data: HRV results, as the main outcome; assessment tools used to measure HRV; HRV

component(s); frequency and duration of the assessment; and gestational age at the assessment.

Effect Measures

The effect measure for all the studies was the mean difference, and a significant difference was considered at a *P* value <.05.

Risk of Bias Assessment

Two independent reviewers, ZS and JO, assessed the methodological quality of the selected studies using the National Heart, Lung, and Blood Institute (NHLBI) Quality Assessment scale for observational cohort and cross-sectional studies [18]. The NHLBI quality assessment tool, consisting of 14 questions, assesses studies in terms of the following criteria: study objectives, study population, sample size, exposure, outcome measures, and key potential confounding variables. Each study was assessed for a risk of bias using responses of "yes," "no," and "cannot determine/not applicable/not reported" for every single criterion.

Synthesis Methods

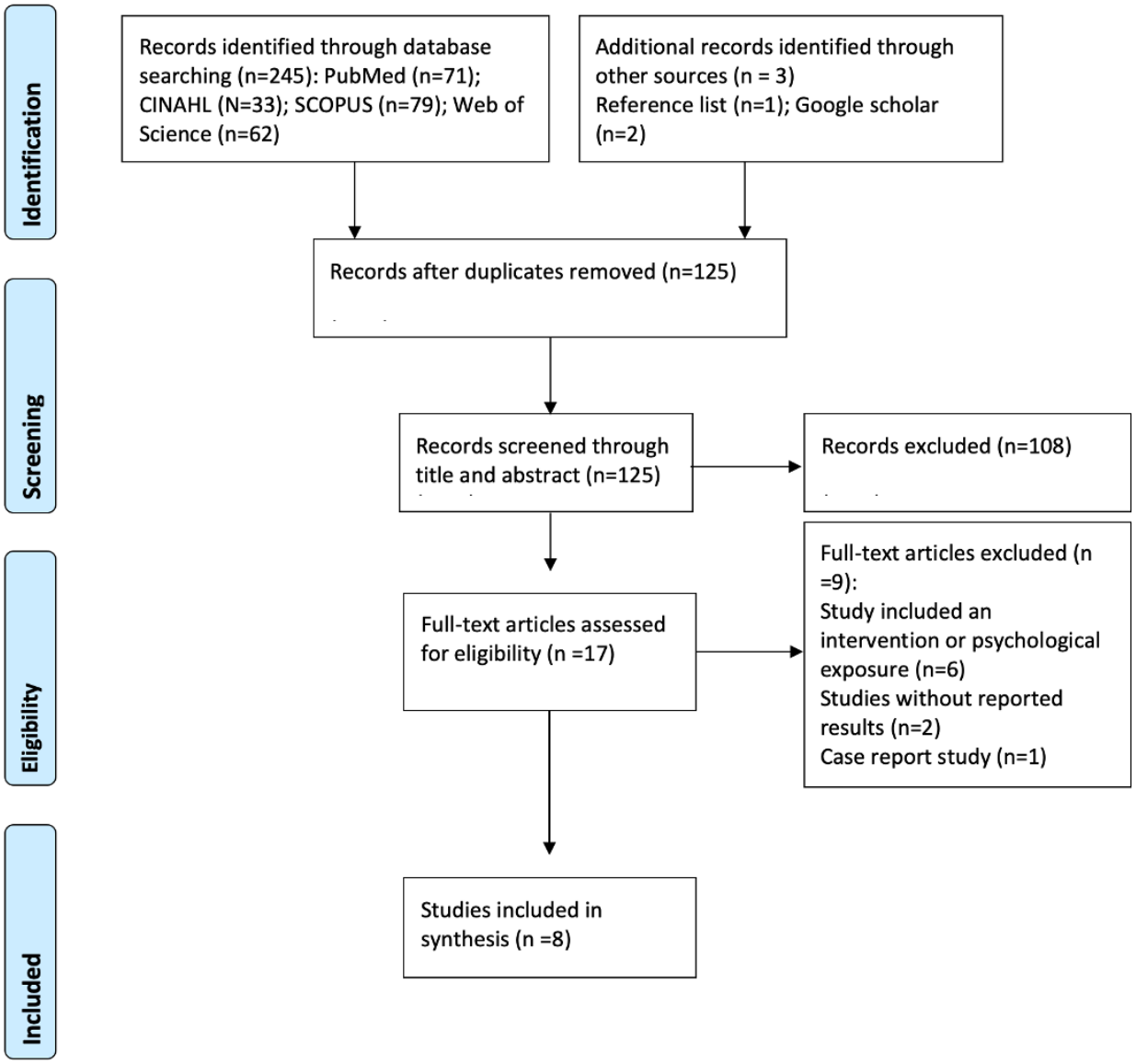
The included studies were not homogenous in terms of the assessment time frame, component, and frequency; thus, a meta-analysis was not possible. A narrative synthesis was chosen to bring together the broad knowledge from a variety of approaches. This type of synthesis is not the same as a narrative description that accompanies many reviews. To synthesize the literature, we applied the guideline from Popay et al [19]. The steps include (1) preliminary analysis, (2) exploration of relationships, and (3) assessment of the robustness of the synthesis. Theory development was not performed due to the exploratory nature of the research synthesized. For the main synthesis, we extracted the descriptive characteristics of the included studies, presented them in a table, and produced a textual summary of the results. These characteristics included first author, publication year, country, study design, population, and sample size. Then, we applied thematic analysis to extract the main themes from all studies. The 3 themes presented in the Results section represent the main areas of knowledge available about HRV in pregnant individuals. These included HRV-related measures during pregnancy (duration and frequency of the HRV assessment, assessment tool used to measure HRV, and assessed HRV components), HRV changes or patterns in the different trimesters of pregnancy as compared with nonpregnant individuals, and HRV changes or patterns across or between the different trimesters of pregnancy.

Results

Study Selection

A total of 245 articles were accessed, of which 120 duplicates were removed by 2 autonomous reviewers. Of the remaining 125 articles, 108 were excluded during the title and abstract screening process. Of the 17 articles that underwent the full-text screening process, 9 were excluded for a variety of reasons (intervention, psychological exposure, findings not reported), resulting in 8 articles that went through the data extraction and synthesis process: See Figure 1 for the PRISMA flow.

Figure 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow chart for study selection.



Results of the Synthesis

Study Characteristics

Participants were mainly pregnant (as the study group) and sometimes included nonpregnant individuals (as a control group) aged from 16 years to 45 years. One of the studies included hypertensive and pre-eclamptic comparison groups [20]. All 8 studies included healthy participants as the main study group. The definition of healthy pregnancy varied from study to study; studies often were selective in considering the American College of Obstetricians and Gynecologists definition or did not specify the characteristics for their definition. For example, one of the studies [21] relied on 2 criteria for distinguishing healthy pregnant individuals: no history of cardiovascular diseases and no drug consumption affecting the ANS. All the studies recruited participants from outpatient settings including prenatal care centers. Of the 8 included studies, 5 were conducted in India, 2 were conducted in the United States, and 1 was conducted in Portugal. See Table 1 for details.

Table 1. Characteristics of included studies based on the population, exposure, comparator, outcome, and study (PECOS) framework.

Author, publication year, country	Study design (S, C)	Population (P, E)	Duration and frequency	Tool	HRV components (O)
Chamchad et al [22], 2007, United States	Cross-sectional with 2 comparison groups	24 nonpregnant healthy and 22 full-term (labor) healthy individuals with a single gestation	A single 10-minute supine position	ECG[a]	LF[b] (ms^{2}), HF[c] (ms^{2}), LF/HF, mean NN interval, SDNN[d], RMSSD[e], pNN50[f], HTI[g], TINN[h]
Puente [20], 2011, Portugal	Longitudinal observational study with 3 groups	217 participants: 135 normal blood pressure, 55 hypertensive, 27 pre-eclamptic	563 recordings of 10-minute measurements at ≤14, 15-19, 20-24, 26-30, 30-35, 36-40 weeks of gestational age in the sitting position	ECG	LF (ms^{2}), HF (ms^{2}), LF/HF, VLF[i]
Garg et al [23], 2020, India	Longitudinal observational with a single group	66 healthy pregnant individuals	A 5-minute measurement at 11-13, 20-22, and 30-32 weeks of gestation in the supine position	ECG	LF (ms^{2}), HF (ms^{2}), LF (nu), HF (nu), LF/HF, SDNN, SDSD[j], pNN50, total power (TP)
Alam et al [24], 2018, India	Cross-sectional with 4 groups	200 healthy participants: nonpregnant (n=50), first trimester (n=50), second trimester (n=50), and third trimester (n=50)	A single 5-minute measurement between 9:00 AM and 12.00 PM in the supine position	ECG	LF (ms^{2}), HF (ms^{2}), HF (nu), LF (nu) LF/HF, mean RR interval, SDNN, RMSSD, NN50, pNN50
Gandhi et al [25], 2014, India	Longitudinal observational with 2 comparison groups	60 healthy participants: pregnant individuals (n=30), nonpregnant individuals (n=30)	A single 5-minute measurement for nonpregnant individuals and twice in the first (6-12 weeks) and third (25-36 weeks) trimesters for pregnant individuals	ECG	VLF, LF (ms^{2}), HF (ms^{2}), LF (nu), HF (nu), LF/HF, SDNN, RMSSD, SDSD, NN50, pNN50, SD_1[k]/SD_2[l], HTI, mean RR interval
Solanki et al [21], 2020, India	Cross-sectional case-control study	119 healthy individuals: pregnant individuals (n=89): T_1 (n=24), T_2 (n=37), T_3 (n=28), and nonpregnant individuals (n=30)	A single 5-minute assessment between 8.30 am and 12.00 pm in the supine position	ECG	VLF, LF (nu), HF (nu), LF/HF, SDNN, RMSSD, SDSD, NN50, pNN50, HTI, SD_1, SD_2
Veerabhadrappa et al [26], 2015, India	Cross-sectional study with 4 groups	156 participants; first trimester (n=25), second trimester (n=47), third trimester (n=52), and postpartum (within a week; n=32)	A single, simultaneous assessment in each group	ECG	LF (nu), HF (nu), LF/HF
Stein et al [27], 1999, United States	Longitudinal observational study with a single group	8 healthy nonpregnant individuals who expect to be pregnant	5 successive 24-hour recordings at prepregnancy and ≤6, 10, 18, and 34 weeks of gestational age	Holter	ULF[m], VLF, LF (ms^{2}), HF (ms^{2}), SDNN, SDANN[n], SDNN, RMSSD, pNN50, TP

[a]ECG: electrocardiogram.

[b]LF: low frequency.

[c]HF: high frequency.

[d]SDNN: standard deviation of the NN interval.

[e]RMSSD: root mean square of successive NN interval differences.

[f]pNN50: percentage of successive NN intervals that differ by more than 50 ms.

[g]HTI: integral of the intensity of the NN interval histogram divided by its height.

[h]TINN: total variation index of the NN intervals.

[i]VLF: very-low frequency.

[j]SDSD: standard deviation of the differences between successive NN intervals.

[k]SD_1: Poincaré plot standard deviation perpendicular to the line of identity.

[l]SD_2: Poincaré plot standard deviation along the line of identity.

[m]ULF: ultra-low frequency.

[n]SDANN: standard deviation of the average NN interval.

HRV Assessment

Tools

To measure HRV components, the majority of the studies (7/8, 88%) computed the short-term beat-to-beat interval using an electrocardiogram (ECG). Only 1 study used a 24-hour Holter for HRV assessment. These assessment tools, however, used different software to analyze the produced algorithm, such as VarioWin_HR (Genesis Medical Systems Pvt Ltd, Telangana, India), NI-DAQ approximate entropy (ApEn; National Instruments Corp, Austin, TX), DATAQ Instruments (Akron, OH), Labchart Pro 7 (ADInstruments, Sydney, Australia), and a Marquette Laser SXP Holter scanner (Marquette Electronics Inc, Milwaukee, WI). One study did not report the software used for signal analysis.

Assessment Characteristics

Assessment processes were performed in health settings such as a hospital or clinic. In 3 of 8 studies, participants were asked to refrain from consuming stimulant substances including tea, coffee, cola, and alcoholic drinks as well as to refrain from smoking for 24 hours before the testing [21,23,24]. In 2 of 8 studies, the time frame for conducting the ECG test was specified to be between 8 am and 9 am to 12 pm [21,24]. In 5 of 8 studies, the ECG test was acquired in the supine position [21-23,25]. In 4 of 8 studies, participants were asked to rest for 5 minutes [21,25], 10 minutes [24], or 20 minutes [23] before undergoing the ECG test. The duration of the ECG record varied in different studies and was either 5 minutes [21,23-25] or 10 minutes [20,22]. One study did not specify the duration [26]. Of the 8 studies, 3 used only lead II to obtain the ECG signals [23-25]. The rest of the studies did not report any information regarding the leads that were used.

Assessed Metrics

In the included studies, in general, HRV was operationalized using 3 types of components: time-domain (TD), frequency-domain (FD), and nonlinear methods. TD is the primary and simplest way to calculate HRV using statistical calculations of several consecutive beat-to-beat (RR) intervals, and a TD graph shows how a signal changes over time. FD represents a model that reflects the strength of the ANS function (specifically the parasympathetic branch) at a given time, and an FD graph shows how much of the signal lies within each given frequency band over a range of frequencies. The representative metrics for FD include the average NN interval, low frequency (LF), high frequency (HF), the LF/HF ratio, very-low frequency (VLF), ultra-low frequency (ULF), and total power (TP). The relevant metrics for TD are the standard deviation of the NN interval (SDNN), root mean square of successive NN interval differences (RMSSD), successive NN intervals that differ by more than 50 ms (NN50), percentage of NN50 (pNN50), standard deviation of the differences between successive NN intervals (SDSD), and integral of the intensity of the NN interval histogram divided by its height (HTI). A few studies also considered nonlinear algorithms such as the Poincaré plot standard deviation perpendicular to the line of identity (SD_1) and Poincaré plot standard deviation along the line of identity (SD_2) to measure HRV during pregnancy [11]. The detailed descriptions of all these HRV components including TD, FD, and nonlinear metrics are provided in the Table 2.

Of the 2 main HRV components (TD and FD), at least 3 indexes of FD metrics (eg, LF, HF, LF/HF, VLF, ULF, average NN [RR] interval, TP) were assessed in all the included studies. TD indexes, including SDNN, RMSSD, NN50, pNN50, SDSD, HTI, and TP, were also assessed. The nonlinear methods of SD_1 and SD_2 were acquired in 2 of 8 studies.

HRV changes or an HRV pattern was defined in the studies as an increase or decrease in the aforementioned HRV components. For example, SDRR is defined as standard deviation of the RR interval. If an article reported an increased SDRR from the first to the second trimesters, the SD of the RR interval was increased from the first trimester to the second trimester.

Table 2. Heart rate variability (HRV) components and metrics.

Components and metrics	Unit	Description
Time domain		
SDNN	ms	Standard deviation of the NN interval
RMSSD	ms	Root mean square of successive NN interval differences
NN50	ms	Mean number of times an hour in which the change in successive normal sinus (NN) intervals exceeds 50 ms
pNN50	%	Percentage of successive NN intervals that differ by more than 50 ms
SDSD	ms	Standard deviation of the differences between successive NN intervals
HTI	N/A[a]	Integral of the intensity of the NN interval histogram divided by its height.
Frequency domain		
LF	ms^2/nu	Absolute/relative power of the low frequency band (0.04-0.15 Hz)
HF	ms^2/nu	Absolute/relative power of the high frequency band (0.15-0.4 Hz)
LF/HF	%	Ratio of LF to HF
ULF	ms^2	Absolute power of the ultra-low frequency band (≤0.003 Hz)
VLF	ms^2	Absolute power of very-low frequency band (0.0033-0.04 Hz)
Avg N-N	ms	Mean of the NN intervals
TP	ms^2	Absolute power of the total frequency band (≤0.4 Hz)
Nonlinear		
SD_1	ms^2	Poincaré plot standard deviation perpendicular to the line of identity
SD_2	ms^2	Poincaré plot standard deviation along the line of identity

[a]N/A: not applicable.

HRV Changes in Different Trimesters as Compared With Nonpregnant Individuals

The nonpregnant comparison group included individuals who did not report pregnancy at the time of assessment or were postpartum. Of the studies, 75% (6/8) reported HRV changes in pregnancy as compared with nonpregnancy. However, not all 6 studies overlapped in terms of the assessed HRV components. With this heterogeneity, the data for a trend assessment across the studies were insufficient for most of the HRV components. To report the frequency of the assessed metrics, we used the report format of "X out of Y increased/decreased" in which Y represents the total number of studies that reported the intended component and X reflects the number of studies in which X decreased or increased. The direction of change for most of the HRV components in both FD and TD indexes varied in different studies, specifically in early pregnancy. However, some FD and TD elements showed the same change direction (increase or decrease) in late pregnancy in the majority of the studies that reported the HRV. For instance, as compared with nonpregnant individual elements, the FD elements of HF (nu) in 75% (3/4) of the studies, HF (ms^2) in 100% (3/3) of the studies, and VLF in 100% (3/3) of the studies decreased, and LF (nu) increased in late pregnancy in 75% (3/4) of the studies. The TD components of SDNN in 80% (4/5) of the studies, RMSSD in 100% (5/5) of the studies, and pNN50 in 80% (4/5) of the studies decreased in late pregnancy as compared with nonpregnant individuals. See Table 3 for more details.

Table 3. Changes in heart rate variability (HRV) components in pregnant individuals as compared with nonpregnant individuals in different trimesters.

First author, year, and measurement period	Linear: frequency domain (FD)								Linear: time domain (TD)							Nonlinear	
	LF[a]	HF[b]	LF (nu)	HF (nu)	LF/HF	ULF[c]	VLF[d]	Average NN	TP[e]	SDNN[f]	RMSSD[g]	NN50[h]	pNN50[i]	SDSD[j]	HTI[k]	SD_1[l]	SD_2[m]
Gandhi [25], 2014																	
T_1[n]	NC[o]	NC	NC	NC	NC	—[p]	NC	—	—	NC	NC	NC	NC	NC	—	NC	NC
T_3[q]	D[r]	D	I[s]	D	D	—	D	—	—	D	D	D	D	D	—	I	I
Stein [27], 1999																	
T_1	D	D	—	—	—	D	D	—	D	D	D	—	D	—	—	—	—
T_3	D	D	—	—	—	D	D	—	D	D	D	—	D	—	—	—	—
Chamchad [22], 2007																	
T_3	NC	NC	—	—	I	—	—	D	—	D	D	—	D	—	D	D	D
Solanki [21] , 2020																	
T_1	—	—	D	I	D	—	D	—	—	D	D	I	I	I	I	D	D
T_2[t]	—	—	D	I	D	—	D	—	—	D	D	I	I	I	I	D	D
T_3	—	—	D	I	D	—	D	—	—	D	D	I	I	I	I	D	D
Alam [24] , 2018																	
T_1	D	I	D	I	D	—	—	I	—	I	I	I	I	—	—	—	—
T_2	I	D	I	D	I	—	—	D	—	I	I	D	D	—	—	—	—
T_3	I	D	I	D	I	—	—	D	—	I	D	D	D	—	—	—	—
Veerabhadrappa [26] , 2015																	
T_1	—	—	I	D	I	—	—	—	—	—	—	—	—	—	—	—	—
T_2	—	—	I	D	I	—	—	—	—	—	—	—	—	—	—	—	—
T_3	—	—	I	D	I	—	—	—	—	—	—	—	—	—	—	—	—

[a]LF: low frequency.

[b]HF: high frequency.

[c]ULF: ultra-low frequency.

[d]VLF: very-low frequency.

[e]TP: total power.

[f]SDNN: standard deviation of the NN interval.

[g]RMSSD: root mean square of successive NN interval differences.

[h]NN50: successive NN intervals that differ by more than 50 ms.

[i]pNN50: percentage of successive NN intervals that differ by more than 50 ms.

[j]SDSD: standard deviation of the differences between successive NN intervals.

[k]HTI: integral of the intensity of the NN interval histogram divided by its height.

[l]SD_1: Poincaré plot standard deviation perpendicular to the line of identity.

[m]SD_2: Poincaré plot standard deviation along the line of identity.

[n]T1: first trimester.

[o]NC: no change.

[p]Not applicable.

[q]T_3: third trimester.

[r]D: decrease.

[s]I: increase.

[t]T_2: second trimester.

HRV Adaptation During Different Trimesters of Pregnancy

According to the literature, ANS regulation during pregnancy starts in the first weeks of pregnancy and continues until the end of pregnancy. These changes vary based on the pregnancy-related time-sensitive demands to ensure fetus development. To understand the potential differences in HRV components between different trimesters, for this literature review, HRV changes were assessed in transition between different trimesters during pregnancy**.**

First Trimester to the Third Trimester

HRV changes during the first and third trimesters of pregnancy were assessed in 7 (7/8, 88%) studies. In all 7 studies, whether they were longitudinal studies with the same population or cross-sectional studies with different populations in different trimesters, the findings for most of the FD and TD indexes were analogous. Most of the TD metrics generally decreased from the first trimester to the third trimester in the majority (5/7, 71%) of the studies. Of the TD indexes, the following decreased during pregnancy: SDNN in 80% (4/5) of the studies, RMSSD in 75% (3/4) of the studies, NN50 in 70% (2/3) of the studies, PNN50 in 80% (4/5) of the studies, SDSD in 70% (2/3) of the studies, and SD_1 and SD_2 in 50% (1/2) of the studies. The FD indexes also decreased most of the time except for normalized LF (nu) and the LF/HF ratio, which showed an ascending trend from early to late pregnancy. Among the various FD indexes, LF power (ms^2) in 60% (3/5) of the studies, HF power (ms^2) in 100% (5/5) of the studies, HF (nu) in 100% (5/5) of the studies, VLF in 75% (3/4) of the studies, TP in 100% (2/2) of the studies, and the average NN interval in 100% (2/2) of the studies decreased, while LF (nu) in 80% (4/5) of the studies and LF/HF in 100% (6/6) studies increased. See Table 4 for more details.

Table 4. Changes in heart rate variability (HRV) components during pregnancy from early (first trimester) to late pregnancy (third trimester).

First author, year	Linear frequency domain (FD)									Linear time domain (TD)						Nonlinear
	LF[a]	HF[b]	LF (nu)	HF (nu)	LF/HF	ULF[c]	VLF[d]	Average NN	TP[e]	SDNN[f]	RMSSD[g]	NN50[h]	pNN50[i]	SDSD[j]	HTI[k]	SD_1[l] and SD_2[m]
Puente [20], 2011	NC[n]	D[o]	—[p]	—	I[q]	—	I	—	—	—	—	—	—	—	—	—
Garg [23], 2020	D	D	I	D	I	—	—	—	D	D	—	—	D	D	—	—
Gandhi [25], 2014	D	D	I	D	I	—	D	D	—	D	D	—	D	D	D	D
Stein [27], 1999	D	D	—	—	—	D	D	—	D	D	D	D	D	—	—	—
Solanki [21], 2020	—	—	D	D	I	—	D	—	—	NC	NC	NC	NC	NC	NC	NC
Alam [24], 2018	I	D	I	D	I	—	—	D	—	D	D	D	D	—	—	—
Veerabhadrappa [26], 2015	—	—	I	D	I	—	—	—	—	—	—	—	—	—	—	—

[a]LF: low frequency.

[b]HF: high frequency.

[c]ULF: ultra-low frequency.

[d]VLF: very-low frequency.

[e]TP: total power.

[f]SDNN: standard deviation of the NN interval.

[g]RMSSD: root mean square of successive NN interval differences.

[h]NN50: successive NN intervals that differ by more than 50 ms.

[i]pNN50: percentage of successive NN intervals that differ by more than 50 ms.

[j]SDSD: standard deviation of the differences between successive NN intervals.

[k]HTI: integral of the intensity of the NN interval histogram divided by its height.

[l]SD_1: Poincaré plot standard deviation perpendicular to the line of identity.

[m]SD_2: Poincaré plot standard deviation along the line of identity.

[n]NC: no change.

[o]D: decrease.

[p]Not applicable.

[q]I: increase.

First Trimester to the Second Trimester

Only one-half (4/8, 50%) of the studies reported HRV changes from the first trimester to the second trimester. The included articles either did not report TD and FD or were divergent in measured TD and FD indexes from the first to the second trimesters of pregnancy. Thus, determining a general pattern relying on the current findings cannot be done. However, the elements often tended to decline or stay unchanged from the first to the second trimesters except for normalized LF (nu) and LF/HF, which increased in 75% (3/4) and 100% (4/4), respectively, of the reported elements. See Table 5 for more details.

Table 5. Changes in heart rate variability (HRV) components from the first to the second trimesters of pregnancy.

First author, year	Linear frequency domain (FD)									Linear time domain (TD)						Non linear
	LF[a]	HF[b]	LF (nu)	HF (nu)	LF/HF	ULF[c]	VLF[d]	Average NN	TP[e]	SDNN[f]	RMSSD[g]	NN50[h]	pNN50[i]	SDSD[j]	HTI[k]	SD_1[l] and SD_2[m]
Garg et al [23], 2020	D[n]	D	I[o]	D	I	—[p]	—	—	D	D	—	—	D	D	—	—
Solanki [21], 2020	—	—	D	D	I	—	D	—	—	NC[q]	NC	NC	NC	NC	NC	NC
Alam et al [24], 2018	I	D	I	D	I	—	—	D	—	D	D	D	D	—	—	—
Veerabhadrappa [26], 2015	—	—	I	D	I	—	—	—	—	—	—	—	—	—	—	—

[a]LF: low frequency.

[b]HF: high frequency.

[c]ULF: ultra-low frequency.

[d]VLF: very-low frequency.

[e]TP: total power.

[f]SDNN: standard deviation of the NN interval.

[g]RMSSD: root mean square of successive NN interval differences.

[h]NN50: successive NN intervals that differ by more than 50 ms.

[i]pNN50: percentage of successive NN intervals that differ by more than 50 ms.

[j]SDSD: standard deviation of the differences between successive NN intervals.

[k]HTI: integral of the intensity of the NN interval histogram divided by its height.

[l]SD_1: Poincaré plot standard deviation perpendicular to the line of identity.

[m]SD_2: Poincaré plot standard deviation along the line of identity.

[n]D: decrease.

[o]I: increase.

[p]Not applicable.

[q]NC: no change.

Second Trimester to the Third Trimester

Only one-half (4/8, 50%) of the studies reported HRV changes from the second to the third trimesters. Divergency in the TD and FD metrics considered in studies did not allow for the comparative assessment between studies. Nevertheless, the changes mostly were similar to those from the first to the second trimesters except for normalized LF (nu), which was neutral; HF (nu), which decreased in 75% (3/4) of the studies; and the LF/HF ratio, which increased in 75% (3/4) of the studies that measured the HRV from the second trimester to the third trimester. See Table 6 for more details.

Table 6. Changes in heart rate variability (HRV) components from the second to the third trimesters of pregnancy.

First author, year	Linear frequency domain (FD)									Linear time domain (TD)						Nonlinear
	LF[a]	HF[b]	LF (nu)	HF (nu)	LF/HF	ULF[c]	VLF[d]	Average NN	TP[e]	SDNN[f]	RMSSD[g]	NN50[h]	pNN50[i]	SDSD[j]	HTI[k]	SD_1[l] and SD_2[m]
Garg et al [23], 2020	D[n]	D	D	I[o]	D	—[p]	—	—	D	D	—	—	D	D	—	—
Solanki [21], 2020	—	—	D	D	I	—	D	—	—	NC[q]	NC	NC	NC	NC	NC	NC
Alam et al [24], 2018	I	D	I	D	I	—	—	D	—	I	D	D	D	—	—	—
Veerabhadrappa [26], 2015	—	—	I	D	I	—	—	—	—	—	—	—	—	—	—	—

[a]LF: low frequency.

[b]HF: high frequency.

[c]ULF: ultra-low frequency.

[d]VLF: very-low frequency.

[e]TP: total power.

[f]SDNN: standard deviation of the NN interval.

[g]RMSSD: root mean square of successive NN interval differences.

[h]NN50: successive NN intervals that differ by more than 50 ms.

[i]pNN50: percentage of successive NN intervals that differ by more than 50 ms.

[j]SDSD: standard deviation of the differences between successive NN intervals.

[k]HTI: integral of the intensity of the NN interval histogram divided by its height.

[l]SD_1: Poincaré plot standard deviation perpendicular to the line of identity.

[m]SD_2: Poincaré plot standard deviation along the line of identity.

[n]D: decrease.

[o]I: increase.

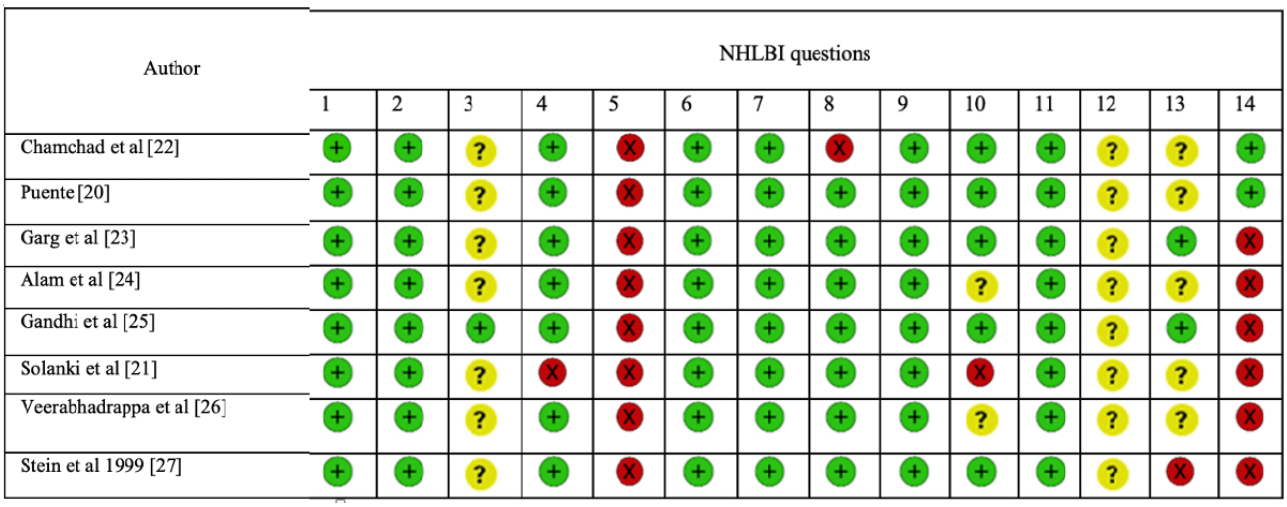
[p]Not applicable.

[q]NC: no change.

Risk of Bias in the Studies

The result of the NHLBI assessment is reported in Figure 2. Articles were peer-reviewed in Covidence online software by 2 independent reviewers. All the studies were screened by both reviewers, ZS and JO, in 2 steps. In the first step, the quality of articles was assessed using the 14 domains of the NHLBI see Multimedia Appendix 2 for more details). Next, the overall quality of each study was assessed based on the addressed NHLBI domains. To resolve raised disagreements, a third reviewer, MB, was involved. The research question, study population, dependent variables, and independent variables were specified in all the studies. Since the independent variable was "being pregnant," it was already established in the study populations. The dependent variable was HRV, which may be affected by pregnancy. Both variables were consistent across the study populations in all the studies. The sample size was justified in none of the studies. In all the studies (8/8, 100%), the timeframe was sufficient for the potential expected association to occur; the study population was pregnant individuals who were already pregnant when the HRV assessment was performed. The overall quality of the studies was based on the number of domains addressed in each study.

Figure 2. National Heart, Lung, and Blood Institute (NHLBI) quality assessment tool, with which the study quality was scored based on the number of addressed domains (≤7 poor; 8 fair; ≥9 good): (1) clear research objective; (2) clear study population; (3) participation rate >50%; (4) internal validity in the population; (5) sample size justification; (6) prospective study; (7) time span between exposure and outcome; (8) exposure aspects; (9) exposure measures; (10) exposure assessment frequency; (11) outcome aspects; (12) blinded outcome assessment; (13) attrition; (14) confounding factors.



Discussion

Principal Findings

In this study, we aimed to review studies concerned with HRV adaptation to evaluate ANS function in healthy pregnant individuals. According to our findings based on the existing data, during pregnancy, almost all the TD and most of the FD bands were decreased except for LF (nu) and the LF/HF ratio. From the second trimester to the third trimester, however, LF change was not consistent across the studies; for example, one-half of the studies showed a decrease, while the other one-half demonstrated an increase in LF. Increased LF/HF and LF during pregnancy indicate a dominance of the sympathetic nervous system over the parasympathetic nervous system. This result is in accordance with the findings in the existing literature, in which methods other than the HRV were used for ANS assessment. For example, Ekholm et al [28] used various maneuvers including the Valsalva maneuver, deep breathing test, orthostatic test, and isometric handgrip test to assess ANS changes in pregnancy. They concluded that the parasympathetic nervous system becomes less active as time progresses in pregnancy [28]. Kochhar et al [29] also applied conventional tests such as the standing-to-lying down ratio, Valsalva maneuver, tachycardia maneuver, hand grip test, and cold pressor test to assess the ANS in pregnancy. The results supported sympathetic activation over parasympathetic activation from the first through the third trimesters and compared with nonpregnant women [29]. Systematic reviews have also investigated the application of HRV for ANS assessment in complicated pregnancies and showed that HRV as a biomarker of the ANS can be affected by common pregnancy complications including hypertensive disorders such as preeclampsia [28,29]. These studies suggested that parasympathetic activity of the ANS decreases more than sympathetic activity in hypertensive pregnant women. This may explain why hypertension often occurs in the late second or third trimester of pregnancy when the physiological response tends to be sympathetic overactivation.

Synthesis of the Results and Limitations

The heterogeneity in the findings from the included studies can potentially be explained by the methodological issues we uncovered, which threaten the internal and external validity of the findings. Due to the divergence in the selected and reported HRV components by different studies, a reliable conclusion regarding ANS function cannot be reached using these insufficient data. Additionally, the wide variability in the length of the recording period may have significantly affected both FD and TD measurements [30]. A short-term epoch (~5 minutes) lacks the prognostic potential for morbidity and mortality. Basically, in published protocols, the recommended assessment periods for HRV recordings vary from 1 minute to 24 hours for various FD and TD metrics [11]. However, since important factors including circadian rhythms, metabolism, the sleep cycle, core body temperature, and the renin-angiotensin system follow a 24-hour cycle, the length of clinical HRV assessments should be at least 24 hours to provide acceptable information [11]. The studies that were reviewed in this study often used 5-minute to 10-minute assessments, which may make it challenging to achieve reliable results. Furthermore, the HRV measurement in the included studies was often conducted in clinic or hospital settings, and the studies often ignored the impact of mental-environmental confounding factors including negative mental situations such as stress, anxiety, and fear and environmental factors (eg, temperature) that can temporarily affect ANS function, potentially leading to false results. These factors could be significantly associated with ANS function and thus may lead to variability in and misinterpretation of ANS function in response to pregnancy. Also, the frequency of HRV measurements varied from 1 to 3 times a trimester in different studies. An episodic assessment (ie, 1 or 3 times a trimester) may not be reflective of actual ANS function in real life in response to the ever-changing pregnancy-related demands. This is because the ANS is a responsive system that continuously undergoes dynamic adaptations in response to the various internal and external situations that one may face from moment to moment [8].

Another inconsistency we found in the included studies was the way they operationalized HRV. As discussed earlier, HRV was operationalized in various ways, via TD, FD, and nonlinear methods, each involving different corresponding components. No justification was provided in any study for the selection of HRV measurement modality and associated components. Specifying the weaknesses and strengths of the applied components for measurement may provide important information for future studies. It is critical to address why some components are commonly used as compared with others to represent HRV and whether the applied component is reliable, valid, and easily measured. For example, recent studies showed that linear algorithms including TD and FD are affected by nonstationarity and thus perhaps not adequate for HRV assessment [31]. Nonlinear (fractal) measurements such as SD_1, SD_2, ApEn, and sample entropy are recommended as they represent the unpredictability of a time series resulting from the complexity of the regulatory mechanisms of HRV. It is suggested that nonlinear HRV measures may enable clinicians and researchers to study the complex interactions between electrophysiological, hemodynamic, and humoral variables as well as their regulation by the ANS and central nervous system [31].

One of the limitations of this study is that its protocol was not registered in PROSPERO.

Implications

This study will help us determine if there are consistent stable patterns of HRV across pregnancy for a sample of healthy pregnant women that reflect "healthy" ANS function. Determination of the pattern can provide the basis for extended research in order to determine if the identified pattern is generalizable to a larger sample of pregnant individuals and then in other, diverse pregnant populations. If a pattern is found, then the next step would be to compare the pattern with complicated pregnancies to find where, when, and how this pattern may be different. Understanding the differences between complicated and healthy pregnancies can provide an opportunity to develop a screening and detective system to screen all pregnant women throughout pregnancy and predict whether they are manifesting the nonhealthy pattern, to be able to perform a timely intervention before it threatens the mother's and baby's lives. Due to the limitations in the existing literature, identifying potential patterns seems critical. Thus, more studies are needed to reflect this pattern by eliminating the methodological limitations in current studies.

Conclusion

This study determined the feasibility of HRV measurements to assess ANS function in pregnant individuals. We found significant heterogeneity in the HRV measurement modalities used, the settings in which measurements were performed, the time frames of the HRV assessment, and assessments done across trimesters. There were inconsistencies in definitions of healthy pregnancy across studies. We found some potential HRV patterns that were consistent within and across studies, but not all the studies were convergent in terms of the reported results, which may be due to the methodological heterogeneity. In summary, we found significant variability in how studies measured HRV and how they identified HRV patterns, which made it impossible to determine potentially normal HRV patterns across trimesters with any degree of validity. More research is needed to overcome the aforementioned limitations and determine the required criteria and methods to assess HRV patterns corresponding to ANS function in healthy pregnant persons.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File , 14 KB - bioinform_v3i1e36791_app1.docx]

Multimedia Appendix 2
Quality assessment domains of the National Heart, Lung, and Blood Institute (NHLBI).
[PNG File , 238 KB - bioinform_v3i1e36791_app2.png]

References

1. Waxenbaum JA, Reddy V, Varacello M. Anatomy, Autonomic Nervous System. Treasure Island, FL: StatPearls Publishing; Jan 01, 2022.
2. Fu Q, Levine B. Autonomic circulatory control during pregnancy in humans. Semin Reprod Med 2009 Jul 15;27(4):330-337 [FREE Full text] [doi: 10.1055/s-0029-1225261] [Medline: 19530067]
3. Cerritelli F, Frasch MG, Antonelli MC, Viglione C, Vecchi S, Chiera M, et al. A review on the vagus nerve and autonomic nervous system during fetal development: searching for critical windows. Front Neurosci 2021 Sep 20;15:721605 [FREE Full text] [doi: 10.3389/fnins.2021.721605] [Medline: 34616274]
4. Conrad KP. Maternal vasodilation in pregnancy: the emerging role of relaxin. Am J Physiol Regul Integr Comp Physiol 2011 Aug;301(2):R267-R275 [FREE Full text] [doi: 10.1152/ajpregu.00156.2011] [Medline: 21613576]

5. Leo CH, Jelinic M, Ng HH, Marshall SA, Novak J, Tare M, et al. Vascular actions of relaxin: nitric oxide and beyond. Br J Pharmacol 2017 May 30;174(10):1002-1014 [FREE Full text] [doi: 10.1111/bph.13614] [Medline: 27590257]
6. Tkachenko O, Shchekochikhin D, Schrier RW. Hormones and hemodynamics in pregnancy. Int J Endocrinol Metab 2014 Apr 01;12(2):e14098 [FREE Full text] [doi: 10.5812/ijem.14098] [Medline: 24803942]
7. Ishida J, Matsuoka T, Saito-Fujita T, Inaba S, Kunita S, Sugiyama F, et al. Pregnancy-associated homeostasis and dysregulation: lessons from genetically modified animal models. J Biochem 2011 Jul 25;150(1):5-14. [doi: 10.1093/jb/mvr069] [Medline: 21613291]
8. Balajewicz-Nowak M, Furgala A, Pitynski K, Thor P, Huras H, Rytlewski K. The dynamics of autonomic nervous system activity and hemodynamic changes in pregnant women. Neuro Endocrinol Lett 2016;37(1):70-77. [Medline: 26994389]
9. Zygmunt A, Stanczyk J. Methods of evaluation of autonomic nervous system function. Arch Med Sci 2010 Mar 01;6(1):11-18 [FREE Full text] [doi: 10.5114/aoms.2010.13500] [Medline: 22371714]
10. Ziemssen T, Siepmann T. The investigation of the cardiovascular and sudomotor autonomic nervous system-a review. Front Neurol 2019 Feb 12;10:53 [FREE Full text] [doi: 10.3389/fneur.2019.00053] [Medline: 30809183]
11. Shaffer F, Ginsberg JP. An overview of heart rate variability metrics and norms. Front Public Health 2017;5:258 [FREE Full text] [doi: 10.3389/fpubh.2017.00258] [Medline: 29034226]
12. Natarajan A, Pantelopoulos A, Emir-Farinas H, Natarajan P. Heart rate variability with photoplethysmography in 8 million individuals: a cross-sectional study. The Lancet Digital Health 2020 Dec;2(12):e650-e657. [doi: 10.1016/s2589-7500(20)30246-6]
13. Ying W, Catov JM, Ouyang P. Hypertensive disorders of pregnancy and future maternal cardiovascular risk. JAHA 2018 Sep 04;7(17):1. [doi: 10.1161/jaha.118.009382]
14. Massaro S, Pecchia L. Heart rate variability (HRV) analysis: a methodology for organizational neuroscience. Organizational Research Methods 2016 Dec 01;22(1):354-393. [doi: 10.1177/1094428116681072]
15. Ernst G. Heart-rate variability-more than heart beats? Front Public Health 2017 Sep 11;5:240 [FREE Full text] [doi: 10.3389/fpubh.2017.00240] [Medline: 28955705]
16. Kataoka K, Tomiya Y, Sakamoto A, Kamada Y, Hiramatsu Y, Nakatsuka M. Altered autonomic nervous system activity in women with unexplained recurrent pregnancy loss. J Obstet Gynaecol Res 2015 Jun 29;41(6):912-918. [doi: 10.1111/jog.12653] [Medline: 25546149]
17. Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, PRISMA-P Group. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev 2015 Jan 01;4(1):1 [FREE Full text] [doi: 10.1186/2046-4053-4-1] [Medline: 25554246]
18. Quality Assessment of Controlled Intervention Studies. National Heart, Lung, and Blood Institute. Accessed 13 Dec 2021. URL: https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools [accessed 2022-11-08]
19. Popay J, Roberts H, Sowden A, Petticrew M, Arai L, Rodgers M, et al. Guidance on the Conduct of Narrative Synthesis in Systematic Reviews: A Product from the ESRC Methods Programme. Lancaster University. 2006 Apr. URL: https://www.lancaster.ac.uk/media/lancaster-university/content-assets/documents/fhm/dhr/chir/NSsynthesisguidanceVersion1-April2006.pdf [accessed 2022-11-13]
20. Puente ET. Heart rate variability analysis during normal and hypertensive pregnancy. University of Porto. 2010. URL: https://repositorio-aberto.up.pt/bitstream/10216/53927/2/teseeduardo%20puente.pdf [accessed 2022-11-06]
21. Solanki J, Desai F, Desai K. Heart rate variability is reduced in normal pregnancy irrespective of trimester: A cross-sectional study from Gujarat, India. J Family Med Prim Care 2020 Feb;9(2):626-631 [FREE Full text] [doi: 10.4103/jfmpc.jfmpc_1123_19] [Medline: 32318393]
22. Chamchad D, Horrow J, Nakhamchik L, Arkoosh V. Heart rate variability changes during pregnancy: an observational study. Int J Obstet Anesth 2007 Apr;16(2):106-109. [doi: 10.1016/j.ijoa.2006.08.008] [Medline: 17270423]
23. Garg P, Yadav K, Jaryal AK, Kachhawa G, Kriplani A, Deepak KK. Sequential analysis of heart rate variability, blood pressure variability and baroreflex sensitivity in healthy pregnancy. Clin Auton Res 2020 Oct 24;30(5):433-439. [doi: 10.1007/s10286-020-00667-4] [Medline: 31981003]
24. Alam T, Choudhary AK, Kumaran DS. Maternal heart rate variability during different trimesters of pregnancy. Natl J Physiol Pharm Pharmacol 2018;8(9):1475. [doi: 10.5455/njppp.2018.8.0723327072018]
25. Gandhi P, Mehta H, Gokhale A, Desai C, Gokhale P, Shah C. A study on cardiac autonomic modulation during pregnancy by non-invasive heart rate variability measurement. Int J Med Public Health 2014;4(4):441. [doi: 10.4103/2230-8598.144131]
26. Veerabhadrappa S, Vyas AL, Anand S. Changes in heart rate variability and pulse wave characteristics during normal pregnancy and postpartum. IJBET 2015;17(2):99. [doi: 10.1504/ijbet.2015.068045]
27. Stein PK, Hagley MT, Cole PL, Domitrovich PP, Kleiger RE, Rottman JN. Changes in 24-hour heart rate variability during normal pregnancy. Am J Obstet Gynecol 1999 Apr;180(4):978-985. [doi: 10.1016/s0002-9378(99)70670-8] [Medline: 10203667]
28. Ekholm EMK, Piha SJ, Erkkola RU, Antila KJ. Autonomic cardiovascular reflexes in pregnancy. A longitudinal study. Clinical Autonomic Research 1994 Aug;4(4):161-165. [doi: 10.1007/bf01826181]
29. Kochhar DS, Jit DS, Ummat DA. Study of autonomic sympatho-vagal modulation in different trimesters of pregnancy. Int J Med Res Rev 2016 May 31;4(5):751-757. [doi: 10.17511/ijmrr.2016.i05.15]

30. Laborde S, Mosley E, Thayer JF. Heart rate variability and cardiac vagal tone in psychophysiological research - recommendations for experiment planning, data analysis, and data reporting. Front Psychol 2017 Feb 20;8:213 [FREE Full text] [doi: 10.3389/fpsyg.2017.00213] [Medline: 28265249]
31. Francesco B, Maria Grazia B, Emanuele G, Valentina F, Sara C, Chiara F, et al. Linear and nonlinear heart rate variability indexes in clinical practice. Comput Math Methods Med 2012;2012:219080 [FREE Full text] [doi: 10.1155/2012/219080] [Medline: 22400047]

Abbreviations

ANS: autonomic nervous system
ApEn: approximate entropy
ECG: electrocardiogram
HF: high frequency
HRV: heart rate variability
HTI: integral of the intensity of the NN interval histogram divided by its height
LF: low frequency
NHLBI: National Heart, Lung, and Blood Institute
NN50: successive NN intervals that differ by more than 50 ms
PECO: population, exposure, comparator, and outcome
pNN50: percentage of NN50 (pNN50)
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RMSSD: root mean square of successive NN interval differences
SD1: Poincaré plot standard deviation perpendicular to the line of identity
SD2: Poincaré plot standard deviation along the line of identity
SDSD: standard deviation of the differences between successive NN intervals
SDNN: standard deviation of the NN interval
TP: total power
ULF: ultra-low frequency
VLF: very-low frequency

Edited by A Mavragani; submitted 25.01.22; peer-reviewed by S Norouzi, Z Galavi; comments to author 29.06.22; revised version received 20.08.22; accepted 03.11.22; published 17.11.22.

Please cite as:
Sharifiheris Z, Rahmani A, Onwuka J, Bender M
The Utilization of Heart Rate Variability for Autonomic Nervous System Assessment in Healthy Pregnant Women: Systematic Review
JMIR Bioinform Biotech 2022;3(1):e36791
URL: https://bioinform.jmir.org/2022/1/e36791
doi:10.2196/36791
PMID:

Review

Use of Artificial Intelligence in the Search for New Information Through Routine Laboratory Tests: Systematic Review

Glauco Cardozo[1*], PhD; Salvador Francisco Tirloni[2*], MSci; Antônio Renato Pereira Moro[2*], PhD; Jefferson Luiz Brum Marques[2*], PhD

[1]Federal Institute of Santa Catarina, Florianópolis, Brazil

[2]Federal University of Santa Catarina, Florianopolis, Brazil

[*]all authors contributed equally

Corresponding Author:
Glauco Cardozo, PhD
Federal Institute of Santa Catarina
Av. Mauro Ramos, 950 - Centro
Florianópolis, 88020-300
Brazil
Phone: 55 48984060740
Email: glauco.cardozo@ifsc.edu.br

Abstract

Background: In recent decades, the use of artificial intelligence has been widely explored in health care. Similarly, the amount of data generated in the most varied medical processes has practically doubled every year, requiring new methods of analysis and treatment of these data. Mainly aimed at aiding in the diagnosis and prevention of diseases, this precision medicine has shown great potential in different medical disciplines. Laboratory tests, for example, almost always present their results separately as individual values. However, physicians need to analyze a set of results to propose a supposed diagnosis, which leads us to think that sets of laboratory tests may contain more information than those presented separately for each result. In this way, the processes of medical laboratories can be strongly affected by these techniques.

Objective: In this sense, we sought to identify scientific research that used laboratory tests and machine learning techniques to predict hidden information and diagnose diseases.

Methods: The methodology adopted used the population, intervention, comparison, and outcomes principle, searching the main engineering and health sciences databases. The search terms were defined based on the list of terms used in the Medical Subject Heading database. Data from this study were presented descriptively and followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses; 2020) statement flow diagram and the National Institutes of Health tool for quality assessment of articles. During the analysis, the inclusion and exclusion criteria were independently applied by 2 authors, with a third author being consulted in cases of disagreement.

Results: Following the defined requirements, 40 studies presenting good quality in the analysis process were selected and evaluated. We found that, in recent years, there has been a significant increase in the number of works that have used this methodology, mainly because of COVID-19. In general, the studies used machine learning classification models to predict new information, and the most used parameters were data from routine laboratory tests such as the complete blood count.

Conclusions: Finally, we conclude that laboratory tests, together with machine learning techniques, can predict new tests, thus helping the search for new diagnoses. This process has proved to be advantageous and innovative for medical laboratories. It is making it possible to discover hidden information and propose additional tests, reducing the number of false negatives and helping in the early discovery of unknown diseases.

(JMIR Bioinform Biotech 2022;3(1):e40473) doi:10.2196/40473

KEYWORDS

review; laboratory tests; machine learning; prediction; diagnosis; COVID-19

Introduction

Background

The large amount of data generated in the last decades has become a great challenge, demanding new forms of analysis and processing of complex and unstructured data, known until now as data mining [1]. The health care domain has great prominence in applying data mining, supporting infection control, epidemiological analysis, treatment and diagnosis of diseases, hospital management, home care, public health administration, and disease management [2]. This predictive analysis is strongly linked to the evolution of artificial intelligence (AI) techniques such as machine learning (ML). These algorithms, able to learn interactively from data, allow systems based on computational intelligence to find information that was initially unknown [3].

Currently, prediction systems [4] and decision-making support have been using web-based medical records and clinical data, analyzing the history of patients to propose models to identify high-risk situations as well as false positives [5]. This precision medicine (in silico) based on electronic health records has gained strength given the possibility of more accessible and efficient treatments aimed at the particular characteristics of each individual. In this sense, Wong et al [6] proposed using ML to structure and organize stored data and for mining and aiding in diagnosis. Similarly, Roy et al [7] used electronic health record data to predict laboratory test results in a pretest.

These works motivated us to study the potential of the use of AI, especially ML techniques, in the area of health.

According to Peek et al [8], in recent decades, there has been a major shift from knowledge-based to data-oriented methods. Analyzing 30 years of publications from the International Conference on Artificial Intelligence in Medicine, an increase in the use of data mining and ML techniques was observed.

In recent years, other reviews have been published presenting the growth and potential of the use of ML methods in the health area. In their review, Rashidi et al [9] addressed the multidisciplinary aspect of this scenario and presented the potential of using ML techniques in data processing in the health area comparing the different methods.

Similarly, Ahmed et al [10] discussed aspects of precision medicine in their review, presenting works with different approaches to the use of ML in addition to discussing ethical aspects and the management of health resources.

However, the work by Houfani et al [11] focused on the prediction of diagnoses, presenting an overview of the methods applied in the prediction of diseases.

In their work, Ma et al [12] present aspects of real-world big data studies with a focus on laboratory medicine. In their review, Ma et al [12] highlighted the lack of standardization in clinical laboratories and the difficulty in using data in real time, mainly because of unstructured and unreliable data. However, the potential is emphasized in the use of laboratory data together with aspects such as the establishment of the reference range, quality control based on patient data, analysis of factors that affect analyte test results, establishment of diagnostic and prognostic models, epidemiological investigation, laboratory management, and data mining. All of this is aimed at helping traditional clinical laboratories develop into smart clinical laboratories.

In contrast to the studies presented, this study aimed to analyze studies that used data from laboratory tests together with AI techniques to predict new results.

Study Questions

Clinical laboratories display most test results as individual numerical values. However, the results of these tests, viewed in isolation, are usually of limited significance for reaching a diagnosis.

In their study of ferritin, Luo et al [5] found that laboratory tests often contain redundant information.

Similarly, Gunčar et al [13] found that ML models can predict hematological diseases using only blood tests. In their study, Gunčar et al [13] stated that laboratory tests have more information than health professionals commonly consider.

Demirci et al [14] and Rosenbaum and Baron [15] also used ML techniques to identify possible errors in the clinical process of performing laboratory tests. In both studies, the authors obtained satisfactory results, demonstrating the ability of computational models based on ML to assist in analyzing laboratory tests. Similarly, Baron et al [16] used an algorithm to generate a decision tree capable of identifying tests with possible problems arising from the preanalytical process during the execution of laboratory tests.

The presentation of these works makes us reflect on how much information can be present in a set of laboratory test data and the potential for the exploration and use of such data. Thus, our objective was to identify scientific studies that used laboratory tests and ML models to predict results.

This study had the following specific research questions: (1) Is it possible to predict specific examinations from other examinations? (2) Which examinations are typically used as input data to predict other results? and (3) What methods are used to predict these tests?

Methods

Search Strategy

Searches were conducted in 7 electronic databases in international journals in the areas of engineering and health sciences—Compendex (Engineering Village), EBSCO (MEDLINE complete), IEEE Xplore, PubMed (MEDLINE), ScienceDirect, Scopus, and Web of Science—in the English language for publications from April 2011 to February 2022. Additional records were further identified during the screening phase of this research by analyzing the references of the eligible articles included.

The population, intervention, comparison, and outcome principles were used to group the search terms. As this study addressed laboratory tests, 3 principal search terms were considered, and 2 Boolean operators were used (OR and AND):

population (“Clinical Laboratory Test” OR “Laboratory Diagnosis” OR “Blood Count, Complete” OR “Routine Diagnostic Test”) AND intervention (“Machine Learning”) AND outcomes (“Clinical Decision-Making” OR “Computer-Assisted Diagnosis” OR “Predictive Value of Tests”).

The search terms were defined based on the list of terms used in the Medical Subject Heading database [17]. The studies were collected from the databases from April 2, 2021, to April 10, 2021; the roots of the words and all the variants of the terms were searched (singular or plural, past tense, gerund, comparative adjective, and superlative, when possible). The following filters were used for the area of activity: medicine, engineering (industrial, biomedical, electrical, manufacturing, and mechanics), robotics, health professions, and multidisciplinary according to the availability in the database.

The following study characteristics were extracted and described: authors’ names, year of publication, title, description, data set, features, methods, and main results. The data of this study were presented descriptively and followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement flow diagram [18] and the National Institutes of Health (NIH) Quality Assessment Tool for Observational Cohort and Cross-Sectional Studies [19].

Inclusion and Exclusion Criteria

The criteria for inclusion and exclusion of studies are outlined in Textbox 1.

The search results were exported to the web-based Mendeley software (Elsevier), where duplicates or triplicates were removed, and full texts were extracted after analyzing the possible eligibility of the articles.

Textbox 1. Study inclusion and exclusion criteria.

Inclusion criteria

- Use of laboratory tests
- Use of machine learning techniques
- Written in English
- Full-text articles published in specialized journals

Exclusion criteria

- No use of laboratory tests
- Not seeking to predict new results

Study Analysis

Regarding the eligibility of the studies, the review process involved an analysis of the title keywords and reading of the abstracts by 2 reviewers independently (the first 2 authors of this paper). When in doubt about eligibility, the full text was reviewed. In cases of disagreement between the 2 reviewers, a decision was made by consensus or a third investigator provided an additional review, and the decision was made by arbitration.

Methodological Quality Assessment of the Studies

Regardless of the inclusion and exclusion criteria, which were directly related to the objective of the study, an analysis of the quality of the selected articles was also conducted.

The quality of the eligible studies was assessed using tools proposed by the NIH of the United States [19]. This study included the cross-sectional study assessment tool (with 14 criteria). The NIH website [19] provides tools and guidelines for assessing the quality of each type of study, containing explanatory information about each item that should be assessed in the study: (1) Was the research question or objective in this study clearly stated? (2) Was the study population clearly specified and defined? (3) Was the participation rate of eligible persons at least 50%? (4) Were all the participants selected or recruited from the same or similar populations (including the same period)? Were inclusion and exclusion criteria for being in the study prespecified and applied uniformly to all participants? (5) Was a sample size justification, power description, or variance and effect estimates provided? (6) For the analyses in this study, were the exposures of interest measured before the outcomes were measured? (7) Was the time frame sufficient so that one could reasonably expect to see an association between exposure and outcome if it existed? (8) For exposures that can vary in amount or level, did the study examine different levels of exposure as related to the outcome (eg, categories of exposure or exposure measured as a continuous variable)? (9) Were the exposure measures (independent variables) clearly defined, valid, reliable, and implemented consistently across all study participants? (10) Was the exposure assessed more than once over time? (11) Were the outcome measures (dependent variables) clearly defined, valid, reliable, and implemented consistently across all study participants? (12) Were the outcome assessors blinded to the exposure status of participants? (13) Was loss to follow-up after baseline 20% or less? and (14) Were key potential confounding variables measured and adjusted statistically for their impact on the relationship between exposure and outcome?

The rating quality was classified as good, fair, or bad, allowing for the general analysis of the evaluators considering all items [19]. Each item in the assessment tool received an “✓” rating when the study was performed, a negative (“–”) when not performed, and other options (cannot be determined, not applicable, and not reported).

According to Wong et al [20], observational studies with a classification of ≥67% of positive items indicated good quality, 34% to 66% of positive verifications indicated regular quality, and ≤33% indicated low quality.

Results

The search results included 513 potentially eligible studies. First, 8% (41/513) of duplicated or triplicated articles were excluded, and of the 472 remaining articles, 43 (9.1%) were considered eligible based on the review of titles, keywords, and abstracts. Additional studies (n=30) were included after searching the references and citations of the eligible articles, totaling 73 full texts for evaluation. After reviewing these 73 studies, 33 (45%) were ineligible, ending the process with 40 (55%) studies for quality assessment (Figure 1).

Table 1 presents the assessment of the methodological quality of the studies. The articles are organized by author and year, by framing of the questions, and by the average points obtained through this analysis performed by the authors of this paper.

Table 2 shows the description of the studies included in this review. It is organized by author and year, title, description, data set, features, methods, and main results.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of study screening and selection.

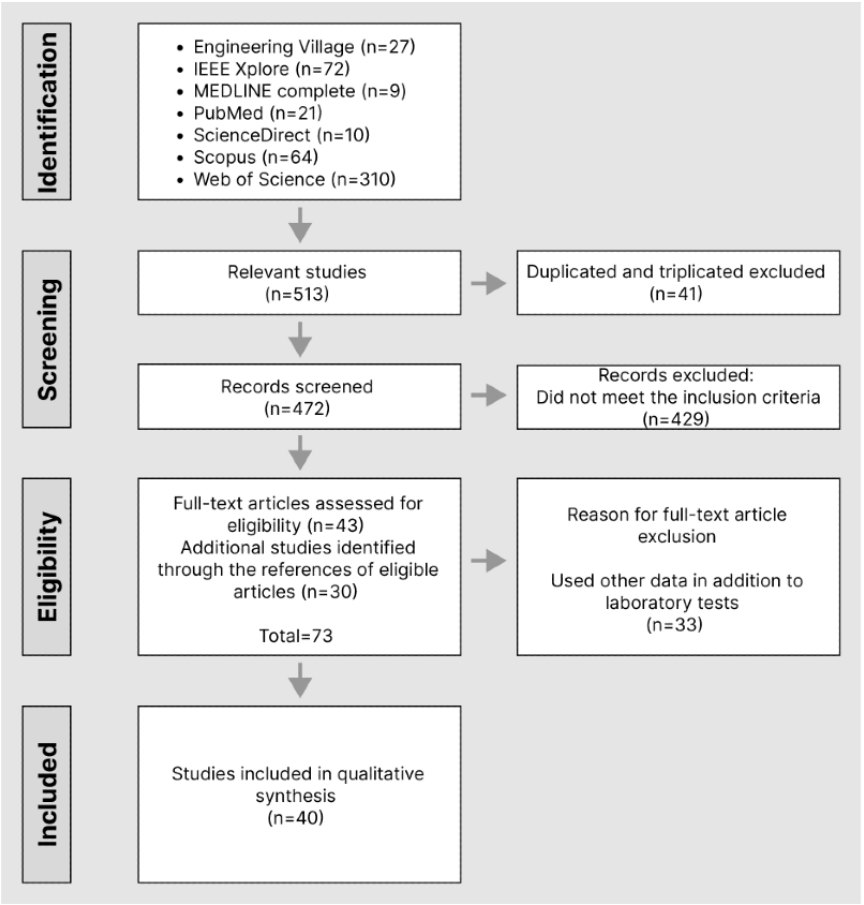


Table 1. Assessment of the methodological quality of the studies[a].

Author, year	Quality assessment tool items														Total assessment tool items, n (%)
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	
Richardson and Lidbury [21], 2013	✓	✓	✓	✓	✓	✓	✓	✓	✓	N/A[b]	✓	✓	✓	✓	13 (93)
Waljee et al [22], 2013	✓	✓	✓	✓	✓	✓	✓	✓	CD[c]	N/A	CD	✓	✓	✓	11 (79)
Kinar et al [23], 2016	✓	✓	✓	CD	✓	✓	✓	CD	✓	✓	✓	✓	✓	✓	12 (86)
Luo et al [5], 2016	✓	✓	✓	✓	✓	✓	✓	✓	✓	N/A	✓	✓	✓	✓	13 (93)
Razavian et al [24], 2016	✓	✓	✓	✓	✓	✓	✓	N/A	✓	✓	✓	✓	✓	✓	13 (93)
Richardson and Lidbury [25], 2017	✓	✓	✓	✓	✓	✓	✓	✓	✓	NR[d]	✓	✓	✓	✓	13 (93)
Birks et al [26], 2017	✓	✓	✓	✓	✓	✓	✓	✓	CD	N/A	✓	✓	✓	✓	12 (86)
Hernandez et al [27], 2017	✓	✓	✓	✓	✓	✓	CD	CD	✓	✓	✓	✓	✓	✓	12 (86)
Roy et al [7], 2018	✓	✓	✓	✓	✓	✓	✓	CD	✓	✓	✓	✓	✓	✓	13 (93)
Rawson et al [28], 2019	✓	✓	✓	✓	✓	✓	✓	✓	✓	N/A	✓	✓	✓	✓	13 (93)
Aikens et al [29], 2019	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	14 (100)
Hu et al [30], 2019	✓	✓	✓	✓	✓	✓	CD	N/A	✓	CD	✓	✓	✓	✓	11 (79)
Bernardini et al [31], 2019	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	14 (100)
Xu et al [32], 2019	✓	✓	✓	✓	✓	✓	✓	CD	✓	✓	✓	✓	✓	✓	13 (93)
Lai et al [33], 2019	✓	✓	✓	✓	✓	✓	✓	✓	✓	N/A	✓	✓	✓	✓	13 (93)
Tamune et al [34], 2020	✓	✓	✓	✓	✓	✓	CD	✓	✓	N/A	✓	✓	✓	✓	12 (86)
Chicco and Jurman [35], 2020	✓	✓	✓	✓	✓	✓	✓	✓	✓	N/A	✓	✓	✓	✓	13 (93)
Yu et al [36], 2020	✓	✓	✓	✓	✓	✓	✓	CD	NR	✓	✓	✓	✓	✓	12 (86)
Banerjee et al [37], 2020	✓	✓	✓	✓	✓	✓	✓	N/A	✓	N/A	✓	✓	✓	✓	12 (86)
Joshi et al [38], 2020	✓	✓	✓	✓	✓	✓	✓	N/A	CD	N/A	✓	✓	✓	✓	11 (79)
Brinati et al [39], 2020	✓	✓	✓	✓	✓	✓	✓	N/A	✓	N/A	✓	✓	✓	✓	12 (86)
Metsker et al [40], 2020	✓	✓	✓	✓	✓	✓	✓	✓	✓	N/A	✓	✓	✓	✓	13 (93)
AlJame et al [41], 2020	✓	✓	✓	✓	✓	✓	✓	✓	✓	N/A	✓	✓	✓	✓	13 (93)
Yadaw et al [42], 2020	✓	✓	✓	✓	✓	✓	✓	N/A	CD	N/A	✓	✓	✓	✓	11 (79)
Cabitza et al [43], 2020	✓	✓	✓	✓	✓	✓	✓	N/A	✓	N/A	✓	✓	✓	✓	12 (86)
Schneider et al [44], 2020	✓	✓	✓	✓	✓	✓	CD	✓	CD	N/A	✓	✓	✓	✓	11 (79)
Yang et al [45], 2020	✓	✓	✓	✓	✓	✓	✓	✓	✓	N/A	✓	✓	✓	✓	13 (93)

Author, year	Quality assessment tool items														Total assessment tool items, n (%)
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	
Plante et al [46], 2020	✓	✓	✓	✓	✓	✓	✓	CD	✓	N/A	✓✓	✓	✓	✓	12 (86)
Mooney et al [47], 2020	✓	✓	✓	✓	✓	✓	✓	CD	✓	✓	✓	✓	✓	✓	13 (93)
Yu et al [48], 2020	✓	✓	✓	✓	✓	✓	✓	✓	✓	N/A	✓	✓	✓	✓	13 (93)
Kaftan et al [49], 2021	✓	✓	✓	✓	✓	✓	✓	✓	✓	N/A	✓	✓	✓	✓	13 (93)
Park et al [50], 2021	✓	✓	✓	✓	✓	✓	CD	✓	CD	N/A	✓	✓	✓	✓	11 (79)
Souza et al [51], 2021	✓	✓	✓	✓	✓	✓	✓	✓	✓	N/A	✓	✓	✓	✓	13 (93)
Kukar et al [52], 2021	✓	✓	✓	✓	✓	✓	✓	✓	✓	N/A	✓	✓	✓	✓	13 (93)
Gladding et al [53], 2021	✓	✓	✓	✓	✓	✓	✓	N/A	CD	N/A	✓	✓	✓	✓	11 (79)
AlJame et al [41], 2021	✓	✓	✓	✓	✓	✓	✓	N/A	✓	N/A	✓	✓	✓	✓	12 (86)
Rahman et al [54], 2021	✓	✓	✓	✓	✓	✓	✓	N/A	✓	N/A	✓	✓	✓	✓	12 (86)
Myari et al [55], 2021	✓	✓	✓	✓	✓	✓	CD	✓	✓	✓	✓	✓	✓	✓	13 (93)
Campagner et al [56], 2021	✓	✓	✓	✓	✓	✓	✓	✓	✓	N/A	✓	✓	✓	✓	13 (93)
Babaei Rikan et al [57], 2022	✓	✓	✓	✓	✓	✓	✓	N/A	✓	N/A	✓	✓	✓	✓	12 (86)

[a]Quality rating: ≥67%=good, 33% to 66%=fair, and ≤33%=poor.

[b]N/A: not applicable.

[c]CD: cannot be determined.

[d]NR: not reported.

Table 2. Description of the studies included in this review (N=40).

Author, year	Title	Description	Data set	Features	Methods	Main results
Richardson and Lidbury [21], 2013	Infection status outcome, machine learning method and virus type interact to affect the optimised prediction of hepatitis virus immunoassay results from routine pathology laboratory assays in unbalanced data	This study investigated the effect of data preprocessing, the use of ensembles constructed by bagging, and a simple majority vote to combine classification predictions from routine pathology laboratory data, particularly to overcome a significant imbalance of negative HBV[a] and HCV[b] cases HBV or HCV immunoassay positive cases.	Used a data set of 18,625 records from 1997 to 2007 made available by ACT Pathology at The Canberra Hospital, ACT[c], Australia	Age, gender, and CBC[d] (FBC[e]) parameters	Implemented the analysis using the RPART[f] algorithm in R (DT[g])	It was easier to predict positive immunoassay cases than negative cases of HBV or HCV.
Waljee et al [22], 2013	Comparison of imputation methods for missing laboratory data in medicine	Compare the accuracy of 4 imputation methods for missing entirely at random laboratory data and compare the effect of the imputed values on the accuracy of 2 clinical predictive models	The cirrhosis cohort had 446 patients, and the inflammatory bowel disease cohort had 395 patients from a tertiary-level care institution in Ann Arbor, Michigan.	CBC (FBC) parameters	MissForest, mean imputation, nearest neighbor imputation, and MICE[h] to impute the simulated missing data	MissForest had the lowest imputation error for both continuous and categorical variables at each frequency of missingness, and it had the smallest prediction difference when models used imputed laboratory values.
Kinar et al [23], 2016	Development and validation of a predictive model for detection of colorectal cancer in primary care by analysis of complete blood counts: a binational retrospective study	Develop and validate a model to identify individuals at increased risk of CRC[i]	Used a data set of 2 million patients from the Maccabi Healthcare Services in Israel and the United Kingdom THIN[j]	Age, gender, and CBC (FBC) parameters	Gradient boosting model and RF[k] classifier	Mean ROC AUC[l] for detecting CRC was 0.82 (SD 0.01) for the Israeli validation set
Luo et al [5], 2016	Using Machine Learning to Predict Laboratory Test Results	Used ML[m] to predict ferritin values from laboratory test results	Used a data set of 5128 inpatients in a tertiary care hospital in Boston, Massachusetts, collected over 3 months in 2013	Age, gender, and 41 laboratory tests	It used LR[n], Bayesian LR, RFR[o], and lasso regression (lasso).	The model could predict ferritin results with high accuracy (AUC[p] as high as 0.97, held-out test data).
Razavian et al [24], 2016	Multi-task Prediction of Disease Onsets from Longitudinal Laboratory Tests	Using longitudinal measurements of laboratory tests, the study evaluated learning to predict disease onsets.	Used a data set from laboratory measurement and diagnosis information of 298,000 individuals from a larger cohort of 4.1 million insurance subscribers between 2005 and 2013	18 laboratory tests	The study trained an LSTM[q] RNN[r] and 2 novel CNNs[s] for multitask prediction of disease onset.	These representation-based approaches significantly outperformed an LR with several hand engineered, clinically relevant features.

Author, year	Title	Description	Data set	Features	Methods	Main results
Richardson and Lidbury [25], 2017	Enhancement of hepatitis virus immunoassay outcome predictions in imbalanced routine pathology data by data balancing and feature selection before the application of support vector machines	The impact of 3 balancing methods and 1 feature selection method was explored to assess the ability of SVMs[t] to classify imbalanced diagnostic pathology data associated with the laboratory diagnosis of HBV and HCV infections.	The data set used in this study originally comprised 18,625 individual cases of hepatitis virus testing over a decade, from 1997 to 2007.	Age, gender, and 26 laboratory tests	RFs	Generating data sets using the SMOTE[u] resulted in significantly more accurate prediction than single downsizing or MDS[v] of the data set.
Birks et al [26], 2017	Evaluation of a prediction model for colorectal cancer: retrospective analysis of 2.5 million patient records	Evaluate an existing risk algorithm derived in Israel that identifies individuals according to CRC risk using FBC data through CPRD[w] data from the United Kingdom	2,550,119 patients who were ≥40 years old from CPRD	Age, gender, and CBC test	Application of the algorithm in case-control analysis of patients undergoing FBC testing during 2012 to estimate predictive values	The algorithm offered an additional means of identifying risk of CRC and could support other approaches to early detection, including screening and active case finding.
Hernandez et al [27], 2017	Supervised learning for infection risk inference using pathology data	Evaluated the performance of different binary classifiers to detect any type of infection from a reduced set of commonly requested clinical measurements	Pathology and microbiology data of patients from all hospital wards at ICHNT[x] were extracted.	Alanine aminotransferase, alkaline phosphatase, bilirubin, creatinine, C-reactive proteins, and WBC[y]	Supervised ML algorithms for binary classification (Gaussian NB[z], DT classifier, RF classifier, and SVM)	ROC AUC (0.80-0.83), sensitivity (0.64-0.75), and specificity (0.92-0.97)
Roy et al [7], 2018	Predicting Low Information Laboratory Diagnostic Tests	The study described the prevalence of common laboratory tests in a hospital environment and the rate of “normal” results to quantify pretest probabilities under different conditions.	Electronic medical records (Epic) of 71,000 patients admitted to Stanford Tertiary Academic Hospital between the years 2008 and 2014	Common laboratory tests (eg, thyroid stimulating hormone, sepsis protocol lactate, ferritin, and NT-PROBNP[aa])	Provided a data-driven, systematic method to identify cases where the incremental value of testing is worth reconsidering	The study found that low-yield laboratory tests were common (eg, approximately 90% of blood cultures were normal).
Rawson et al [28], 2019	Supervised machine learning for the prediction of infection on admission to hospital: A prospective observational cohort study	An SML[ab] algorithm was developed to classify cases into infection versus no infection using microbiology records and 6 available blood parameters.	This study took place at ICHNT, comprising 3 university teaching hospitals. The study took place between October 2017 and March 2018 with 160,203 individuals.	C-reactive protein, WCC[ac], bilirubin, creatinine, ALT[ad], and alkaline phosphatase	A (SVM) binary classifier algorithm was developed and incorporated into the EPIC IMPOC[ae] CDSS[af] for investigation within this study following validation and pilot assessment.	The infection group had a likelihood of 0.80 (SD 0.09), and the noninfection group had a likelihood of 0.50 (0.29, 95% CI 0.20-0.40; $P<.01$). ROC AUC was 0.84 (95% CI 0.76-0.91).
Aikens et al [29]	A machine learning approach to predicting the stability of inpatient lab test results	Development of a predictive model that can identify low-information laboratory tests before they are ordered	Analyzed 6 years (2008-2014) of inpatient data from Stanford University Hospital, a tertiary academic hospital	Troponin, thyroid stimulating hormone, platelet count, phosphate in serum or plasma, partial thromboplastin time, NT-PROBNP, magnesium, lipase, lactase, heparin activity, ferritin, creatinine kinase, and C-reactive protein	Six different ML models for classification: a DT, a boosted tree classifier (AdaBoost), an RF, a Gaussian NB classifier, a lasso-regularized LR, and a linear regression followed by rounding to 0 or 1	A large proportion of repeat tests were within an SD of 10% or 0.1 of the previous measurement, indicating that a large volume of repetitive testing may be contributing little new information.

Author, year	Title	Description	Data set	Features	Methods	Main results
Hu et al [30], 2019	Using Biochemical Indexes to Prognose Paraquat-Poisoned Patients: An Extreme Learning Machine-Based Approach	Explore useful indexes from biochemical tests and identify their predictive value in prognosis of patients poisoned with PQ[ag]	The biochemical indexes of 101 patients poisoned with PQ who were hospitalized in the emergency room of First Affiliated Hospital of Wenzhou Medical University from 2013 to 2017	Total bilirubin, direct bilirubin, indirect bilirubin, total protein, albumin, albumin-globulin ratio, alanine aminotransferase, aspartate aminotransferase, the ratio of AST[ah] to ALT, blood glucose, urea nitrogen, and creatinine	An effective ELM[ai] model was developed for classification tasks.	A new method for prognosis of PQ poisoning with accuracy of 79.6%
Bernardini et al [31], 2019	TyG-er: An ensemble Regression Forest approach for identification of clinical factors related to insulin resistance condition using Electronic Health Records	The study aimed to discover nontrivial clinical factors in EHR[aj] data to determine where the insulin resistance condition is encoded.	A total of 2276 records from 968 patients not affected by T2D[ak]; the longitudinal patient observational period was from 2010 to 2018 (FIMMG_obs data set)	Gender, age, blood pressure, height, weight, and 73 laboratory exams	Highly interpretable ML approach (ie, ensemble regression forest combined with data imputation strategies), named TyG-er	High agreement (from 0.664 to 0.911 of the Lin correlation coefficient) of the TyG-er and predictive power of the TyG-er approach (up to a mean absolute error of 5.68% and correlation coefficient=0.666; P<.05)
Xu et al [32], 2019	Prevalence and Predictability of Low-Yield Inpatient Laboratory Diagnostic Tests	Identify inpatient diagnostic laboratory testing with predictable results that are unlikely to yield new information	A total of 116,637 inpatients treated at Stanford University Hospital from January 2008 to December 2017; 60,929 inpatients treated at the University of Michigan from January 2015 to December 2018; and 13,940 inpatients treated at the University of California, San Francisco from January 2018 to December 2018 were assessed.	The core features included patient demographics, change of the most recent test, number of recent tests, history of Charlson Comorbidity Index categories, which specialty team was treating the patient, time since admission, statistical data, and laboratory test results.	Regularized LR, regression and round, NB, NN[al] multilayer perceptrons, DT, RF, AdaBoost, and XGB[am]	The findings suggest that low-yield diagnostic testing is common and can be systematically identified through data-driven methods and patient context–aware predictions.
Lai et al [33], 2019	Predictive models for diabetes mellitus using machine learning techniques	The objective of this study was to build an effective predictive model with high sensitivity and selectivity to better identify Canadian patients at risk of having diabetes mellitus based on patient demographic data and the laboratory test results during their visits to medical facilities.	13,309 Canadian patients aged between 18 and 90 years	Age, sex, fasting blood glucose, BMI, high-density lipoprotein, triglycerides, blood pressure, and low-density lipoprotein	Predictive models using LR and GBM[an] techniques	The ROC AUC for the proposed GBM model was 84.7% with a sensitivity of 71.6%, and the ROC AUC for the proposed LR model was 84% with a sensitivity of 73.4%.
Tamune et al [34], 2020	Efficient Prediction of Vitamin B Deficiencies via Machine-Learning Using Routine Blood Test Results in Patients with Intense Psychiatric Episode	Predict vitamin B deficiency using ML models from patient characteristics and routine blood test results that can be obtained within 1 hour	Reviewed 497 patients admitted to the Department of Neuropsychiatry at Tokyo Metropolitan Tama Medical Center between September 2015 and August 2017	Age, sex, and 29 routine blood tests	ML models (KNN[ao], LR, SVM, and RF)	The study demonstrated that ML can efficiently predict some vitamin deficiencies in patients with active psychiatric symptoms.

Author, year	Title	Description	Data set	Features	Methods	Main results
Chicco and Jurman [35], 2020	Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone	ML in particular can predict patients' survival from their data and individuate the most important features among those included in their medical records.	Medical records of 299 patients with heart failure collected at the Faisalabad Institute of Cardiology and the Allied Hospital in Faisalabad (Punjab, Pakistan) from April 2015 to December 2015	Age, anemia, high blood pressure, creatinine phosphokinase, diabetes, ejection fraction, sex, platelets, serum creatinine, serum sodium, smoking, and follow-up period	Apply several ML classifiers to both predict the patient's survival and rank the features corresponding to the most important risk factors	The results of these 2-feature models show not only that serum creatinine and ejection fraction are sufficient to predict survival of patients with heart failure from medical records but also that using these 2 features alone can lead to more accurate predictions than using the original data set features in their entirety.
Yu et al [36], 2020	Predict or draw blood: An integrated method to reduce lab tests	Propose a novel DL[ap] method to jointly predict future laboratory test events to be omitted	The data set (MIMIC III) contained 598,444 laboratory test results and 5,598,079 vital sign records from a total of 41,113 adult patients (aged ≥16 years) admitted to critical care units between 2001 and 2012.	Sodium, potassium, chloride and serum bicarbonate, total calcium, magnesium, phosphate, BUN[aq], creatinine, hemoglobin, platelet count, and WBC.	The study ran a novel DL method combining 4 features: lab (laboratory test data), D (demographic data), V (vital data, which were mean and SD in the vicinity of the corresponding laboratory tests), and C (encoding to indicate missing values).	Was able to omit 15% of laboratory tests with <5% prediction accuracy loss
Banerjee et al [37], 2020	Use of Machine Learning and Artificial Intelligence to predict SARS-CoV-2 infection from Full Blood Counts in a population	The aim of the study was to use ML, an ANN[ar], and a simple statistical test to identify patients who were SARS-CoV-2–positive from FBCs without knowledge of symptoms or history of the individuals.	The data set included in the analysis and training contained anonymized FBC results from 5664 patients seen at the Hospital Israelita Albert Einstein (São Paulo, Brazil) from March 2020 to April 2020 and who had samples collected to perform the SARS-CoV-2 RT-PCR[as] test during a visit to the hospital.	Age and CBC (FBC) parameters	RF and lasso-based regularized generalized linear models and ANN	The study found that, with FBCs, RF, shallow learning, and a flexible ANN model predict patients with SARS-CoV-2 with high accuracy between populations on regular wards (AUC=94%-95%) and those not admitted to the hospital or in the community (AUC=80%-86%).
Joshi et al [38], 2020	A predictive tool for identification of SARS-CoV-2 PCR-negative emergency department patients using routine test results	Predict SARS-CoV-2 PCR[at] positivity based on CBC components and patient sex	357 CBC data from January 2020 to March 2020 ordered within 24 hours of a SARS-CoV-2 PCR test (based off the WHO[au] assay)	Absolute neutrophil count, absolute lymphocyte count, and hematocrit	The study trained an L2[av]-regularized LR model.	Prediction of SARS-CoV-2 PCR positivity demonstrated a C-statistic of 78% and an optimized sensitivity of 93%.
Brinati et al [39], 2020	Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study	Develop a predictive model based on ML techniques to predict positivity or negativity for COVID-19	Data set available from the IRCCS[aw] Ospedale San Raffaele 2 with 279 cases randomly extracted from the end of February 2020 to mid-March 2020	Gender, age, leukocytes, platelets, C-reactive protein, transaminases, gamma-glutamyltransferase, lactate dehydrogenase, neutrophils, lymphocytes, monocytes, eosinophils, and basophils	DT, ETs[ax], KNN, LR, NB, RF, and SVMs	Their accuracy ranged from 82% to 86%, and sensitivity ranged from 92% to 95%.

Author, year	Title	Description	Data set	Features	Methods	Main results
Metsker et al [40], 2020	Identification of risk factors for patients with diabetes: diabetic polyneuropathy case study	Implementation of ML methods for identifying the risk of diabetes polyneuropathy based on structured electronic medical records collected from databases of medical information systems	Laboratory records from 5425 patients between 2010 and 2017	16 laboratory tests plus a CBC	ANN, SVM, DT, linear regression, and LR classifier	79.82% precision, 81.52% recall, 80.64% F_1-score, 82.61% accuracy, and 89.88% AUC using the NN classifier
AlJame et al [41], 2020	Ensemble learning model for diagnosing COVID-19 from routine blood tests	The study proposed ERLX, which is an ensemble learning model for COVID-19 diagnosis from routine blood tests.	The study used 5644 data samples with 559 confirmed COVID-19 cases from a publicly available data set from Albert Einstein Hospital in Brazil.	24 laboratory tests, including INR[ay], albumin, D-dimer, and prothrombin time	The proposed model used 3 classifiers—extra trees, RF, and LR—combining their predictions with an XGB.	The ensemble model achieved outstanding performance, with an overall accuracy of 99.88%, AUC of 99.38%, sensitivity of 98.72%, and specificity of 99.99%.
Yadaw et al [42], 2020	Clinical Predictive Models for COVID-19: Systematic Study	The aim of this study was to develop, study, and evaluate clinical predictive models that estimate, using ML and based on routinely collected clinical data, which patients are likely to receive a positive SARS-CoV-2 test or require hospitalization or intensive care.	The study used anonymized data from a cohort of 5644 patients seen at the Hospital Israelita Albert Einstein in São Paulo, Brazil, in the early months of 2020.	The study used 106 routine clinical, laboratory, and demographic measurements.	LR, NN, RF, SVM, and gradient boosting (XGB)	Predicted positive tests for SARS-CoV-2 a priori at a sensitivity of 75% and a specificity of 49%, patients who were SARS-CoV-2–positive who required hospitalization with 0.92 AUC, and patients who were SARS-CoV-2–positive who required critical care with 0.98 AUC
Cabitza et al [43], 2020	Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests	Routine blood tests can be exploited using the authors' method to diagnose COVID-19.	1925 patients on admission to the ED[az] at the San Raffaele Hospital (OSR[ba]) from February 2020 to May 2020	72 features: CBC, biochemical, coagulation, hemogas analysis and CO-oximetry values, age, sex, and specific symptoms at triage	RF, NB, LR, SVM, and KNN	For the complete OSR data set, the AUC for the algorithms ranged from 0.83 to 0.90; for the COVID-19–specific data set, it ranged from 0.83 to 0.87.
Schneider et al [44], 2020	Validation of an Algorithm to Identify Patients at Risk for Colorectal Cancer Based on Laboratory Test and Demographic Data in Diverse, Community-Based Population	Validate a predictive score generated by an ML algorithm with common laboratory test data to identify patients at high risk of CRC in a large, community-based, ethnically diverse cohort	The eligible study cohort population included 2,855,994 KPNC[bb] Health Plan members between 1996 and 2015.	Gender, year of birth, and at least one CBC test, including cell parameters	Validate the ability of an algorithm that uses laboratory and demographic information to identify patients at increased risk of CRC	The algorithm identified 3% of the population who required an investigation and 35% of patients who received a diagnosis of CRC within the following 6 months.
Yang et al [45], 2020	Routine Laboratory Blood Tests Predict SARS-CoV-2 Infection Using Machine Learning	Develop an ML model integrating age, gender, race, and routine laboratory blood tests, which are readily available with a short TAT[bc]	5893 patients evaluated at the NYPH[bd] and WCM[be] from March 2020 to April 2020	26 laboratory tests, including C-reactive protein, ferritin, lactic acid dehydrogenase, and magnesium	Used a GBDT[bf] model	The model achieved an AUC of 0.854. The model, too, predicted initial SARS-CoV-2 RT-PCR positivity in 66% of individuals whose RT-PCR result changed from negative to positive within 2 days.

Author, year	Title	Description	Data set	Features	Methods	Main results
Plante et al [46], 2020	Development and External Validation of a Machine Learning Tool to Rule Out COVID-19 Among Adults in the Emergency Department Using Routine Blood Tests: A Large, Multicenter, Real-World Study	Develop an ML model to rule out COVID-19 using only routine blood tests among adults in EDs	Model training used 2183 PCR-confirmed cases from 43 hospitals during the pandemic; negative controls were 10,000 prepandemic patients from the same hospitals. External validation used 23 hospitals with 1020 PCR-confirmed cases and 171,734 prepandemic negative controls.	14 laboratory tests, including sodium, bicarbonate, BUN, and chloride	XGB ML model	The model found high discrimination across age, race, sex, and disease severity subgroups and had high diagnostic yield at low score cutoffs in a screening population with a disease prevalence of <10%. Such a model could rapidly identify those at low risk of COVID-19 in a "rule out" method and might reduce the need for PCR testing in such patients.
Mooney et al [47], 2020	Predicting bacteraemia in maternity patients using full blood count parameters: A supervised machine learning algorithm approach	Use ML tools to identify if bacteremia in pregnant or postpartum women could be predicted using FBC parameters other than the WCC	129 women from the Rotunda Hospital in 2019, a stand-alone tertiary-level maternity hospital in Ireland	WCC, absolute neutrophils, lymphocytes, monocytes, eosinophils, basophils, NLR[bg], platelets, MPV[bh], MPV to platelet ratio, and monocyte to lymphocyte ratio	LDA[bi], KNN, SVM with a linear kernel, and RF along with CART[bj]	Sensitivity of 27.9% (95% CI 20.3-36.4), specificity of 94.1% (95% CI 93.3-94.8), PPV[bk] of 13.9% (95% CI 10.6-17.9), and NPV[bl] of 97.4% (95% CI 97.2-97.7)
Yu et al [48], 2020	A deep learning solution to recommend laboratory reduction strategies in ICU	Build an ML model that predicts laboratory test results and provides a promising laboratory test reduction strategy using spatial-temporal correlations	The Medical Information Mart for Intensive Care III data set with 53,423 distinct hospital admissions of adult patients to intensive care units at Beth Israel Deaconess Medical Center	Sodium, potassium, chloride, serum bicarbonate, total calcium, magnesium, phosphate, BUN, creatinine, hemoglobin, platelet count, WBC, age, gender, and race	Built a DL model with 5 variants for each of the combinations of input features	The model predicted normality or abnormality of laboratory tests with a 98.27% accuracy (AUC=0.9885; sensitivity 97.84%; specificity 98.8%; PPV=99.01%; NPV=97.39%) on 20.26% reduced laboratory tests and recommended 98.1% of transitions to be checked.
Kaftan et al [49], 2021	Predictive Value of C-reactive Protein, Lactate Dehydrogenase, Ferritin and D-dimer Levels in Diagnosing COVID-19 Patients: a Retrospective Study	The study aimed to evaluate the diagnostic accuracy of CRP[bm], ferritin, LDH[bn], and D-dimer in predicting positive cases of COVID-19 in Iraq.	The sample size was based on a minimum sensitivity and specificity of 95%; the study randomly selected medical records of 938 patients suspected to have COVID-19 between May 2020 and December 2020.	Age, gender, C-reactive protein, ferritin, LDH, and D-dimer.	A retrospective observational cohort study based on STARD[bo] guidelines to determine the diagnostic accuracy of COVID-19	A combination of routine laboratory biomarkers (CRP, LDH, and ferritin ±D-dimer) can be used to predict the diagnosis of COVID-19 with an accepted sensitivity and specificity before proceeding to definitive diagnosis through RT-PCR.

Author, year	Title	Description	Data set	Features	Methods	Main results
Park et al [50], 2021	Development of machine learning model for diagnostic disease prediction based on laboratory tests	Build a new optimized ensemble model by blending a DNN[bp] model with 2 ML models for disease prediction using laboratory test results	The study analyzed data sets provided by the Department of Internal Medicine from 5145 patients visiting the emergency room and those admitted to Catholic University of Korea St. Vincent's Hospital in Suwon, Korea, between 2010 and 2019.	The study confirmed a total of 88 attributes, including sex and age.	The study developed a new ensemble model by combining their DL (DNN) model with their 2 ML models (SVM and RF) to improve AI[bq] performance.	The optimized ensemble model achieved an F_1-score of 81% and a prediction accuracy of 92% for the 5 most common diseases.
Souza et al [51], 2021	Simple hemogram to support the decision-making of COVID-19 diagnosis using clusters analysis with self-organising maps neural network	Identify potential variables in routine blood tests that can support clinician decision-making during COVID-19 diagnosis at hospital admission	5644 patients allocated to the Albert Einstein Hospital in São Paulo, Brazil, in the Kaggle platform on March 2020	14 variables present in the blood test	Nonsupervised clustering analysis with NN SOM[br] as a strategy of decision-making	It was possible to detect a group of units of the map with a discrimination power of approximately 83% to patients who were SARS-CoV-2–positive.
Kukar et al [52], 2021	COVID-19 diagnosis by routine blood tests using machine learning	The aim of this study was to determine the diagnostic accuracy of an ML model built specifically for the diagnosis of COVID-19 using the results of routine blood tests.	52,306 patients admitted to the Department of Infectious Diseases, UMCL[bs], Slovenia, in March 2020 and April 2020	Age, gender, and 35 laboratory tests	SBA[bt] algorithm: a CRISP-DM[bu]–based ML pipeline consisting of 5 processing stages and using an XGB model	The model exhibited a high sensitivity of 81.9%, a specificity of 97.9%, and an AUC of 0.97.
Gladding et al [53], 2021	A machine learning PROGRAM to identify COVID-19 and other diseases from haematology data	The study proposed a method for screening FBC metadata for evidence of communicable and noncommunicable diseases using ML.	A total of 156,570 hematology raw data were collected between July 2019 and June 2020 from Waitakere Hospital and North Shore Hospital.	A maximum of 247 FBC features from CSV[bv] data were used; 134 were categorical, and 101 were numeric.	MDCalc software was used to analyze and apply ML models using DTs and ensembles, LR, and DNNs.	Urinary tract infection: ROC AUC=0.68, sensitivity=52%, and specificity=79%; COVID-19: ROC AUC=0.8, sensitivity=82%, and specificity=75%; heart failure: ROC AUC=0.78, sensitivity=72%, and specificity=72%
AlJame et al [41], 2021	Deep forest model for diagnosing COVID-19 from routine blood tests	Develop an ML prediction model to accurately diagnose COVID-19 from clinical or routine laboratory test data	5644 patient records that were collected from March 2020 to April 2020 (Albert Einstein Israelita Hospital, located in São Paulo, Brazil) and 279 patients who were admitted to San Raffaele Hospital, Milan, Italy, from the end of February 2020 to mid-March 2020	Age, gender, and 13 laboratory tests	DF[bw] model constructed from 3 different classifiers: extra trees, XGB, and LightGBM	Experimental results show that the proposed DF model has an accuracy of 99.5%, sensitivity of 95.28%, and specificity of 99.96%.

Author, year	Title	Description	Data set	Features	Methods	Main results
Rahman et al [54], 2021	Mortality Prediction Utilising Blood Biomarkers to Predict the Severity of COVID-19 Using Machine Learning Technique	Development of a prediction model of high mortality risk for patients both with and without COVID-19	654 patients with and without COVID-19 were admitted to the ED in Boston (March 2020 to April 2020) and Tongji Hospital in China (January 2020 to February 2020).	Age, lymphocyte count, D-dimer, CRP, and creatinine	RF, SVM, KNN, XGB, extra trees, and LR	For the development cohort and the internal and external validation cohorts using LR, the AUCs were 0.987, 0.999, and 0.992, respectively.
Myari et al [55], 2021	Diagnostic value of white blood cell parameters for COVID - 19: Is there a role for HFLC and IG?	Investigate the ability of WBC and its subsets, HFLC[bx], IG[by], and C-reactive protein to aid diagnosis of COVID-19 during the triage process and as indicators of disease progression to serious and critical condition	A retrospective case-control study conducted with data collected from patients admitted to the ED of University General Hospital of Ioannina (Ioannina, Epirus, Greece) from March 2020 to March 2021	Age, gender, and 13 laboratory tests	Enter binary LR analysis was conducted to determine the influence of the parameters on the outcome.	The combined WBC-HFLC marker was the best diagnostic marker for both mild and serious disease. CRP and lymphocyte count were early indicators of progression to serious disease, whereas WBC, NEUT[bz], IG, and the NLR were the best indicators of critical disease.
Campagner et al [56], 2021	External validation of Machine Learning models for COVID-19 detection based on Complete Blood Count	Evaluate whether ML models for COVID-19 diagnosis based on CBC data could be robust to cross-site transportability and, thus, could be reliably deployed as medical decision support tools	Data from 1736 patients collected at the EDs of the IRCCS Hospital San Raffaele and the IRCCS Istituto Ortopedico Galeazzi of Milan (Italy)	Age, gender, and 23 routine laboratory tests	RF, LR, KNN, SVM, NB, and ensemble	The study reported an average AUC of 95%. The best-performing model (SVM) reported an average AUC of 97.5%.

Author, year	Title	Description	Data set	Features	Methods	Main results
Babaei Rikan et al [57], 2022	COVID-19 diagnosis from routine blood tests using artificial intelligence techniques	The study presented the development and comparison of various models for diagnosing positive cases of COVID-19 using 3 data sets of routine laboratory blood tests.	A total of 3 open-access study data sets from 2498 patients containing routine blood test data from COVID-19 and non–COVID-19 cases were used.	Routine laboratory tests according to each of the 3 data sets	Seven ML methods —LR, KNN, DT, SVM, NB, ET, RF. In addition to XGB —along with 4 DL methods: DNN, CNN, RNN, and LSTM	On average, accuracy, specificity, and AUC were 92.11%, 84.56%, and 92.2% for the first data set; 93.16%, 93.02%, and 93.2% for the second data set; and 92.5%, 85%, and 92.2% for the third data set, respectively.

[a]HBV: hepatitis B virus.
[b]HCV: hepatitis C virus.
[c]ACT: Australian Capital Territory.
[d]CBC: complete blood count.
[e]FBC: full blood count.
[f]RPART: Recursive Partitioning.
[g]DT: decision tree.
[h]MICE: Multivariate Imputation by Chained Equations.
[i]CRC: colorectal cancer.
[j]THIN: The Health Improvement Network.
[k]RF: random forest.
[l]ROC AUC: area under the receiver operating characteristic curve.
[m]ML: machine learning.
[n]LR: logistic regression.
[o]RFR: RF regression.
[p]AUC: area under the curve.
[q]LSTM: long short-term memory.
[r]RNN: recurrent neural network.
[s]CNN: convolutional neural network.
[t]SVM: support vector machine.
[u]SMOTE: Synthetic Minority Over-sampling Technique.
[v]MDS: multiple downsizing.
[w]CPRD: Clinical Practice Research Datalink.
[x]ICHNT: Imperial College Healthcare National Health Service Trust.
[y]WBC: white blood count.
[z]NB: naïve Bayes.
[aa]NT-PROBNP: N-terminal pro–brain natriuretic peptide.
[ab]SML: supervised machine learning.
[ac]WCC: white cell count.
[ad]ALT: alanine aminotransferase.
[ae]EPIC IMPOC: Enhanced, Personalized, and Integrated Care for Infection Management at the Point-of-Care.
[af]CDSS: clinical decision support system.
[ag]PQ: Paraquat.
[ah]AST: aspartate transaminase.
[ai]ELM: extreme learning machine.
[aj]EHR: electronic health record.
[ak]T2D: type 2 diabetes.
[al]NN: neural network.
[am]XGB: extreme gradient boosting.
[an]GBM: gradient boosting machine.
[ao]KNN: k-nearest neighbor.
[ap]DL: deep learning.
[aq]BUN: blood urea nitrogen.
[ar]ANN: artificial NN.

[as]RT-PCR: reverse transcription polymerase chain reaction.
[at]PCR: polymerase chain reaction.
[au]WHO: World Health Organization.
[av]L2: L2-penalization.
[aw]IRCCS: Scientific Institute for Research, Hospitalization and Healthcare.
[ax]ET: extremely randomized trees.
[ay]INR: international normalized ratio.
[az]ED: emergency department.
[ba]OSR: San Raphael Hospital.
[bb]KPNC: Kaiser Permanente Northern California.
[bc]TAT: turnaround time.
[bd]NYPH: New York Presbyterian Hospital.
[be]WCM: Weill Cornell Medicine.
[bf]GBDT: gradient boosting DT.
[bg]NLR: neutrophil to lymphocyte ratio.
[bh]MPV: mean platelet volume.
[bi]LDA: linear discriminant analysis.
[bj]CART: classification and regression trees.
[bk]PPV: positive predictive value.
[bl]NPV: negative predictive value.
[bm]CRP: C-reactive protein.
[bn]LDH: lactate dehydrogenase.
[bo]STARD: Standards for the Reporting of Diagnostic Accuracy Studies.
[bp]DNN: deep NN.
[bq]AI: artificial intelligence.
[br]SOM: self-organizing map.
[bs]UMCL: University Medical Centre Ljubljana.
[bt]SBA: Smart Blood Analytics.
[bu]CRISP-DM: cross-industry process for data mining.
[bv]CSV: comma-separated value.
[bw]DF: deep forest.
[bx]HFLC: high-fluorescence lymphocyte cell.
[by]IG: immature granulocyte count.
[bz]NEUT: neutrophil count.

Discussion

Principal Findings

This study aimed to identify studies that used laboratory tests to predict new results. Our interest in this line of study was motivated by the possibility that laboratory tests can be used more comprehensively to search for hidden information, discovering previously unknown pathologies. This methodology is highly advantageous for the diagnostic process of medical laboratories. In this sense, intelligent systems could automatically analyze the examinations performed on a patient and make predictions in the search for hidden pathologies. In positive cases, alarms would be generated, and complementary examinations would be suggested. In most cases, the collected sample could be used to carry out new tests.

The use of laboratory tests to predict results has been increasingly explored. In recent years, several studies have obtained good results using clinical data to search for diagnoses [58]. In addition to laboratory tests, the studies in this review used patient histories, imaging tests, and medical diagnoses. For example, Wu et al [59] and Hische et al [60], in addition to laboratory tests, also made use of other clinical data in the search for a diagnosis. Some studies, such as those by Ravaut et al [61] and Le et al [62], aimed to determine whether a patient was likely to develop the disease in the future, which is quite relevant as part of a process in predictive medicine. These studies obtained good results but used clinical or diagnostic data. This information is generated through the analysis by a physician, unlike most laboratory tests such as the complete blood count, which follows an automated analytical process without the intervention of human factors.

However, in this research, we only looked for studies that emphasized laboratory tests to predict new information. This methodology can innovate the diagnostic processes of medical laboratories and has attracted the interest of several researchers over time, especially in recent years owing to the COVID-19 pandemic. In total, we found 40 studies referring to the last decade that met the established criteria, with most studies published in 2020 (15/40, 38%) and 2021 (10/40, 25%).

All (40/40, 100%) the studies presented in this review used laboratory tests as input data in addition to some clinical data such as gender and age. Some (12/40, 30%) studies used >20

parameters, such as the study by Yadaw et al [42], who used >100 different parameters. Others (6/40, 15%) used very few parameters, as is the case of the work by Joshi et al [38], who used only 3 parameters (absolute neutrophil count, absolute lymphocyte count, and hematocrit). However, most (22/40, 55%) studies used approximately 10 parameters, with the complete blood count as the primary data source. Finally, 22% (9/40) of the studies used full blood count data only.

When analyzing the quality assessment tool (Table 1), all studies showed good results, with an average value of 88%. As most of the studies were characterized as retrospective cohort studies, the data used were generated before the research. Thus, questions 8 and 10 of the questionnaire [19], referring to the levels and amount of exposure, were answered mainly with *not applicable* or *cannot be determined*. This fact lowered the average slightly in the evaluation process of most (38/40, 95%) studies. However, 5% (2/40) of the studies [29,31] were evaluated with 100%. Another 45% (18/40) of the studies were evaluated with 93%, 32% (13/40) of the studies were evaluated with 86%, and 18% (7/40) of the studies were evaluated with 79%.

Table 2 presents a summary of the main characteristics of the studies. In addition to a brief description of the research, it is possible to know the methodology and the main results in a simplified way.

It is not possible to make a comparison between the methodology and results of the selected studies as they had different objectives. Our goal was to confirm the possibility of predicting specific examinations from other examinations and which ML methods and parameters were most used.

Regarding the models, most (39/40, 98%) studies used ML methods with supervised training, almost always aiming at the exam responsible for the diagnosis. Of the 40 studies selected, only 3 (8%) used regression methods, whereas the other 37 (92%) used classification methods. Among the most used models, we can mention logistic regression, random forest, support vector machine, and k-nearest neighbor, trained as binary classifiers. In the case of neural networks, they were almost always used with deep learning techniques (deep neural networks [DNNs]).

The random forest method was the most tested, with 50% (20/40) of the studies using it. The next most tested methods were logistic regression with 45% (18/40) of the studies and support vector machine with 35% (14/40) of the studies, followed by naïve Bayes, decision tree, and XGBoost with 25% (10/40) of the studies each. By contrast, artificial neural networks were tested in 18% (7/40) of the studies, in addition to DNN methods in another 15% (6/40) of the studies.

In general, the most efficient method was the DNN, such that, of the 6 studies that used this method, 5 (83%) had better results with it. Next, there was the XGBoost method, such that, of the 10 studies that used this method, 7 (70%) considered it better, followed by random forest, where, of the 20 studies that tested this method, 12 (60%) had better results with it. In a simplified way, we can say that the DNN method was 83% better than the others, followed by XGBoost (70% better) and random forest (60% better).

Although the DNN model presents better results, the random forest method is quite attractive, not only because it is simple and fast but also because it presents the path taken in the search for the result, which is quite relevant in research in the health care domain.

Research that initially caught our attention was conducted by Luo et al [5] to predict ferritin levels to detect patients with anemia. The research used 41 laboratory tests from 989 patients admitted to the tertiary care hospital in Boston, Massachusetts, for 3 months in 2013. The work had good results, with an area under the curve (AUC) of 97%. The most interesting thing is that, even in cases where the ferritin tests were false negatives, the system could detect anemia. This result shows that laboratory tests may have more information when analyzed holistically than when referring to the specific test performed.

Rawson et al [28] used laboratory tests to identify cases of bacterial infection among 160,203 hospitalized patients over 6 months. An interesting feature of this research is that only 6 tests were used as input parameters (C-reactive protein, white blood cell count, bilirubin, creatinine, alanine aminotransferase, and alkaline phosphatase), achieving good results, with an area under the receiver operating characteristic curve of 0.84. The use of a low number of examinations was an important factor in building the model. This situation makes it possible to use tests already performed on patients, making the screening process fast and straightforward without collecting more blood samples from a patient.

Of the selected studies, 8% (3/40) focused on the prediction of colorectal cancer. Colorectal cancer has a high incidence rate, accounting for many deaths worldwide. The early identification of this type of pathology can be very advantageous to governments and health systems, who can provide adequate treatment to prevent the worsening of the disease. Kinar et al [23] obtained good results in identifying patients with a propensity to develop colorectal cancer 1 year before the development of the disease. In this study, 20 parameters from the complete blood count of approximately 2 million patients were used. Similarly, Birks et al [26] used the complete blood count of 2.5 million patients, obtaining an AUC of 75% for more extended periods (3 years) and 85% for shorter periods (6 months). More recently, Schneider et al [44] also obtained a mean AUC of 78% in a study of approximately 2.8 million patients seen between 1996 and 2015.

Another 12% (5/40) of the studies [7,29,32,36,48] aimed to identify tests that would not change over time, remaining classified as normal without the need to be repeated. In general, all of them showed good results; however, we highlight the work by Xu et al [32], who obtained an AUC of >90% for 12 months of analysis.

A recent publication that also caught our attention was the work by Park et al [50]. The authors used deep learning models to predict 39 different diseases in their research, reaching an accuracy of >90% and an F_1-score of 81% for the 5 most

common diseases. They used 88 features from 5145 patients who visited the emergency room.

The use of laboratory tests and ML techniques has increased in recent years, mainly owing to the COVID-19 pandemic. Of the 40 studies in this review, 27 (68%) published between 2020 and 2022 were selected. Of these 27 studies, 19 (70%) studies were related to SARS-CoV-2, a total of 8 (30%) studies were published in 2020, a total of 9 (33%) studies were published in 2021, and 1 (4%) study was published in 2022. All of them used laboratory tests to predict some unknown information, and most (34/40, 85%) studies focused on the search for a diagnosis.

Analyzing aspects related to training and the potential for bias based on the data sets, a common feature among most studies was the fact that 92% (37/40) of them were treated as a classification problem using supervised models. In this process, a point to be considered is the fact that the target classes of the models are almost always defined by a medical diagnosis or a reference value. In class prediction, the results of values close to the classification margins may be affected, influencing the final result of the model.

Another aspect that draws attention is the fact that the data sets were highly unbalanced, with some (3/40, 8%) studies [21,23,26] where the target represented <1% of the data set, implying some care to avoid errors in the training and evaluation process. In this sense, most (34/40, 85%) of the analyzed studies used the area under the receiver operating characteristic curve as the main evaluation metric, with an average value of approximately 85%. Although this metric is quite common in health-related problems, some authors defend [63] the use of the area under the precision-recall curve as the most appropriate metric for strongly unbalanced bases.

Considering the aspects discussed, we question whether, in the search for a diagnosis, it would not be more appropriate to treat the prediction of new tests as a regression problem, leaving the responsibility of decision-making to health professionals.

Limitations

One of the limitations of this study was how the articles were selected, analyzing only the data from the titles, keywords, and abstracts initially reviewed.

Another limitation was the nonuse of studies whose data source consisted of imaging examinations and clinical history and where the objective was not a prediction.

These criteria greatly reduced the number of selected studies. However, our objective was to analyze only studies that had a main focus on the use of laboratory tests. These requirements are fundamental in building models that can automatically analyze test results without affecting the processes of medical laboratories.

Conclusions

In the search for scientific research that used laboratory tests and ML models to predict new information, 40 studies were found that fit the established criteria. Among these, all (40/40, 100%) sought to predict unknown information, with most (34/40, 85%) focused on the search for a diagnosis.

We have seen a large increase in the use of this methodology in recent years, mainly motivated by the COVID-19 pandemic. Of the 40 works selected from 2010 onward, 27 (68%) focused on SARS-CoV-2, published between 2020 and 2022.

All (40/40, 100%) studies used only laboratory tests, and the complete blood count was the most used. The use of routine examinations is encouraged, mainly as they are more frequently performed and have greater availability. Among the prediction methods, most (39/40, 98%) studies used ML models with supervised learning. These techniques have been spreading and obtaining good results over the years, and binary classification models are still the most used, with XGBoost and DNNs being the models with the best results. These models almost always seek to determine the occurrence or not of a specific event, which has proved to be very useful in the triage of hospitalized patients and in the search for a diagnosis.

In general, all the evaluated studies presented good results, making predictions according to the research objective. Responding to the objectives of this work, we conclude that it is possible to predict specific tests from other laboratory tests, with the complete blood count being the most used in the prediction of new results. The most used method was binary classification with supervised learning.

Thus, the use of laboratory tests and ML techniques represents an innovative potential for the process of medical laboratories, allowing for a more comprehensive analysis of the tests performed, enabling the early discovery of unknown pathologies or errors in the tests performed. This automatic analysis is very advantageous as it is low-cost and does not interfere with the processes already established by medical laboratories.

Conflicts of Interest

None declared.

References

1. Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques. Amsterdam, Netherlands: Elsevier Science; 2011. [doi: 10.1016/C2009-0-61819-5]
2. Sharma A, Mansotra V. Emerging applications of data mining for healthcare management - A critical review. In: Proceedings of the 2014 International Conference on Computing for Sustainable Global Development (INDIACom). 2014 Presented at: 2014 International Conference on Computing for Sustainable Global Development (INDIACom); Mar 5-7, 2014; New Delhi, India. [doi: 10.1109/indiacom.2014.6828163]

3. Hall P, Phan W, Whitson K. Opportunities and Challenges for Machine Learning in Business. Sebastopol: O'Reilly Media; 2016.
4. Castrillón OD, Sarache W, Castaño E. Sistema Bayesiano para la Predicción de la Diabetes. Inf tecnol 2017;28(6):161-168. [doi: 10.4067/s0718-07642017000600017]
5. Luo Y, Szolovits P, Dighe AS, Baron JM. Using machine learning to predict laboratory test results. Am J Clin Pathol 2016 Jun;145(6):778-788. [doi: 10.1093/ajcp/aqw064] [Medline: 27329638]
6. Wong J, Horwitz MM, Zhou L, Toh S. Using machine learning to identify health outcomes from electronic health record data. Curr Epidemiol Rep 2018 Dec;5(4):331-342 [FREE Full text] [doi: 10.1007/s40471-018-0165-9] [Medline: 30555773]
7. Roy SK, Hom J, Mackey L, Shah N, Chen JH. Predicting low information laboratory diagnostic tests. AMIA Jt Summits Transl Sci Proc 2018;2017:217-226 [FREE Full text] [Medline: 29888076]
8. Peek N, Combi C, Marin R, Bellazzi R. Thirty years of artificial intelligence in medicine (AIME) conferences: a review of research themes. Artif Intell Med 2015 Sep;65(1):61-73. [doi: 10.1016/j.artmed.2015.07.003] [Medline: 26265491]
9. Rashidi HH, Tran NK, Betts EV, Howell LP, Green R. Artificial intelligence and machine learning in pathology: the present landscape of supervised methods. Acad Pathol 2019 Sep 03;6:2374289519873088 [FREE Full text] [doi: 10.1177/2374289519873088] [Medline: 31523704]
10. Ahmed Z, Mohamed K, Zeeshan S, Dong X. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. Database (Oxford) 2020 Jan 01;2020:baaa010 [FREE Full text] [doi: 10.1093/database/baaa010] [Medline: 32185396]
11. Houfani D, Slatnia S, Kazar O, Saouli H, Merizig A. Artificial intelligence in healthcare: a review on predicting clinical needs. Int J Healthc Manag 2021 Feb 28;15(3):267-275. [doi: 10.1080/20479700.2021.1886478]
12. Ma C, Wang X, Wu J, Cheng X, Xia L, Xue F, et al. Real-world big-data studies in laboratory medicine: current status, application, and future considerations. Clin Biochem 2020 Oct;84:21-30 [FREE Full text] [doi: 10.1016/j.clinbiochem.2020.06.014] [Medline: 32652094]
13. Gunčar G, Kukar M, Notar M, Brvar M, Černelč P, Notar M, et al. An application of machine learning to haematological diagnosis. Sci Rep 2018 Jan 11;8(1):411 [FREE Full text] [doi: 10.1038/s41598-017-18564-8] [Medline: 29323142]
14. Demirci F, Akan P, Kume T, Sisman AR, Erbayraktar Z, Sevinc S. Artificial neural network approach in laboratory test reporting: learning algorithms. Am J Clin Pathol 2016 Aug 27;146(2):227-237. [doi: 10.1093/ajcp/aqw104] [Medline: 27473741]
15. Rosenbaum M, Baron J. Using machine learning-based multianalyte delta checks to detect wrong blood in tube errors. Am J Clin Pathol 2018 Oct 24;150(6):555-566. [doi: 10.1093/ajcp/aqy085] [Medline: 30169595]
16. Baron JM, Mermel CH, Lewandrowski KB, Dighe AS. Detection of preanalytic laboratory testing errors using a statistically guided protocol. Am J Clin Pathol 2012 Sep 01;138(3):406-413. [doi: 10.1309/ajcpqirib3ct1ejv]
17. Welcome to Medical Subject Headings. NIH U.S. National Library of Medicines. URL: https://www.nlm.nih.gov/mesh/meshhome.html [accessed 2022-02-21]
18. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
19. Study quality assessment tools. NIH National Heart, Lung and Blood Institute. 2021. URL: https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools [accessed 2022-02-21]
20. Wong WC, Cheung CS, Hart GJ. Development of a quality assessment tool for systematic reviews of observational studies (QATSO) of HIV prevalence in men having sex with men and associated risk behaviours. Emerg Themes Epidemiol 2008 Nov 17;5:23 [FREE Full text] [doi: 10.1186/1742-7622-5-23] [Medline: 19014686]
21. Richardson AM, Lidbury BA. Infection status outcome, machine learning method and virus type interact to affect the optimised prediction of hepatitis virus immunoassay results from routine pathology laboratory assays in unbalanced data. BMC Bioinformatics 2013 Jun 25;14:206 [FREE Full text] [doi: 10.1186/1471-2105-14-206] [Medline: 23800244]
22. Waljee AK, Mukherjee A, Singal AG, Zhang Y, Warren J, Balis U, et al. Comparison of imputation methods for missing laboratory data in medicine. BMJ Open 2013 Aug 01;3(8):e002847 [FREE Full text] [doi: 10.1136/bmjopen-2013-002847] [Medline: 23906948]
23. Kinar Y, Kalkstein N, Akiva P, Levin B, Half EE, Goldshtein I, et al. Development and validation of a predictive model for detection of colorectal cancer in primary care by analysis of complete blood counts: a binational retrospective study. J Am Med Inform Assoc 2016 Sep 15;23(5):879-890 [FREE Full text] [doi: 10.1093/jamia/ocv195] [Medline: 26911814]
24. Razavian N, Marcus J, Sontag D. Multi-task prediction of disease onsets from longitudinal laboratory tests. Proc Mach Learn Res 2016;56:73-100 [FREE Full text]
25. Richardson AM, Lidbury BA. Enhancement of hepatitis virus immunoassay outcome predictions in imbalanced routine pathology data by data balancing and feature selection before the application of support vector machines. BMC Med Inform Decis Mak 2017 Aug 14;17(1):121 [FREE Full text] [doi: 10.1186/s12911-017-0522-5] [Medline: 28806936]
26. Birks J, Bankhead C, Holt TA, Fuller A, Patnick J. Evaluation of a prediction model for colorectal cancer: retrospective analysis of 2.5 million patient records. Cancer Med 2017 Oct;6(10):2453-2460 [FREE Full text] [doi: 10.1002/cam4.1183] [Medline: 28941187]

27. Hernandez B, Herrero P, Rawson TM, Moore LS, Evans B, Toumazou C, et al. Supervised learning for infection risk inference using pathology data. BMC Med Inform Decis Mak 2017 Dec 08;17(1):168 [FREE Full text] [doi: 10.1186/s12911-017-0550-1] [Medline: 29216923]
28. Rawson T, Hernandez B, Moore L, Blandy O, Herrero P, Gilchrist M, et al. Supervised machine learning for the prediction of infection on admission to hospital: a prospective observational cohort study. J Antimicrob Chemother 2019 Apr 01;74(4):1108-1115. [doi: 10.1093/jac/dky514] [Medline: 30590545]
29. Aikens RC, Balasubramanian S, Chen JH. A machine learning approach to predicting the stability of inpatient lab test results. AMIA Jt Summits Transl Sci Proc 2019;2019:515-523 [FREE Full text] [Medline: 31259006]
30. Hu L, Yang P, Wang X, Lin F, Chen H, Cao H, et al. Using biochemical indexes to prognose paraquat-poisoned patients: an extreme learning machine-based approach. IEEE Access 2019;7:42148-42155. [doi: 10.1109/access.2019.2907272]
31. Bernardini M, Morettini M, Romeo L, Frontoni E, Burattini L. TyG-er: an ensemble Regression Forest approach for identification of clinical factors related to insulin resistance condition using Electronic Health Records. Comput Biol Med 2019 Sep;112:103358. [doi: 10.1016/j.compbiomed.2019.103358] [Medline: 31336327]
32. Xu S, Hom J, Balasubramanian S, Schroeder LF, Najafi N, Roy S, et al. Prevalence and predictability of low-yield inpatient laboratory diagnostic tests. JAMA Netw Open 2019 Sep 04;2(9):e1910967 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.10967] [Medline: 31509205]
33. Lai H, Huang H, Keshavjee K, Guergachi A, Gao X. Predictive models for diabetes mellitus using machine learning techniques. BMC Endocr Disord 2019 Oct 15;19(1):101 [FREE Full text] [doi: 10.1186/s12902-019-0436-6] [Medline: 31615566]
34. Tamune H, Ukita J, Hamamoto Y, Tanaka H, Narushima K, Yamamoto N. Efficient prediction of vitamin B deficiencies via machine-learning using routine blood test results in patients with intense psychiatric episode. Front Psychiatry 2019 Feb 20;10:1029 [FREE Full text] [doi: 10.3389/fpsyt.2019.01029] [Medline: 32153432]
35. Chicco D, Jurman G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Med Inform Decis Mak 2020 Feb 03;20(1):16 [FREE Full text] [doi: 10.1186/s12911-020-1023-5] [Medline: 32013925]
36. Yu L, Zhang Q, Bernstam EV, Jiang X. Predict or draw blood: an integrated method to reduce lab tests. J Biomed Inform 2020 Apr;104:103394 [FREE Full text] [doi: 10.1016/j.jbi.2020.103394] [Medline: 32113004]
37. Banerjee A, Ray S, Vorselaars B, Kitson J, Mamalakis M, Weeks S, et al. Use of machine learning and artificial intelligence to predict SARS-CoV2 infection from full blood counts in a population. Int Immunopharmacol 2020 Sep;86:106705 [FREE Full text] [doi: 10.1016/j.intimp.2020.106705] [Medline: 32652499]
38. Joshi RP, Pejaver V, Hammarlund NE, Sung H, Lee SK, Furmanchuk A, et al. A predictive tool for identification of SARS-CoV-2 PCR-negative emergency department patients using routine test results. J Clin Virol 2020 Aug;129:104502 [FREE Full text] [doi: 10.1016/j.jcv.2020.104502] [Medline: 32544861]
39. Brinati D, Campagner A, Ferrari D, Locatelli M, Banfi G, Cabitza F. Detection of COVID-19 infection from routine blood exams with machine learning: a feasibility study. J Med Syst 2020 Jul 01;44(8):135 [FREE Full text] [doi: 10.1007/s10916-020-01597-4] [Medline: 32607737]
40. Metsker O, Magoev K, Yakovlev A, Yanishevskiy S, Kopanitsa G, Kovalchuk S, et al. Identification of risk factors for patients with diabetes: diabetic polyneuropathy case study. BMC Med Inform Decis Mak 2020 Aug 24;20(1):201 [FREE Full text] [doi: 10.1186/s12911-020-01215-w] [Medline: 32831065]
41. AlJame M, Imtiaz A, Ahmad I, Mohammed A. Deep forest model for diagnosing COVID-19 from routine blood tests. Sci Rep 2021 Aug 17;11(1):16682 [FREE Full text] [doi: 10.1038/s41598-021-95957-w] [Medline: 34404838]
42. Yadaw A, Li Y, Bose S, Iyengar R, Bunyavanich S, Pandey G. Clinical features of COVID-19 mortality: development and validation of a clinical prediction model. Lancet Digital Health 2020 Oct;2(10):e516-e525 [FREE Full text] [doi: 10.1016/s2589-7500(20)30217-x]
43. Cabitza F, Campagner A, Ferrari D, Di Resta C, Ceriotti D, Sabetta E, et al. Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests. Clin Chem Lab Med 2020 Oct 21;59(2):421-431 [FREE Full text] [doi: 10.1515/cclm-2020-1294] [Medline: 33079698]
44. Schneider JL, Layefsky E, Udaltsova N, Levin TR, Corley DA. Validation of an algorithm to identify patients at risk for colorectal cancer based on laboratory test and demographic data in diverse, community-based population. Clin Gastroenterol Hepatol 2020 Nov;18(12):2734-41.e6. [doi: 10.1016/j.cgh.2020.04.054] [Medline: 32360824]
45. Yang HS, Hou Y, Vasovic LV, Steel PA, Chadburn A, Racine-Brzostek SE, et al. Routine laboratory blood tests predict SARS-CoV-2 infection using machine learning. Clin Chem 2020 Nov 01;66(11):1396-1404 [FREE Full text] [doi: 10.1093/clinchem/hvaa200] [Medline: 32821907]
46. Plante TB, Blau AM, Berg AN, Weinberg AS, Jun IC, Tapson VF, et al. Development and external validation of a machine learning tool to rule out COVID-19 among adults in the emergency department using routine blood tests: a large, multicenter, real-world study. J Med Internet Res 2020 Dec 02;22(12):e24048 [FREE Full text] [doi: 10.2196/24048] [Medline: 33226957]
47. Mooney C, Eogan M, Ní Áinle F, Cleary B, Gallagher JJ, O'Loughlin J, et al. Predicting bacteraemia in maternity patients using full blood count parameters: a supervised machine learning algorithm approach. Int J Lab Hematol 2021 Aug 21;43(4):609-615. [doi: 10.1111/ijlh.13434] [Medline: 33347714]

48. Yu L, Li L, Bernstam E, Jiang X. A deep learning solution to recommend laboratory reduction strategies in ICU. Int J Med Inform 2020 Dec;144:104282. [doi: 10.1016/j.ijmedinf.2020.104282] [Medline: 33010730]
49. Kaftan AN, Hussain MK, Algenabi AA, Naser FH, Enaya MA. Predictive value of C-reactive protein, lactate dehydrogenase, ferritin and D-dimer levels in diagnosing COVID-19 patients: a retrospective study. Acta Inform Med 2021 Mar;29(1):45-50 [FREE Full text] [doi: 10.5455/aim.2021.29.45-50] [Medline: 34012213]
50. Park DJ, Park MW, Lee H, Kim Y, Kim Y, Park YH. Development of machine learning model for diagnostic disease prediction based on laboratory tests. Sci Rep 2021 Apr 07;11(1):7567 [FREE Full text] [doi: 10.1038/s41598-021-87171-5] [Medline: 33828178]
51. Souza AA, Almeida DC, Barcelos TS, Bortoletto RC, Munoz R, Waldman H, et al. Simple hemogram to support the decision-making of COVID-19 diagnosis using clusters analysis with self-organizing maps neural network. Soft comput 2021 May 17:1-12 [FREE Full text] [doi: 10.1007/s00500-021-05810-5] [Medline: 34025211]
52. Kukar M, Gunčar G, Vovko T, Podnar S, Černelč P, Brvar M, et al. COVID-19 diagnosis by routine blood tests using machine learning. Sci Rep 2021 May 24;11(1):10738 [FREE Full text] [doi: 10.1038/s41598-021-90265-9] [Medline: 34031483]
53. Gladding PA, Ayar Z, Smith K, Patel P, Pearce J, Puwakdandawa S, et al. A machine learning PROGRAM to identify COVID-19 and other diseases from hematology data. Future Sci OA 2021 Aug;7(7):FSO733 [FREE Full text] [doi: 10.2144/fsoa-2020-0207] [Medline: 34254032]
54. Rahman T, Al-Ishaq FA, Al-Mohannadi FS, Mubarak RS, Al-Hitmi MH, Islam KR, et al. Mortality prediction utilizing blood biomarkers to predict the severity of COVID-19 using machine learning technique. Diagnostics (Basel) 2021 Aug 31;11(9):1582 [FREE Full text] [doi: 10.3390/diagnostics11091582] [Medline: 34573923]
55. Myari A, Papapetrou E, Tsaousi C. Diagnostic value of white blood cell parameters for COVID-19: is there a role for HFLC and IG? Int J Lab Hematol 2022 Feb 08;44(1):104-111 [FREE Full text] [doi: 10.1111/ijlh.13728] [Medline: 34623763]
56. Campagner A, Carobene A, Cabitza F. External validation of Machine Learning models for COVID-19 detection based on Complete Blood Count. Health Inf Sci Syst 2021 Dec;9(1):37 [FREE Full text] [doi: 10.1007/s13755-021-00167-3] [Medline: 34721844]
57. Babaei Rikan S, Sorayaie Azar A, Ghafari A, Bagherzadeh Mohasefi J, Pirnejad H. COVID-19 diagnosis from routine blood tests using artificial intelligence techniques. Biomed Signal Process Control 2021 Nov 01:103263 [FREE Full text] [doi: 10.1016/j.bspc.2021.103263] [Medline: 34745318]
58. Hossain ME, Khan A, Moni MA, Uddin S. Use of electronic health data for disease prediction: a comprehensive literature review. IEEE/ACM Trans Comput Biol Bioinf 2021 Mar 1;18(2):745-758. [doi: 10.1109/tcbb.2019.2937862]
59. Wu YT, Zhang CJ, Mol BW, Kawai A, Li C, Chen L, et al. Early prediction of gestational diabetes mellitus in the Chinese population via advanced machine learning. J Clin Endocrinol Metab 2021 Mar 08;106(3):e1191-e1205 [FREE Full text] [doi: 10.1210/clinem/dgaa899] [Medline: 33351102]
60. Hische M, Luis-Dominguez O, Pfeiffer AF, Schwarz PE, Selbig J, Spranger J. Decision trees as a simple-to-use and reliable tool to identify individuals with impaired glucose metabolism or type 2 diabetes mellitus. Eur J Endocrinol 2010 Oct;163(4):565-571. [doi: 10.1530/EJE-10-0649] [Medline: 20693184]
61. Ravaut M, Harish V, Sadeghi H, Leung KK, Volkovs M, Kornas K, et al. Development and validation of a machine learning model using administrative health data to predict onset of type 2 diabetes. JAMA Netw Open 2021 May 03;4(5):e2111315 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.11315] [Medline: 34032855]
62. Le TM, Vo TM, Pham TN, Dao SV. A novel wrapper–based feature selection for early diabetes prediction enhanced with a metaheuristic. IEEE Access 2021;9:7869-7884. [doi: 10.1109/access.2020.3047942]
63. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015 Mar 4;10(3):e0118432 [FREE Full text] [doi: 10.1371/journal.pone.0118432] [Medline: 25738806]

Abbreviations

AUC: area under the curve
DNN: deep neural network
ML: machine learning
NIH: National Institutes of Health
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by A Mavragani; submitted 22.06.22; peer-reviewed by J Nievola, P Dunn; comments to author 19.07.22; revised version received 28.08.22; accepted 31.10.22; published 23.12.22.

Please cite as:
Cardozo G, Tirloni SF, Pereira Moro AR, Marques JLB
Use of Artificial Intelligence in the Search for New Information Through Routine Laboratory Tests: Systematic Review
JMIR Bioinform Biotech 2022;3(1):e40473
URL: https://bioinform.jmir.org/2022/1/e40473
doi:10.2196/40473
PMID:36644762

Original Paper

The Easy-to-Use SARS-CoV-2 Assembler for Genome Sequencing: Development Study

Martina Rueca[1], MSc; Emanuela Giombini[1], PhD; Francesco Messina[2], PhD; Barbara Bartolini[2], PhD; Antonino Di Caro[2,3], MSc; Maria Rosaria Capobianchi[1,3], MSc; Cesare EM Gruber[1], PhD

[1]Laboratory of Virology and Biosafety Laboratories, National Institute for Infectious Diseases "Lazzaro Spallanzani", Istituto di Ricovero e Cura a Carattere Scientifico, Rome, Italy

[2]Laboratory of Microbiology and Biological Bank, National Institute for Infectious Diseases "Lazzaro Spallanzani", Istituto di Ricovero e Cura a Carattere Scientifico, Rome, Italy

[3]UniCamillus - Saint Camillus International University of Health Sciences, Roma, Italy

Corresponding Author:
Francesco Messina, PhD
Laboratory of Microbiology and Biological Bank
National Institute for Infectious Diseases "Lazzaro Spallanzani"
Istituto di Ricovero e Cura a Carattere Scientifico
Via Portuense 292
Rome, 00149
Italy
Phone: 39 0655170668
Email: francesco.messina@inmi.it

Abstract

Background: Early sequencing and quick analysis of the SARS-CoV-2 genome have contributed to the understanding of the dynamics of COVID-19 epidemics and in designing countermeasures at a global level.

Objective: Amplicon-based next-generation sequencing (NGS) methods are widely used to sequence the SARS-CoV-2 genome and to identify novel variants that are emerging in rapid succession as well as harboring multiple deletions and amino acid–changing mutations.

Methods: To facilitate the analysis of NGS sequencing data obtained from amplicon-based sequencing methods, here, we propose an easy-to-use SARS-CoV-2 genome assembler: the Easy-to-use SARS-CoV-2 Assembler (ESCA) pipeline.

Results: Our results have shown that ESCA could perform high-quality genome assembly from Ion Torrent and Illumina raw data and help the user in easily correct low-coverage regions. Moreover, ESCA includes the possibility of comparing assembled genomes of multisample runs through an easy table format.

Conclusions: In conclusion, ESCA automatically furnished a variant table output file, fundamental to rapidly recognizing variants of interest. Our pipeline could be a useful method for obtaining a complete, rapid, and accurate analysis even with minimal knowledge in bioinformatics.

(JMIR Bioinform Biotech 2022;3(1):e31536) doi:10.2196/31536

KEYWORDS

SARS-CoV-2 genome; bioinformatics tool; NGS data analysis; COVID-19; genome; health informatics; bioinformatic; digital tools; algorithms

Introduction

Next-generation sequencing (NGS) has reached a pivotal role in the field of emerging infectious diseases by enhancing the development capacity of new diagnostic methods, vaccines, and drugs [1,2]. Moreover, a key role has been recognized for sequence data production and sharing in outbreak response and management [3-5]. In the current COVID-19 epidemic, more than 6 million full genome sequences of SARS-CoV-2 have been deposited in publicly accessible databases in the arc of 1 year (ie, GISAID) [6,7]. SARS-CoV-2 genome surveillance on a global scale is permitting real-time analysis of the outbreak, with a direct impact on the public health response. This contribution includes the tracing of SARS-CoV-2 spread over

time and space, evidence of emerging variants that may affect pathogenicity, transmission capacity, diagnostic methods, therapeutics, or vaccines [8-11]. Recently divergent SARS-CoV-2 variants are emerging in rapid succession, harboring multiple deletions and amino acid mutations. Some mutations occur in the receptor-binding domain of the spike protein and are associated with an increase of angiotensin-converting enzyme 2 (ACE2) affinity as well as a potential reduction of polyclonal human plasma antibody efficacy [12,13]. The growing contribution of sequence information to public health is driving global investment in sequencing facilities and scientific programs [14,15]. The falling cost of generating genomic NGS data provides new chances for sequencing capacity expansion; however, many laboratories have low sequencing capacity and even a lack of expertise for data elaboration.

While sequencing runs can be performed without consolidated experience in the infectious disease field, virus genomic sequence assembly is often a demanding task. Translating SARS-CoV-2 raw read data into reliable and informative results is complex and requires solid bioinformatics knowledge, particularly for low-coverage samples. Some steps can lead to incorrect variant calling and produce erroneous assembled sequences.

Supervision of the sequence assembly to avoid inconsistent or misleading assignment of a virus to a taxonomic lineage or clade [9,10] as well as evaluation of low-coverage samples to prevent loss of epidemiological information are mandatory.

Many tools have been developed to support whole genome sequence reconstruction, starting with reads produced by different NGS platforms. However, most tools have been designed for genome assembly of other viruses and often are able to elaborate only a specific type of data. Some of these tools, for example, have implemented the assembly method for one specific platform (ie, Loretta for PacBio data) [16] or for a specific sequencing approach (ie, UNAGI for Nanopore and Illumina data) [17]. Some sequencing platform manufacturers have proposed pipelines for SARS-CoV-2 genome reconstruction that have been designed to obtain the most accurate sequence from one specific technological output. For example, Illumina developed the DRAGEN tool for SARS-CoV-2 genome analysis, a commercial tool that is temporarily free and available online, while Ion Torrent suggests the iterative refinement meta-assembler (IRMA) for SARS-CoV-2 data analyses, an open-source program developed by the Centers for Disease Control and Prevention (CDC) [18].

We propose the Easy-to-use SARS-CoV-2 Assembler (ESCA) pipeline: a novel reference-based genome assembly pipeline specifically designed for SARS-CoV-2 data analysis. This pipeline was created to support laboratories with limited experience in bioinformatics for SARS-CoV-2 analysis. ESCA can be easily installed and runs in most Linux environments.

Methods

Overview

The ESCA pipeline is a reference-based assembly algorithm written for Linux environments and requires only raw reads as input files, without any other information. Two versions of the software are available: one for Illumina paired-end reads in the “fastq.gz” file format and the other for Ion Torrent reads in the “ubam” file format.

The software is designed to process several samples in a single run. All reads (paired or unpaired) must be copied into the same working directory, and then, the program is launched through the command line by typing “StartEasyTorrent” for IonTorrent input or “StartEasyIllumina” for Illumina input. The pipeline than performs all the other passages automatically, as described in the following paragraphs.

The program processes all input reads, dividing them into different samples using file names as identifiers. Illumina paired-end reads are expected to be divided into 2 files that contain “R1” or “R2” to distinguish forward reads from reverse reads.

Sample preprocessing is performed by filtering out all reads with a mean Phred quality score lower than 20 and that are less than 30 nucleotides long.

Filtered reads are mapped on the SARS-CoV-2 reference genome Wuhan-Hu-1 (GenBank Accession Number NC_045512.2) with bwa-mem software [15]; all reads that do not map on the reference genome are then discarded.

Genome coverage is then analyzed: The read-mapping file is converted into “sorted-bam” and “mpileup” files using samtools software [19], and these data are translated into a detailed coverage table that reports the count of nucleotides observed at each position.

The consensus sequence is then reconstructed on the basis of 3 parameters: (1) frequency of nucleotides observed at each position, (2) nucleotide coverage, (3) reference genome sequence.

Briefly, sample parameters for consensus sequence reconstruction are designed to call the nucleotide observed with >50% frequency and with a coverage of >50 reads, but the minimum coverage is reduced at >10 reads if the most frequent nucleotide observed is identical to the nucleotide observed in the reference genome.

For all positions where these parameters are not satisfied, the ESCA pipeline is designed to call “N” to indicate a low coverage position or an intrasample nucleotide variant.

After whole genome reconstruction of all samples, the consensus sequences are aligned with the Wuhan-Hu-1 reference genome using MAFFT software [20], and a mutation table is generated, reporting nucleotide mutations of all the genomes assembled.

Illumina Data

To test the efficiency of the ESCA pipeline, 228 SARS-CoV-2–positive samples were sequenced with Illumina

platforms using the Ion AmpliSeq SARS-CoV-2 Research Panel following the manufacturer's instructions (ThermoFisher, Waltham, MA). For Illumina samples, whole SARS-CoV-2 genome sequences were assembled using both ESCA and DRAGEN RNA Pathogen Detection v.3.5.15 (BaseSpace) with default parameters.

Ion Torrent Data

A resequencing assay on Ion Torrent platforms was carried out for the same 228 SARS-CoV-2–positive samples using the Ion AmpliSeq SARS-CoV-2 Research Panel following the manufacturer's instructions (ThermoFisher).

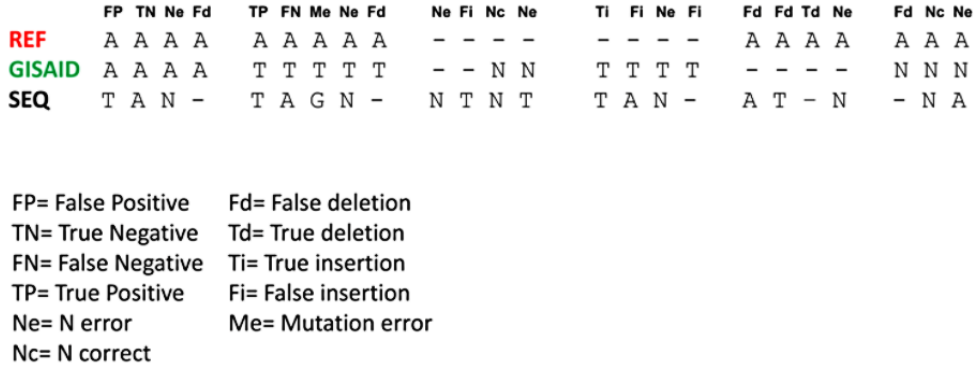
For Ion Torrent samples, whole SARS-CoV-2 genome sequences were assembled with ESCA and IRMA software [18] using the setting parameters indicated by ThermoFisher, in order to test the consistency of ESCA and IRMA outputs.

Performance Test

The respective results were compared, aligning the sequences obtained using the 2 methods with the reference sequence Wuhan-Hu-1 (NCBI Acc. Numb. NC_045512.2), and the corrected sequence was submitted to GISAID, using MAFFT [20]. Then, each discordant position was evaluated following the classification reported in Figure 1. In particular, we evaluated true positives (TP; mutations correctly classified as real); false negatives (FN; mutations correctly classified as unreal); false positives (FP; mutations incorrectly classified as real); true negatives (TN; mutations correctly classified as unreal); corrected TN (positions unknown correctly classified as N); and TN error (positions unknown, incorrectly classified as N).

To test the performances with respect to mean coverage, linear regression correlation analysis was carried out for mean coverage and specific measures of accuracy.

Figure 1. Classification scheme for genome assemblers, in which assembled genome sequences (SEQ) were compared with the corresponding submitted sequences (on GISAID) and with reference genome sequence "Wuhan-Hu-1" (REF). Nucleotide threesomes were classified using the following 11 categories: false deletion (Fd), false insertion (Fi), false negative (FN), false positive (FP), mutation error (Me), N correct (Nc), N error (Ne), true deletion (Td), true insertion (Ti), true negative (TN), true positive (TP).



Results

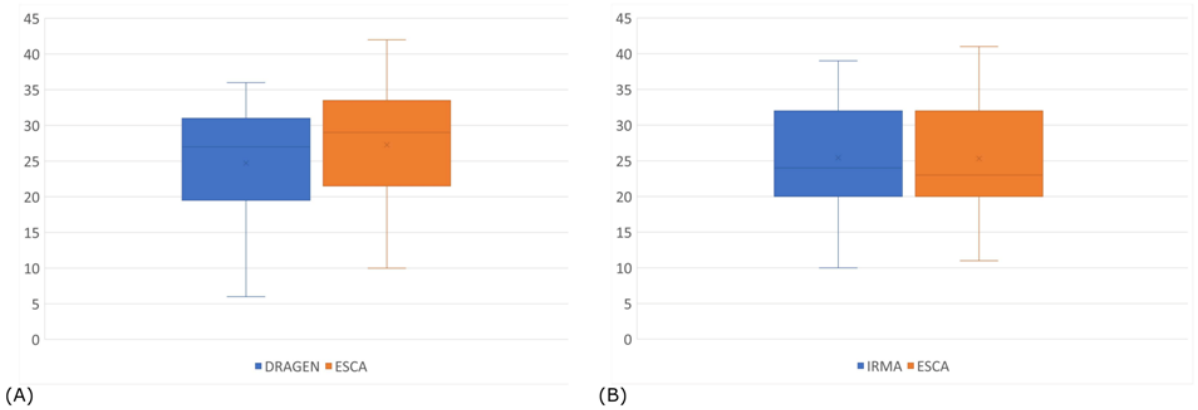
In the computational evaluation, ESCA software was compared with the most often used assemblers for SARS-CoV-2 genome analysis on 228 SARS-CoV-2–positive samples.

Sequencing was performed on Illumina MiSeq for 65 libraries, obtaining a median of 1.50 x 106 paired-end reads per sample (range: 0.02 x 106 to 4.56 x 106), and on Ion Gene Studio S5 Sequencer for 163 libraries, obtaining a median of 0.61 x 106 single-end reads per sample (range: 0.02 x 106 to 3.02 x 106). Using the ESCA reconstruction, the coverage point by point was calculated, and we observed that, in the Illumina sample, the point coverage was not uniform, although the mean coverage was quite high in all samples (average 3508X; range: 70-10,733). This could introduce error in genome reconstruction using some software. In this context, ESCA could reduce the error in regions with low coverage. In parallel, mean coverage obtained with Ion Torrent was 4966X (range: 94-19,917), but a higher uniformity was observed. The comparison of the coverage distribution is shown in Figure 2.

To evaluate the ESCA and DRAGEN/IRMA results, assembled genomes, the reference Wuhan-Hu-1, and the corrected genome of GISAID (Accession IDs available in Multimedia Appendix 1) were aligned with MAFFT [20].

At each position along the SARS-CoV-2 genome, the 24 available nucleotide combinations were classified in 11 mutation categories (Figure 1). For all sequences, the number of occurrences of mutation categories for each assembly software was then evaluated.

Figure 2. Comparison of true positive mutations between our Easy-to-use SARS-CoV-2 Assembler (ESCA) and the (A) Illumina DRAGEN tool and (B) iterative refinement meta-assembler (IRMA) recommended by Ion Torrent.



Illumina Data

The comparison of ESCA with DRAGEN showed that, as expected, the mean number of mutations in genomes was very low (in the mean 28 position) and ESCA could correctly identify a mean 27 of 28 mutations (Figure 2A). Moreover, no FN positions were identified by ESCA. This is due to the pipeline design that reduces the error of introducing N where the coverage is not sufficient. The DRAGEN genome, instead, showed a mean of 25 of 28 TP and 3 FN positions. The absence of a mutation in specific positions could be essential to assigning the lineage, and the presence of FNs could modify the identification of the variants.

On the other hand, both ESCA and DRAGEN did not introduce FN, identifying 29,308 and 28,027 TN positions, respectively.

These results show an accuracy of 100% for ESCA and 99.99% for DRAGEN. Moreover, the sensitivity of ESCA compared with DRAGEN was 96.43% for ESCA and 89.29% for DRAGEN, and the specificity with both methods was 100%.

Ion Torrent Data

Parallel to the previous comparison, ESCA compared with IRMA showed that both methods identified a mean 25 of 26 TP positions (Figure 2B) but did not induce FN. However, IRMA introduced a certain number of errors. In fact, the FP was 20 for IRMA, compared with 0 for ESCA. Once again, the introduction of mutations could induce error in the lineage assignation.

The accuracy of IRMA was calculated to be 99.93%, while it was 100% for ESCA.

Moreover, although the sensitivity was identical with the 2 methods (96.15%), the specificity was 99.93% for IRMA and 100% for ESCA.

Performance Test

To evaluate the performance of each of the methods, linear regression correlation analysis was carried out with respect to mean coverage (Multimedia Appendix 2).

For IonTorrent single-end sequencing data, a significant positive correlation was found comparing coverage and TN for both IRMA and ESCA ($r>0.15$, $P<.05$), while for Illumina pair-end sequencing data, such a correlation was found only for DRAGEN ($r>0.40$, $P<.05$). This difference could be caused by a different error rate for the 2 sequencing techniques. These data suggest that all assembly methods are comparable in the case of high coverage samples, while ESCA seems to perform better for low coverage data.

Discussion

Principal Findings

The importance of rapidly obtaining and sharing high-quality whole genomes of SARS-CoV-2 is increasing with the emerging variant strains [14]. For this reason, the use of NGS custom amplicon panels can be a rapid and performant method for identifying viral variants. However, a lack of bioinformatic skills could be a problem in handling NGS raw data. Our pipeline ESCA provides help to laboratories with low bioinformatic capacity using a single command. Both of the more common methods for the analysis of Ion Torrent and Illumina data (IRMA and DRAGEN, respectively) have shown a certain amount of error that could induce false identification in variant assignation. On the contrary, the SARS-CoV-2 genome obtained by ESCA shows a reduced number of false insertions and false mutations and a higher number of real mutations.

Limitations

This pipeline should be tested on a larger number of sequences and with other sequencing technologies.

Conclusions

ESCA automatically produces a variant table output file, fundamental for rapidly recognizing variants of interest.

These results show how ESCA could be a useful method for obtaining a rapid, complete, and correct analysis even with minimal skill in bioinformatics.

Acknowledgments

We gratefully acknowledge the contributors of the genome sequences of the newly emerging coronavirus (ie, the originating laboratories) for sharing sequences and other metadata through the GISAID Initiative, on which this research is based. We also acknowledge Ornella Butera, Francesco Santini, and Giulia Bonfiglio for their contribution to sample preparation. National Institute for Infectious Diseases Lazzaro Spallanzani IRCCS received financial support from the Italian Ministry of Health grants “Ricerca Corrente” (Progetto 1-2763705) and “5 PER MILLE 2020” (iSNV study for early detection and risk analysis of SARS-CoV-2 mutations with impact on clinical management of COVID-19) research funds.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Comparison of Easy-to-use SARS-CoV-2 Assembler (ESCA) and (first tab: Ion Torrent) iterative refinement meta-assembler (IRMA) and (second tab: Illumina) DRAGEN results for every assembled SARS-CoV-2 genome. DELerror: deletion unknown incorrectly identified; errorMut: number of incorrected mutations described; FN: false negative; FP: false positive; INScorr: insertion unknown correctly identified; INSerror: insertion unknown incorrectly identified; NCorr: positions unknown correctly classified as N; Nerror: position unknown incorrectly classified as N; TN: true negative; TP: true positive.
[XLSX File (Microsoft Excel File), 36 KB - bioinform_v3i1e31536_app1.xlsx]

Multimedia Appendix 2

Pairwise linear regression correlation analysis (shown with the *P* value) for every parameter: DELerror: deletion unknown incorrectly identified; FN: false negative; FP: false positive; INScorr: insertion unknown correctly identified; INSerror: insertion unknown incorrectly identified; NCorr: positions unknown correctly classified as N; Nerror: position unknown incorrectly classified as N; TN: true negative; TP: true positive.
[XLSX File (Microsoft Excel File), 41 KB - bioinform_v3i1e31536_app2.xlsx]

References

1. World Health Organization (WHO). Genomic sequencing of SARS-CoV-2: a guide to implementation for maximum impact on public health. World Health Organization. 2021 Jan 08. URL: https://www.who.int/publications/i/item/9789240018440 [accessed 2022-02-22]
2. Greaney AJ, Loes AN, Crawford KH, Starr TN, Malone KD, Chu HY, et al. Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. Cell Host Microbe 2021 Mar 10;29(3):463-476.e6 [FREE Full text] [doi: 10.1016/j.chom.2021.02.003] [Medline: 33592168]
3. Smith GJD, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, Pybus OG, et al. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature 2009 Jun 25;459(7250):1122-1125. [doi: 10.1038/nature08182] [Medline: 19516283]
4. Revez J, Espinosa L, Albiger B, Leitmeyer KC, Struelens MJ, ECDC National Microbiology Focal Points and Experts Group. Survey on the use of whole-genome sequencing for infectious diseases surveillance: rapid expansion of European National Capacities, 2015-2016. Front Public Health 2017;5:347 [FREE Full text] [doi: 10.3389/fpubh.2017.00347] [Medline: 29326921]
5. Nadon C, Van Walle I, Gerner-Smidt P, Campos J, Chinen I, Concepcion-Acevedo J, FWD-NEXT Expert Panel. PulseNet International: Vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. Euro Surveill 2017 Jun 08;22(23):1 [FREE Full text] [doi: 10.2807/1560-7917.ES.2017.22.23.30544] [Medline: 28662764]
6. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob Chall 2017 Jan 10;1(1):33-46 [FREE Full text] [doi: 10.1002/gch2.1018] [Medline: 31565258]
7. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2018 Jan 04;46(D1):D8-D13 [FREE Full text] [doi: 10.1093/nar/gkx1095] [Medline: 29140470]
8. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 2020 May;20(5):533-534 [FREE Full text] [doi: 10.1016/S1473-3099(20)30120-1] [Medline: 32087114]
9. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 2018 Dec 01;34(23):4121-4123 [FREE Full text] [doi: 10.1093/bioinformatics/bty407] [Medline: 29790939]
10. Rambaut A, Holmes EC, O'Toole A, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol 2020 Nov;5(11):1403-1407 [FREE Full text] [doi: 10.1038/s41564-020-0770-5] [Medline: 32669681]
11. Outbreak.info. URL: https://outbreak.info/ [accessed 2022-02-22]

12. Cameroni E, Bowen JE, Rosen LE, Saliba C, Zepeda SK, Culap K, et al. Broadly neutralizing antibodies overcome SARS-CoV-2 Omicron antigenic shift. Nature 2022 Feb 10;602(7898):664-670 [FREE Full text] [doi: 10.1038/s41586-021-04386-2] [Medline: 35016195]
13. Focosi D, Maggi F. Neutralising antibody escape of SARS-CoV-2 spike protein: Risk assessment for antibody-based Covid-19 therapeutics and vaccines. Rev Med Virol 2021 Nov;31(6):e2231 [FREE Full text] [doi: 10.1002/rmv.2231] [Medline: 33724631]
14. SARS-CoV-2 genomic sequencing for public health goals: Interim guidance, 8 January 2021. World Health Organization. 2021 Jan 08. URL: https://www.who.int/publications/i/item/WHO-2019-nCoV-genomic_sequencing-2021.1 [accessed 2022-02-22]
15. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009 Jul 15;25(14):1754-1760 [FREE Full text] [doi: 10.1093/bioinformatics/btp324] [Medline: 19451168]
16. Al Qaffas A, Nichols J, Davison AJ, Ourahmane A, Hertel L, McVoy MA, et al. LoReTTA, a user-friendly tool for assembling viral genomes from PacBio sequence data. Virus Evol 2021 Jan;7(1):veab042 [FREE Full text] [doi: 10.1093/ve/veab042] [Medline: 33996146]
17. Al Kadi M, Jung N, Ito S, Kameoka S, Hishida T, Motooka D, et al. UNAGI: an automated pipeline for nanopore full-length cDNA sequencing uncovers novel transcripts and isoforms in yeast. Funct Integr Genomics 2020 Jul 18;20(4):523-536 [FREE Full text] [doi: 10.1007/s10142-020-00732-1] [Medline: 31955296]
18. Shepard SS, Meno S, Bahl J, Wilson MM, Barnes J, Neuhaus E. Viral deep sequencing needs an adaptive approach: IRMA, the iterative refinement meta-assembler. BMC Genomics 2016 Sep 05;17:708 [FREE Full text] [doi: 10.1186/s12864-016-3030-6] [Medline: 27595578]
19. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009 Aug 15;25(16):2078-2079 [FREE Full text] [doi: 10.1093/bioinformatics/btp352] [Medline: 19505943]
20. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 2002 Jul 15;30(14):3059-3066 [FREE Full text] [doi: 10.1093/nar/gkf436] [Medline: 12136088]

Abbreviations

ACE2: angiotensin-converting enzyme 2
CDC: Centers for Disease Control and Prevention
ESCA: Easy-to-use SARS-CoV-2 Assembler
FN: false negative
FP: false positive
IRMA: iterative refinement meta-assembler
NGS: next-generation sequencing
TN: true negative
TP: true positive

Edited by A Mavragani; submitted 24.06.21; peer-reviewed by S Tausch, Y Miao; comments to author 30.07.21; revised version received 02.11.21; accepted 05.02.22; published 14.03.22.

Please cite as:
Rueca M, Giombini E, Messina F, Bartolini B, Di Caro A, Capobianchi MR, Gruber CEM
The Easy-to-Use SARS-CoV-2 Assembler for Genome Sequencing: Development Study
JMIR Bioinform Biotech 2022;3(1):e31536
URL: https://bioinform.jmir.org/2022/1/e31536
doi:10.2196/31536
PMID:35309411

Original Paper

Diagnosis of a Single-Nucleotide Variant in Whole-Exome Sequencing Data for Patients With Inherited Diseases: Machine Learning Study Using Artificial Intelligence Variant Prioritization

Yu-Shan Huang[1], MSc; Ching Hsu[2], MSc; Yu-Chang Chune[1], MSc; I-Cheng Liao[1], MSc; Hsin Wang[2], MSc; Yi-Lin Lin[3], MSc; Wuh-Liang Hwu[4], MD, PhD; Ni-Chung Lee[3], MD, PhD; Feipei Lai[1,2], PhD

[1]Department of Computer Science and Information Engineering, National Taiwan University, Taipei City, Taiwan

[2]Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei City, Taiwan

[3]Department of Medical Genetics, National Taiwan University Hospital, Taipei City, Taiwan

[4]Department of Pediatrics, National Taiwan University Hospital, Taipei City, Taiwan

Corresponding Author:
Feipei Lai, PhD
Graduate Institute of Biomedical Electronics and Bioinformatics
National Taiwan University
Number 1, Roosevelt Road, Section 4
Taipei City, 106319
Taiwan
Phone: 886 2 33664924
Fax: 886 2 23628167
Email: flai@ntu.edu.tw

Abstract

Background: In recent years, thanks to the rapid development of next-generation sequencing (NGS) technology, an entire human genome can be sequenced in a short period. As a result, NGS technology is now being widely introduced into clinical diagnosis practice, especially for diagnosis of hereditary disorders. Although the exome data of single-nucleotide variant (SNV) can be generated using these approaches, processing the DNA sequence data of a patient requires multiple tools and complex bioinformatics pipelines.

Objective: This study aims to assist physicians to automatically interpret the genetic variation information generated by NGS in a short period. To determine the true causal variants of a patient with genetic disease, currently, physicians often need to view numerous features on every variant manually and search for literature in different databases to understand the effect of genetic variation.

Methods: We constructed a machine learning model for predicting disease-causing variants in exome data. We collected sequencing data from whole-exome sequencing (WES) and gene panel as training set, and then integrated variant annotations from multiple genetic databases for model training. The model built ranked SNVs and output the most possible disease-causing candidates. For model testing, we collected WES data from 108 patients with rare genetic disorders in National Taiwan University Hospital. We applied sequencing data and phenotypic information automatically extracted by a keyword extraction tool from patient's electronic medical records into our machine learning model.

Results: We succeeded in locating 92.5% (124/134) of the causative variant in the top 10 ranking list among an average of 741 candidate variants per person after filtering. AI Variant Prioritizer was able to assign the target gene to the top rank for around 61.1% (66/108) of the patients, followed by Variant Prioritizer, which assigned it for 44.4% (48/108) of the patients. The cumulative rank result revealed that our AI Variant Prioritizer has the highest accuracy at ranks 1, 5, 10, and 20. It also shows that AI Variant Prioritizer presents better performance than other tools. After adopting the Human Phenotype Ontology (HPO) terms by looking up the databases, the top 10 ranking list can be increased to 93.5% (101/108).

Conclusions: We successfully applied sequencing data from WES and free-text phenotypic information of patient's disease automatically extracted by the keyword extraction tool for model training and testing. By interpreting our model, we identified which features of variants are important. Besides, we achieved a satisfactory result on finding the target variant in our testing data set. After adopting the HPO terms by looking up the databases, the top 10 ranking list can be increased to 93.5% (101/108).

The performance of the model is similar to that of manual analysis, and it has been used to help National Taiwan University Hospital with a genetic diagnosis.

(JMIR Bioinform Biotech 2022;3(1):e37701) doi:10.2196/37701

KEYWORDS

next-generation sequencing; genetic variation analysis; machine learning; artificial intelligence; whole-exome sequencing

Introduction

Background

Modern next-genome sequencing (NGS) technology makes rapid human genome sequencing within a day possible [1,2]. Because of its speed and low cost in comparison with the traditional Sanger sequencing method [3], NGS is being rapidly introduced into clinical and public health laboratory practice, especially for the diagnosis of hereditary disorders.

Although NGS has extremely high throughput and could generate huge amounts of genomic data in a short time, interpreting these data and finding the disease-causing candidates among thousands of variants remain a challenge. To determine the true causal variants of a patient with genetic disease, physicians often need to view numerous features on every variant manually and search for literature in different databases to understand the effect of a genetic variation. Another challenge is in finding the genetic variants that have a strong correlation with patient's phenotype. Physicians often select useful keywords from patient's electronic medical records (EMRs) manually to search for articles in several genetic databases such as Online Mendelian Inheritance in Man (OMIM) [4] and GeneReviews [5] to decide whether a variant is correlated with a genetic disease. It is thus a burden for physicians to go through these laborious and time-consuming processes case-by-case, especially when the number of inherited disease–associated germline mutations published per year has increased exponentially in the last decade [6].

Nowadays, many studies use machine learning methods to solve numerous problems in genomics and genetics. The field of machine learning promises to enable computers to assist humans in making sense of large, complex data sets. After variant annotation, there is a variant list with hundreds of columns that humans are not capable of interpreting one-by-one. As machine learning significantly surpasses human-level performance, especially with structured data, we consider using a machine learning method to analyze variants from NGS and find the target gene.

To address these problems, it is important and necessary to have a high-performance method to filter candidate variants from NGS results and immediately find target variants related to a patient's disease. Recently, many tools such as Exomiser [7], DeepPVP [8], Xrare [9], VarSight [10], Phenolyzer [11], Fabric GEM [12], MOON [2], CADD [13], and MetaSVM [14] have been developed to identify potentially causative variants that are relevant to patient's phenotype in rare disease diagnosis. Exomiser integrates information including calculated gene-specific phenotype score, variant allele frequency (Multimedia Appendix 1), and predicted pathogenicity of several alleles to prioritize disease-causative variants/interactions. Fabric GEM utilizes Bayes factor to prioritize variants with the support of a gene-phenotype score calculated by Phevor [15] and variant prioritization result of several tools including ANNOVAR, VAAST, and Phen-Gen. MOON integrates the result of annotation of several variants and prioritization tools to achieve variant prioritization using several kinds of machine learning models. Gene-phenotype scores calculated by Phevor using Human Phenotype Ontology (HPO) terms extracted from electronic health records (EHRs) of patients are also considered by MOON. CADD utilizes logistic regression to integrate information including context of surrounding sequence, biological constraints, epigenetic measurements, and result of several variant annotation tools to build a predictive model for variant deleteriousness. MetaSVM [14] gathers result of 9 deleteriousness prediction scores including PolyPhen-2 [16], SIFT [17], MutationTaster [18] to build a support vector machine (SVM) deleteriousness predictive model. Although these tools adopt different approaches, including logistic regression and deep neural networks, to prioritize variants, most can only recognize the phenotypes defined in the HPO term [19]. In this work, we developed the AI Variant Prioritizer module based on a machine learning approach that can output the rank of single-nucleotide variants (SNVs) and small insertions/deletions (indels) from whole-exome sequencing (WES) data with the input of free-text phenotypic description or EHR.

In this research, we aimed to implement a website, AI Variant Prioritizer, that uses data from NGS bioinformatics pipelines with machine learning to make a prediction about the most possible disease-causing variants among SNVs and patient's phenotype. The data generated from NGS pipelines are all structured with annotations from several tools including ANNOVAR, Nirvana, Variant Effect Predictor (VEP), and InterVar and additional information from multiple databases queried by MViewer (Mutation Viewer) [20]. To simplify the interpretation process, we integrate the keyword extraction tool to generate the phenotype from EMRs automatically. Our system takes candidate variants filtered by MViewer and patient's EMRs as its input and outputs a list of SNVs with rank and probability of being disease causing. Instead of checking every variant manually, this system can assist researchers and physicians in focusing on those with higher disease-causing probability and save a lot of time. Moreover, we implement a web application programming interface (API) for our system so that the ranking function could be integrated into MViewer. Thus, physicians are able to interpret the results of genetic variation with a single application instead of adopting numerous services.

Data Description

In our research, we focus on patients who have been diagnosed with rare Mendelian diseases. Our data are collected mainly from the rapid exome project of Department of Medical Genetics, National Taiwan University Hospital (NTUH). To build the model with more data, we also applied for several WES data that are deposited in the dbGaP database (project ID 20911). The data we use are the dbGaP accession phs000711.v5.p1 by Baylor Hopkins Center for Mendelian Genomics.

The conditions under which we collect patients' sequencing data to meet the requirements of this research are as follows:

- Patients who were diagnosed with genetic disorders.
- Patients who received WES or targeted panel sequencing and diagnosed with at least one disease-causing variant.
- Patients whose phenotype information is available.

Our data from NTUH include patient demographics, variant call format (VCF) file output by the NGS bioinformatics pipeline, and phenotype information from electrical medical records. Data from dbGaP also include patient demographics, VCF file, and clinical conditions. All data are deidentified and will not invade patients' privacy. We include sex in patient demographic information as a feature in our model because some human genetic disorders are sex linked. Sex-linked diseases are caused by mutations in genes on X or Y chromosomes and passed down through families.

Variant Call Format File

As the end product of the NGS bioinformatics pipeline, the VCF is a generic format for storing DNA polymorphism data such as SNPs, insertions, deletions, and structural variants. The format was developed for the 1000 Genomes Project and has also been widely adopted by other projects. Every VCF file consists of 2 two parts: header section and data section. The header contains metadata about the tags and annotations in the data part. It can be also used to provide information related to the history of the data and file. The last line in the header contains the column headings for the data part. The data section is tab separated into 9 columns and reports a mutation for each row. Columns include CHROM, POS, ID, REF, ALT, QUAL, FILTER, INFO, and FORMAT.

Phenotype Information

For the data from NTUH, we extract patient's phenotypic information from clinicians' history summary. It mainly records a brief summary of patient's illness, clinical diagnosis, and the reason(s) why each patient was admitted. We also collect the phenotype keywords provided by doctors based on the symptom of each patient for model validation. For the data from dbGaP, because EHRs are not available, we will use the clinical condition of the patient instead. For the clinical condition that can be found in OMIM databases, we will extract the corresponding description of phenotypes as the phenotypic information to be used in our research.

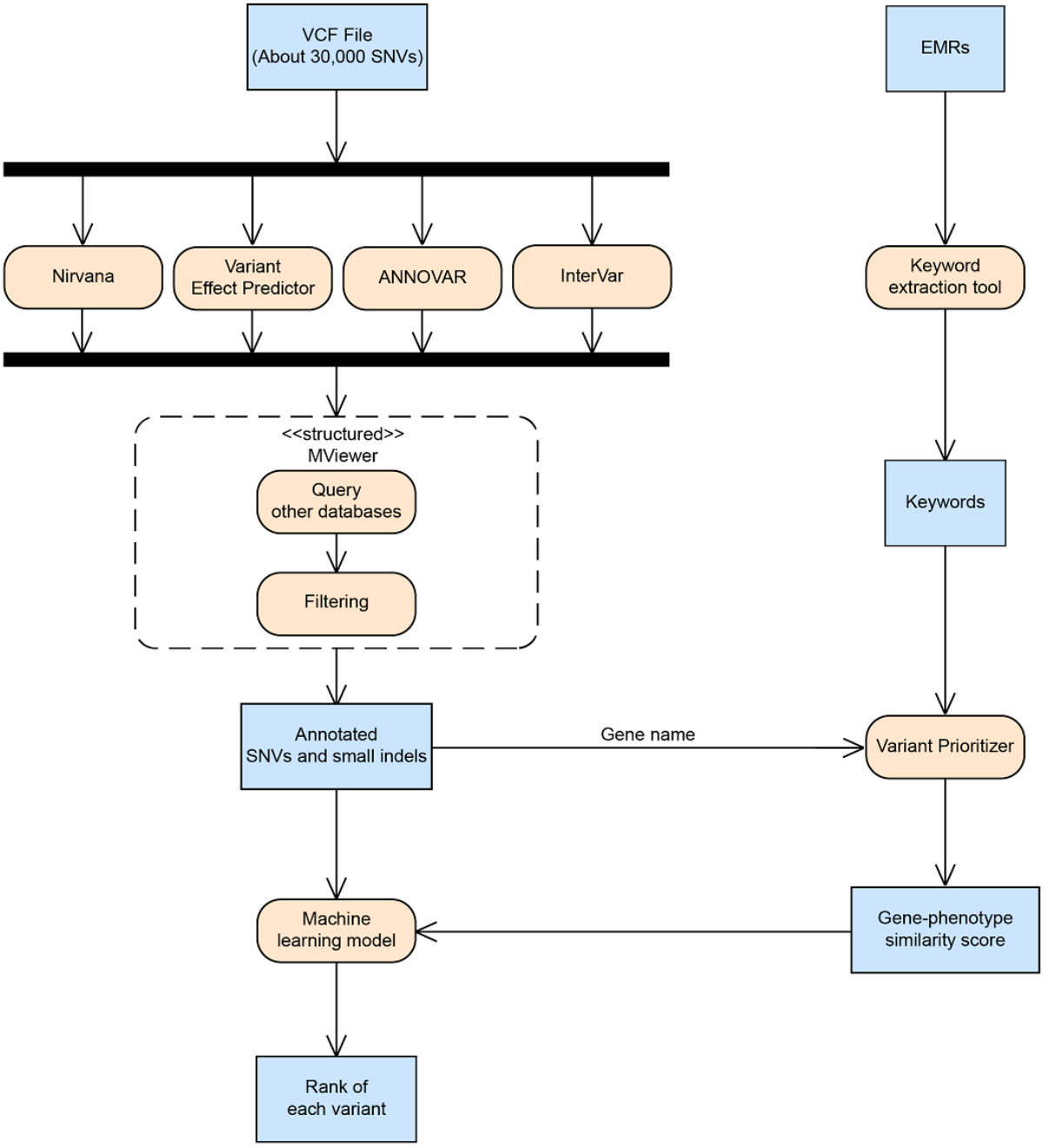
Methods

Workflow

Overview

Figure 1 shows the workflow of our research. We collected VCF of each patient from WES and panel sequencing and then annotated the variants using several tools. After variant annotation, we used our in-house software (MViewer [20]) to query additional external databases and filter for candidate variants. We then used the gene name of these candidate variants and keywords extracted by keyword extraction tools from EMRs to query Variant Prioritizer [21]. The gene similarity scores generated by Variant Prioritizer and columns of annotated variants were used as features to train a machine learning model. This model ranks each variant that represents its disease-causing probability. We will demonstrate the details of each step in the following sections.

Figure 1. The workflow of research. EMR: electronic medical record; indel: insertion/deletion; MViewer: Mutation Viewer; SNV: single-nucleotide variant; VCF: variant call format.



Variant Annotation

We collected each patient's NGS sequencing data in the VCF file and got annotations from several tools, including ANNOVAR [22], VEP [23], Nirvana [24], and InterVar [25]. For additional information that the aforementioned tools will not provide, we used software to import some public data sources, including ClinVar [26], Human Genome Mutation Database (HGMD) [27], and Taiwan Biobank [28]. A detailed description of these annotation fields is summarized in Textbox 1.

Textbox 1. Description of annotation fields.

Allele Frequency

This describes the fraction of gene copies of a particular allele in a defined population. Allele frequency is calculated by dividing the number of copies of a particular allele in a population by the total number of all alleles for that gene in a population. It refers to how common an allele is in a population.

Functional Prediction Score

A range of scoring algorithms with capability to predict the potential deleteriousness of variants based on different information in them, such as their sequence homology, protein structure, and evolutionary conservation. These scoring methods include function prediction scores, conservation scores, and ensemble scores.

Pathogenicity

Clinical significance variants reported in 2 public databases, ClinVar and Human Gene Mutation Database (HGMD), that store information on gene mutation(s) related to human-inherited disease. Both classify variants as disease causing or disease associated by manual curation.

Clinical Interpretation

The American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) published standards and guidelines for the clinical interpretation of sequence variants with respect to human diseases on the basis of 28 criteria [29]. These criteria are as follows: the criteria (16 overall) for classifying variants as pathogenic or likely pathogenic are very strong (PVS1), strong (PS1-PS4), moderate (PM1-PM6), or supporting (PP1-PP5), whereas the criteria (12 overall) for classifying variants as benign or likely benign are standalone (BA1), strong (BS1-BS4), or supporting (BP1-BP7).

Gene-Level Constraint

Constraint on gene expression levels has been shown to influence patterns of genetic variation within humans [30]. For example, some genes are unusually depleted for loss of function and are thought to be constraint with respect to their expression. The Genome Aggregation Database (gnomAD) provides predicted constraint metrics track set that contains metrics of pathogenicity per gene as predicted and identifies genes subject to strong selection against various classes of mutation. These include several subtracks of constraint metrics calculated at gene, transcript, and transcript region levels.

Disease Inheritance

Patterns of inheritance that a trait or disorder associated with a variant can be passed down through families, such as autosomal dominant, autosomal recessive, X-linked, and mitochondrial inheritance. We used the patterns defined in OMIM (Online Mendelian Inheritance in Man) as our data.

Others

Additional information about genetic variants such as the gene name, genotype, and the functional consequence on the different transcripts for a gene or in proximal regulatory regions.

Variant Filtering

There are on average 40,000 variants per proband in WES data. However, most of them are benign and not related to the symptoms. Only a small number of these variants are likely to be deleterious or relevant to the patient's disease. In a standard clinical analysis process, physicians only focus on variants that might be pathogenic or unknown. As our model aims to assist researchers and physicians with their clinical exome reading, reducing the number of variants and focusing on the variants that are more likely to be responsible for the disease are necessary.

For the purpose of generating candidate variants, we used the filter provided by MViewer to remove the variants that are not likely to be deleterious. The filters and criteria are listed in Table 1. For filters that contain more than 1 column, if a variant meets any of their criterion, it will remain in the data. We got approximately 700 SNVs per patient after variant filtering.

Table 1. Filter criteria.

Filter	Column	Criteria
Max allele frequency	• Max Allele Frequency	• ≤0.01 (include no data)
Nonsynonymous missense mutation	• ExonicFunc.refgene	• "nonsynonymous"
Stop gain	• Consequence • ExonicFunc.refgene	• "stop_gained" • "stopgain"
Splice	• Consequence • Func.refgene	• "splice_region_variant" • "splice_acceptor_variant" • "splice_donor_variant" • "splicing"
Frameshift	• Consequence • ExonicFunc.refgene	• "frameshift_variant" • "feature_truncation" • "feature_elongation" • "frameshift"
Initial codon	• Consequence	• "start_lost"
Deletion	• Type • Consequence • ExonicFunc.refgene	• "deletion"
Insertion	• Type • Consequence • ExonicFunc.refgene	• "insertion"
Inframe deletion	• Consequence • ExonicFunc.refgene	• "inframe_deletion" • "nonframeshift deletion"
Exon/splice site	• Func.refgene • Consequence	• "exonic" • "splicing" • "coding_sequence_variant" • "frameshift_variant" • "incomplete_terminal_codon_variant" • "inframe_deletion" • "inframe_insertion" • "missense_variant" • "splice_acceptor_variant" • "splice_donor_variant" • "splice_region_variant"

Phenotype Extraction

Overview

The phenotype information used in this research is from clinicians' history summary. The records were all free text and the length of texts varied from less than 10 to more than 300 words. In the clinical analysis process, it is time consuming for physicians to go through the medical records and identify the phenotype keywords manually. To solve this problem, we used several keyword extraction tools to automatically generate keywords related to phenotype from free-text medical records. The keyword extraction tools applied in our research are listed in the following sections.

MetaMap

MetaMap [31] is a widely used application providing access to the concepts in the Unified Medical Language System (UMLS) Metathesaurus [32]. The UMLS Metathesaurus is a compilation of names, relationships, and associated information from a variety of biomedical naming systems representing different views of biomedical practice or research. It comprises over 1 million biomedical concepts and 5 million concept names [33]. MetaMap is able to map every word in the texts to UMLS concepts, but we just wanted to focus on those associated with phenotypes and diseases. Thus, we extracted the words that are classified as the semantic types of the following: (1) injury or poisoning, (2) cell or molecular dysfunction, (3) genetic function, (4) disease or syndrome, (5) sign or symptom, (6) tissue.

Doc2Hpo

Doc2Hpo [34] is a web application using natural language processing (NLP) techniques to parse clinical note and get the phenotype concept curation as the HPO term. There is a parsing engine that will automatically recognize the phenotype concepts from the input. Doc2Hpo applies an algorithm called NegBio

for negation detection in the input data. After that, there are several NLP engines responsible for HPO concept extraction. We used 3 of these engines and compared the performance of each of them. The first NLP engine is a string-based method that leverages the algorithm for concept extraction. The second engine is the online NCBO Annotator [35] API for HPO concept recognition. The last engine we adopt is MetaMap Lite, which is a fast version of MetaMap that provides near–real-time named entity recognition. The MetaMap Lite engine first identifies clinical terms in the texts and maps them to standard UMLS concepts. The UMLS concepts will then be further mapped to HPO concepts. Results generated by Doc2Hpo are HPO terms, whereas keywords extracted by MetaMap are nonHPO terms.

Phenotype-Gene Similarity Score

Another method to construct the connections between genes and keywords is using the Okapi BM25 ranking function. Okapi BM25 is usually used by search engines, such as Google and Bing, to rank matching documents according to their relevance to a given search. One of the most prominent instantiations of the function is as the following equation:

where score(D, Q) represents the Okapi BM25 score of a document D when given a query Q, containing keywords $q1$, $q2$,...,qn; $f(qi, D)$ is qi's term frequency in the document D; $|D|$ is the length of document D in words; avgdl is the average document length among all documents; $k1$ and b are constants (=1.2 and 0.8, respectively); and IDF(qi) is the inverse document frequency (IDF) weight of the query term qi and is usually defined as:

$$\text{IDF}(qi) = \ln [(N - n(qi) + 0.5]/[n(qi) + 0.5 + 1]$$

where N is the number of documents and n is the number containing the keywords.

In this research, we propose an idea using gene description from OMIM and GeneReviews as documents and keywords as query to implement the Okapi BM25 ranking function. Therefore, we can use the Okapi BM25 score to represent the relationship between gene description and keywords. The higher score that gene description gets from keywords indicates stronger connection between that gene and keywords. Rank values were based on the Okapi BM25 ranking function mentioned before with some other parameters. Compared with the Okapi BM25 regular formula, rank value replaces the IDF function with Robertson-Spärck-Jones weight [36]. The IDF function is a measure of how much information the word provides, that is, whether the word is common or rare across all documents. For example, the term "the" is very common in every document, so term frequency will be inclined to falsely highlight the documents that happen to use the word "the" more frequently. Hence, the IDF function is dedicated to reducing the weight of words that appear very frequently among all documents. In contrast to the regular IDF function, the Robertson-Spärck-Jones weight adds relevant parameters of documents and increases the precision of rank score.

We get the phenotype-gene similarity score of each SNV from Variant Prioritizer, a text mining tool that outputs the rank and score of genes by entering symptoms as keywords. Variant Prioritizer uses the Okapi BM25 ranking function [37] to construct the connections between genes and keywords. Gene descriptions from OMIM, GeneReviews, Entrez Gene [38], and PubTator [39] serve as data sources and keywords as query to implement the Okapi BM25 score using the full-text search method. It returns a column called RANK that includes ordinal value from 0 to 1000. The RANK score is based on the following formula:

where ω is the Robertson-Spärck-Jones weight [36], which is defined as $\omega = \log [(r + 0.5)\cdot(N - n - R + r + 0.5)]/[(R - r + 0.5)\cdot(n - r + 0.5)]$, in which R is the number of known relevant documents and r is the number of these containing the term; tf is the frequency of the word in the property queried within an article; qtf is the frequency of the term in the query; and K is defined as follows:

$$K = k_1[(1 - b) + b(dl/\text{avgdl})]$$

where dl is the property length, in word occurrence; avgdl is the average length of the property being queried, in word occurrence; and k_1, b, and k_3 are constants (=1.2, 0.75, and 8.0, respectively).

We employed the Variant Prioritizer API to get the RANK value from each data source as gene similarity score to represent the association between each SNVs and extracted keywords. We kept the maximum and minimum scores of rank values (4 overall) as 2 separate features for model building.

Ethical Considerations

This retrospective cohort study was approved by the Institutional Review Board (IRB) of the National Taiwan University Hospital (IRB number: 201710066RINB). We confirm that all experiments were performed in accordance with relevant guidelines and regulations. The data retrieved from EHRs were deidentified and could not be linked to the patients' identity by the research team. The need for written informed consent was waived and confirmed by the National Taiwan University Hospital IRB (201710066RINB) because this was a retrospective cohort study with deidentified data. This regulation complies to Health Insurance Portability and Accountability Act (HIPAA) that there are no restrictions on the use or disclosure of deidentified health information.

Data Preprocessing

Overview of Steps

After variant annotation of the VCF file, we preprocessed our data into a model-acceptable format. Data preprocessing is an extremely important step in machine learning because the quality of data can directly affect the ability of a model to learn. It includes various operations and each operation aims to help machine learning build better predictive models. The data preprocessing operations used in this research are explained in the following sections.

Missing Value Handling

In real world, the data usually have missing values. AsFor example, in the genotype variable most machine learning methods cannot deal with null value, it is pivotal to identify and correctly handle the missing values. Basically, the missing values can be handled using various techniques such as deletion or imputation [40]. Deletion removes all data for an observation that has 1 or more missing values. However, if there are many columns with missing values, then deletion will result in the lack of data. Therefore, for some columns we used imputation by substituting the missing values in our data set with mean and for some columns we just simply replaced the missing value with a valid value such as 0.

One Hot Encoding

Many machine learning algorithms cannot operate on categorical data directly. They require all input features to be numeric. Basically, categorical data contain label values rather than numeric values. As a consequence, categorical data must be converted into a numerical form so that they can be used in the machine leaning model. One hot encoding is a widespread approach for dealing with categorical data. One hot encoding transforms a categorical column to a multidimensional vector. It creates new columns, indicating the presence of each possible value from the original data.

For example, in the genotype variable, there are 3 categories: homozygous (hom), heterozygous (het), and hemizygous (hem). Therefore, 3 binary variables [hom, het, hem] are needed. If genotype of a variant is heterozygous, we use [0,1,0] to represent it.

Data Normalization

For continuous data, there are a few with different ranges. If we apply features in very different ranges to some machine learning models such as logistic regression, the feature with broader range will intrinsically influence the result more owing to its larger value. However, this does not necessarily mean that this feature is more important as a predictor. Therefore, we used normalization techniques as a solution to overcome this problem. Normalization is the rescaling of the data from the original range so that all values are within the range of 0 and 1. We rescale all continuous values by min-max normalization. The general formula is as follows:

$$X\text{norm} = (X - X\text{min})/(X\text{max} - X\text{min})$$

where X is the original value and Xnorm is the normalized value. This will make the maximal value map to 1 and the minimal value map to 0. In addition to the aforesaid data preprocessing techniques, we handled different data types in different ways and created some new features for model building. In the following sections, we elaborate on each data type preprocessing and combine them in the end.

Functional Prediction Score

Functional prediction scores including SIFT [17], PolyPhen2 HDIV [16], PolyPhen2 HVAR [16], LRT [41], MutationTaster [18], MutationAssessor [42], FATHMM [43], PROVEAN [44], MetaSVM [14], MetaLR [14], M-CAP [45], CADD [13], GERP++ [46], DANN [47], fathmm-MKL [48], GenoCanyon [49], fitCons [50], PhyloP [51], PhastCons [52], and SiPhy [53] were from ANNOVAR. We used converted rank scores provided by ANNOVAR instead of the original score because all these scores are always within the range of 0 and 1. Besides, converted rank scores from different algorithms are monotonic in the same direction. That is, a higher score indicates that the variant is more likely to be damaging [54]. For splice site prediction, we imported the MaxEntScan score using the VEP plugin. We defined a new column called MaxEntScan significance. The value is 1 when the value of MaxEntScan alt is smaller than 3 and MaxEntScan variation is smaller than 30%; otherwise the value is 0. We used clinical significance reported in ClinVar and computed rank score from the HGMD. The HGMD computed rank score is a probability of pathogenicity between 0 and 1, with 1 being most likely disease causing compared with other HGMD entries.

Clinical Interpretation

We employed clinical interpretation of each genetic variant based on the American College of Medical Genetics and Genomics/Association for Molecular Pathology (ACMG/AMP) 2015 guideline, which is generated by InterVar. We calculated the ACMG score developed by Xrare to represent its overall pathogenicity. The ACMG score is a weighted sum score based on multiple evidence (n=14) with the following weights for each term: PVS1:6, PS1:4, PM1:2, PM2:2, PM4:2, PM5:2, PP2:1, PP3:1, BA1:9, BS1:3, BS2:3, BP3:1, BP4:1, BP7:2 [9]. We collected gene-level constraint features including pLI, pRec, syn_z, and mis_z from the Genome Aggregation Database (gnomAD). We used the patterns of inheritance defined in OMIM as our data. For variants that contain multiple patterns, we calculated the occurrences of each pattern and stored it as a feature. We also get some additional information about each variant from ANNOVAR such as genotype, regions that a variant hits, and read depths. The quality of each variant is also collected from the VCF file. As the genotype annotated by ANNOVAR does not contain hemizygous alleles, we replaced the genotype feature of all male patients' chromosome X with hemizygous alleles. In addition, we collected functional consequence on the different transcripts for a gene or in proximal regulatory regions using Nirvana.

Labels

The goal of our research was to identify the disease-causing variants with SNVs (ie, we classify a variant as disease causing or not). As machine learning algorithms learn how to assign a class label to a test case from examples, it is necessary to assign a class label to all input training sets. We used the 0/1 label to represent whether a variant is disease causing or not. If a variant is causative, we assigned label 1 to it; otherwise the label is 0. Details about all the features used in our model are presented in Multimedia Appendix 2.

Feature Selection

After data preprocessing, we got 94 features for each variant. To reduce the high dimension of the input data set while retaining the discriminatory information for classification problems, we applied univariate feature selection techniques from scikit-learn [55] packages to identify the relevant variables

in a data set and eliminate the useless variables. Feature selection helps to reduce the noise in the data set and lets the model focus on the relevant signals.

There are several scoring functions provided by scikit-learn univariate feature selection modules. We used mutual information classifier to select the most relevant variables. Mutual information [56] between 2 random variables is a nonnegative value, which measures the general dependence of variables without making any assumptions about the nature of their underlying relationships [57]. The mutual information between 2 discrete random variables X and Y is defined as follows:

where $p(x, y)$ is the joint probability density function of X and Y, and $p(x)$ and $p(y)$ are the marginal density function. The mutual information determines the similarity between the joint distribution $p(x, y)$ and the products of the factored marginal distributions. The larger the value means the greater the relationship between the 2 variables. The calculated value is equal to 0 if and only if the 2 variables are independent.

We performed the feature selection process using only the training set to determine the relevant variable. Further, the number of features we selected is based on model evaluation with 10fold cross validation

Building Model

To construct a model by machine learning algorithm, we split the data into 2 groups. As our model aims to assist physicians with their clinical exome data interpretation process, the exome data from the dbGaP database and the targeted gene panel sequencing data from NTUH were set as training set, and the WES data from NTUH were set as testing data. which can only be used on model evaluation. The external validation set consisted of 90 most recent NTUH WES data, which help to make sure that our model can make predictions in future clinical use. Details about the training and testing sets are listed in Table 2.

To build the machine learning model, we implemented the random forests algorithm [58] provided by scikitlearn packages. The selection of hyperparameters is based on a grid search with 10fold cross validation. Random forest was first proposed by Leo Breiman in 2001 [58]. It is an ensemble classifier that evolves from decision trees. Actually, random forests are a combination of decision trees such that each tree depends on the values of a random vector sampled independently, with the same distribution for all trees in the forest [59]. A forest of trees is grown as follows:

- The training set is a bootstrap sample from the original training set.
- The number of trees to build and the number of variables randomly sampled as candidates at each split m-try are set by the user, where m-try is less than the total number of variables.
- At each node, m-try variables are selected at random, and the node is split on the best split point among m-try. This process iterates until the tree grows to its maximal depth.
- For test case prediction, as a test vector **x** is put down at each tree, it is assigned the average of **y** values at the node it stops at. The average of these overall trees in the forest is the predicted value for **x**. The predicted value for classification is the class getting the plurality of the forest votes..

The function we used to measure the quality of a split is Gini impurity. Gini impurity is the probability of incorrectly classifying a randomly chosen element in the data set if it were randomly labeled according to the class distribution in the data set [60]. In decision tree learning it is defined as  , where c is the number of classes and $p(i|t)$ is the probability of randomly picking an object of class i at node t. The optimal split from a root node when training a decision tree is chosen by maximizing the Gini gain, which is calculated by subtracting the weighted impurities of the branches from the original impurity.

Table 2. The training, testing, and external validation sets used in this study.

Data	Training set	Testing set	External validation set
Source	dbGaP[a], NTUH[b] panel	NTUH WES[c]	New NTUH WES
Patients, n	381	108	90
Filtered variants, n	125,693	80,083	109,857
Causative variants, n	478	134	100

[a]dbGaP: Database of Genotypes and Phenotypes.

[b]NTUH: National Taiwan University Hospital.

[c]WES: whole-exome sequencing.

Performance Evaluation

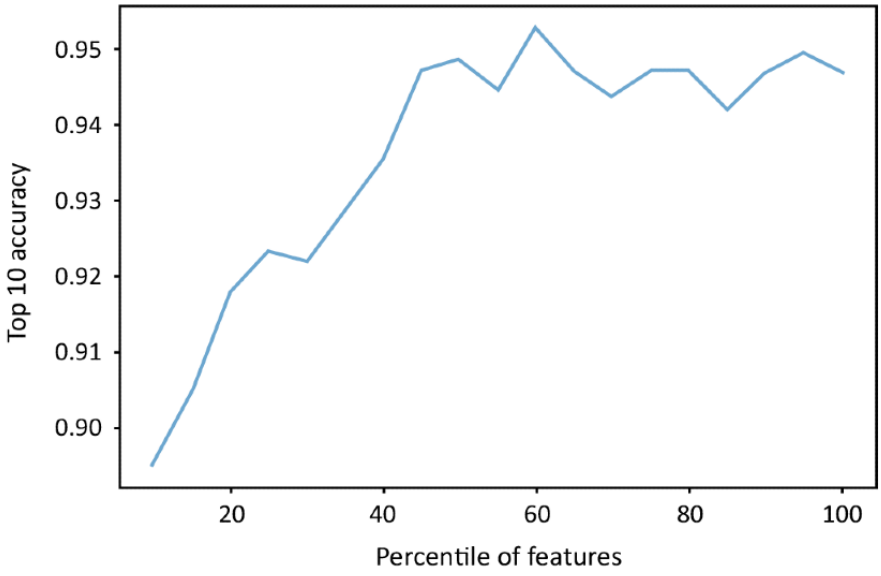
To evaluate our model performance of true causative variant prioritization, we used the ranking statistics mentioned in VarSight. After we applied the binary classification process to all variants, we got a probability for each variant that represents the probability of this variant to be disease causing. We ranked the variants for each patient from the highest to lowest probability and quantified the percentage of the target variants that were ranked in the top 1, 5, 10, 20.

Results

Feature Selection

For the feature selection, we used univariate feature selection based on the SelectPercentile method in scikitlearn package. The classifier we chose is the mutual information classifier. Only the training set was used for selecting the most relevant features. Further, we applied 10fold cross validation to decide the number of features for model training. In Figure 2, we present the top 10 accuracy on 10fold cross validation using different percentages of features. As using 60% of features achieves the highest accuracy, 56 features (60% of total 94 features) with the highest estimated mutual information were selected for the final model building.

Figure 2. The top 10 accuracy on 10-fold cross validation using different percentage of features.



Model Performance

We evaluated the model with our testing set. As mentioned in Table 2, the testing set comprised 108 patients who received WES with at least one disease-causing variant diagnosed by doctors. Multimedia Appendix 3 presents detailed information about their causative variants, keywords, and the corresponding HPO term. The keywords and HPO term are determined by doctors based on the phenotype of each patient.

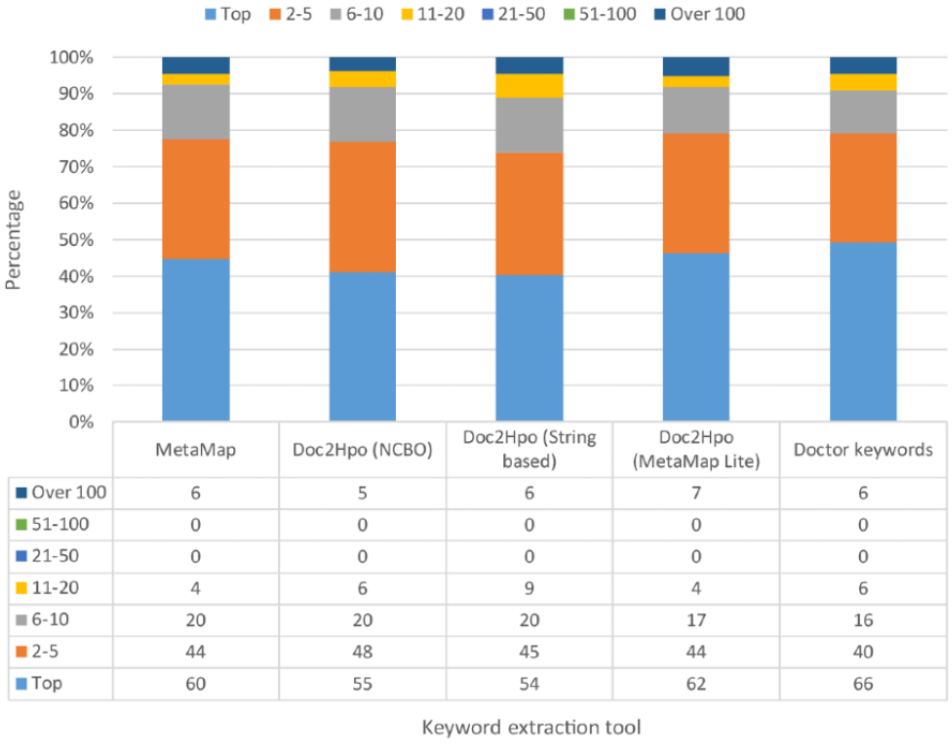
Prediction With Different Keyword Extraction Tools

Figure 3 shows the percentage distribution of the ranking of target variants and Figure 4 shows the cumulative rank result of models using different keyword extraction tools. When using tools to extract phenotypes from abstracts, our model can assign the target variants to the top rank for over 40% (60/134, 44.8%) of the total variants. The top 10 accuracies of models are around 90% (124/134, 92.5%), irrespective of the keyword extraction tool used. Compared with the keywords provided by professional doctors, applying tools to extract keywords had lower top 1 accuracy but comparable top 10 accuracy. This indicated that in most cases our model can successfully rank the true causative variants in the front of the variant lists and the rank is slightly influenced by the input keywords.

We built a random forest model based on the method described in the previous section and evaluated it with our testing set based on different keyword extraction tools. We succeeded in locating 92.5% (124/134) of the causative variant in the top 10 ranking list among an average of 741 candidate variants per person after filtering. The performance of the model is similar to that of manual analysis, and it has been used to help National Taiwan University Hospital with a genetic diagnosis.

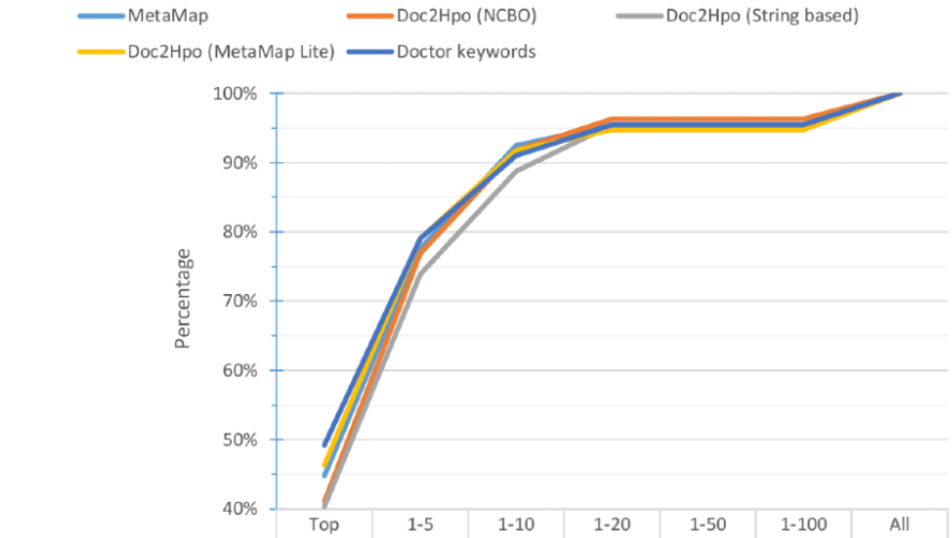
Figures 3 and 4 show the percentage distribution of the ranking of target variants and the cumulative rank result of models using different keyword extraction tools, respectively. When using tools to extract phenotypes from abstracts, our model can assign the target variants to the top rank for over 40% (60/134, 44.8%) of the total variants. The top 10 accuracies of models are around 90% (124/134, 92.5%), irrespective of the keyword extraction tool used. Compared with the keywords provided by professional doctors, applying tools to extract keywords has lower top 1 accuracy but comparable top 10 accuracy. It represents that in most cases our model can successfully rank the true causative variants in the front of the variant lists and the rank is slightly influenced by the input keywords.

Figure 3. Percentage distribution of ranks.



	MetaMap	Doc2Hpo (NCBO)	Doc2Hpo (String based)	Doc2Hpo (MetaMap Lite)	Doctor keywords
Over 100	6	5	6	7	6
51-100	0	0	0	0	0
21-50	0	0	0	0	0
11-20	4	6	9	4	6
6-10	20	20	20	17	16
2-5	44	48	45	44	40
Top	60	55	54	62	66

Figure 4. Cumulative percentage distribution of ranks. NCBO: National Center for Biomedical Ontology.

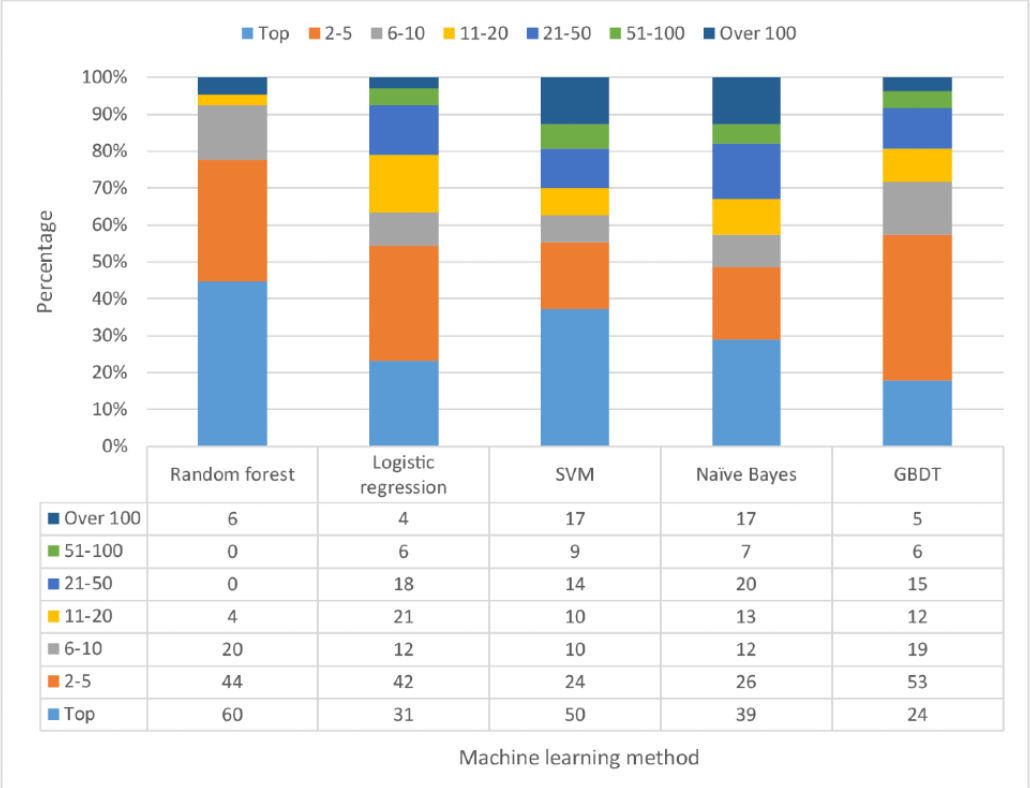


	Top	1-5	1-10	1-20	1-50	1-100	All
MetaMap	44.78%	77.61%	92.54%	95.52%	95.52%	95.52%	1
Doc2Hpo (NCBO)	41.04%	76.87%	91.79%	96.27%	96.27%	96.27%	1
Doc2Hpo (String based)	40.30%	73.88%	88.81%	95.52%	95.52%	95.52%	1
Doc2Hpo (MetaMap Lite)	46.27%	79.10%	91.79%	94.78%	94.78%	94.78%	1
Doctor keywords	49.25%	79.10%	91.04%	95.52%	95.52%	95.52%	1

Other Machine Learning Methods

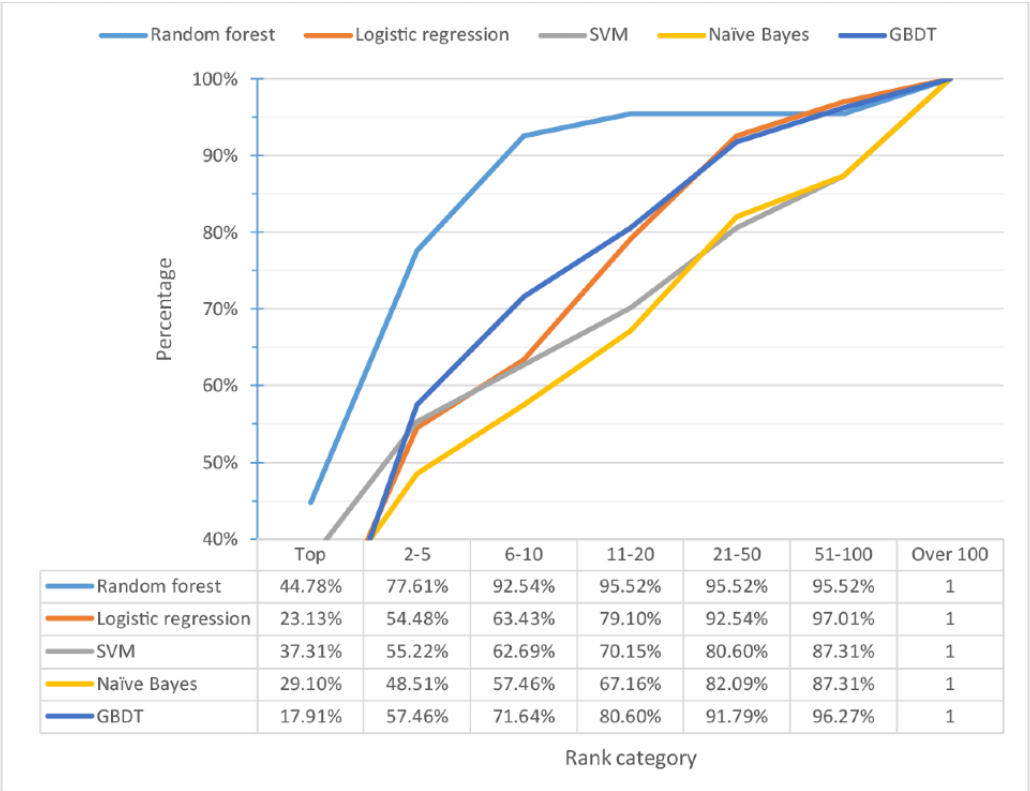
We also evaluated other machine learning methods and compared their performance with random forest. These methods include logistic regression, Gaussian naive Bayes, SVM with RBF kernel, and gradient boosted decision trees. The selection of hyperparameters for each algorithm was based on grid search with 10-fold cross validation. We used MetaMap as the keyword extraction tool and our testing data to test the performance of each method. The percentage distribution of the ranking of target variants by each machine learning method and the cumulative rank result of each model are shown in Figures 5 and 6, respectively. As random forest got the highest top 10 accuracy, we finally chose random forest as our machine learning algorithm.

Figure 5. Percentage distribution of ranks. GBDT: gradient boosting decision tree; SVM: support vector machine.



	Random forest	Logistic regression	SVM	Naïve Bayes	GBDT
Over 100	6	4	17	17	5
51-100	0	6	9	7	6
21-50	0	18	14	20	15
11-20	4	21	10	13	12
6-10	20	12	10	12	19
2-5	44	42	24	26	53
Top	60	31	50	39	24

Figure 6. Cumulative percentage distribution of ranks. GBDT: gradient boosting decision tree; SVM: support vector machine.



	Top	2-5	6-10	11-20	21-50	51-100	Over 100
Random forest	44.78%	77.61%	92.54%	95.52%	95.52%	95.52%	1
Logistic regression	23.13%	54.48%	63.43%	79.10%	92.54%	97.01%	1
SVM	37.31%	55.22%	62.69%	70.15%	80.60%	87.31%	1
Naïve Bayes	29.10%	48.51%	57.46%	67.16%	82.09%	87.31%	1
GBDT	17.91%	57.46%	71.64%	80.60%	91.79%	96.27%	1

Discussion

Principal Findings

We have implemented a website, AI Variant Prioritizer, which uses data from NGS bioinformatics pipelines with machine learning to make a prediction about most possible disease-causing variants among SNVs and patient's phenotype data. This system can assist researchers and physicians by focusing on those with higher disease-causing probability and reducing the average turnaround time (by 1 day) of the entire WES pipeline, from DNA extraction to clinical diagnosis.

Moreover, we have implemented a web API for our system so that the ranking function could be integrated into MViewer. Thus, physicians can interpret the results of genetic variation with a single application instead of adopting numerous services.

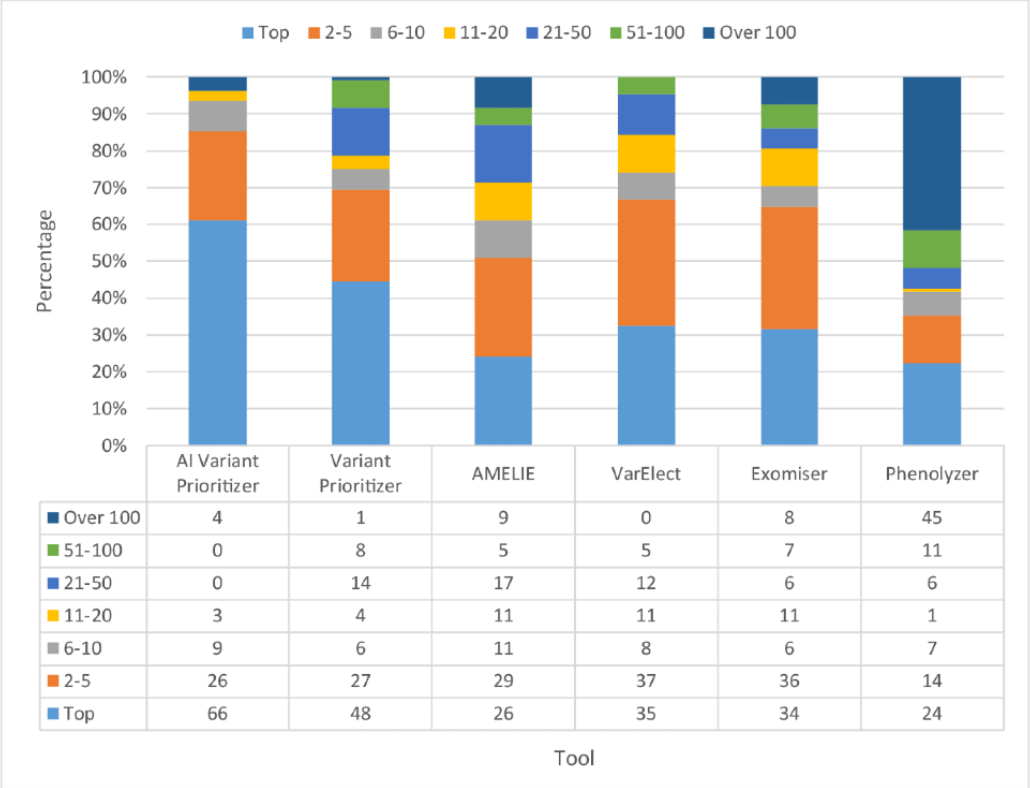
For comparison, we used our testing data to run several prioritization tools including AMELIE [61], VarElect [62], Exomiser, Phenolyzer, and Variant Prioritizer. As AMELIE and Exomiser can only accept phenotypes defined in HPO terms, we entered HPO terms determined by doctors as their input. Phenolyzer can identify both disease terms and HPO terms. However, if the terms do not match in their databases, it will not return any record. Hence, we also used HPO terms as input for Phenolyzer. VarElect, Variant Prioritizer, and our model can identify free-text descriptions. Therefore, we imputed the keywords provided by doctors as input for testing. AMELIE, VarElect, and Variant Prioritizer only prioritize the gene list instead of the variant list. Hence, we evaluated the result for gene-based prioritization instead of variant-based prioritization. That is, for each patient, if the tools prioritize target gene in the top 1, 5, 10, 20, 50, and 100 of our filtered gene lists, this patient will be counted. All the tools use the default settings provided in their websites to run.

Figures 7 and 8 show the percentage and cumulative percentage distribution of the target gene ranking for each tool, respectively. From Figure 8, we can see that AI Variant Prioritizer is able to assign the target gene to the top rank for 61.1% (66/108) of the patients, followed by Variant Prioritizer (48/108, 44.4%). It also shows the cumulative rank result, which reveals that our AI Variant Prioritizer has the highest accuracy at ranks 1, 5, 10, and 20. Further, AI Variant Prioritizer shows better performance than other tools. After adopting the HPO terms by looking up the databases, the top 10 ranking list can be increased to 93.5% (101/108).

In comparison with extraction of phenotypic features from SNOMED through manual mapping of HPO terms to SNOMED Clinical Terms (SNOMED CT) [63], our automation approach explores various phenotypic feature extraction tools and focuses on rare disease interpretation. We have also looked into several AI-driven variant prioritization approaches published in the last 3 years, including Fabric GEM [12], MOON [2], and Exomiser. There are several differences between our approach and each of these approaches, including the algorithms used to build the prioritization model, the features considered, and databases integrated. However, the major difference of our approach from others is the method used to turn the relationships between genes and phenotypes into numerical values, which makes way for later prediction. Fabric GEM and MOON utilize Phevor [15] to turn phenotype-gene relationship into numerical values, whereas Exomiser uses PhenoDigm [64] to achieve this goal.

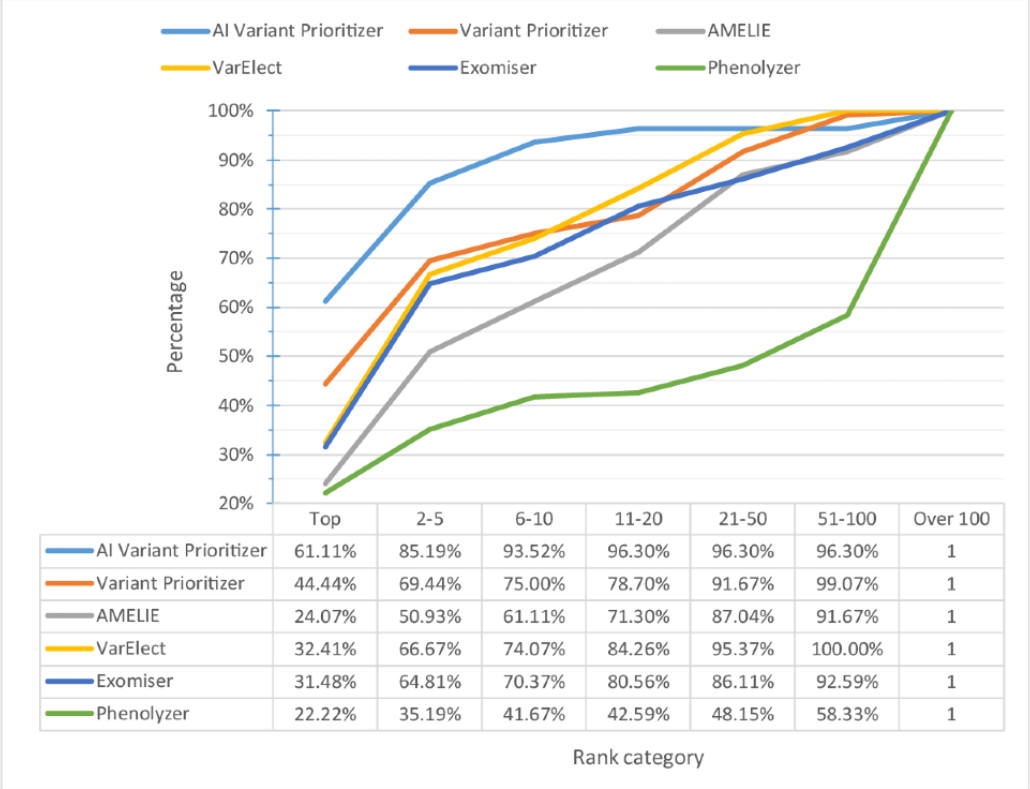
Both Phevor and PhenoDigm construct new methods that bridge HPO and other ontologies to discover more gene-disease associations. Phevor gathers all correlation of diseases and genes provided by HPO and Gene Ontology (GO) and constructs several networks (graphs) and distributes decreasing weights along the paths found. The sum of weights on the specific gene node represents the correlation score of the gene with the corresponding HPO term. PhenoDigm utilizes OWLSim [65] to calculate the similarity among different phenotypes in different ontologies and uses similarity to estimate the magnitude of correlation of given HPO terms and different genes. By contrast, Variant Prioritizer used in our approach extracts the phenotype-gene relationship from a different kind of knowledge source: free text of database. We make a simple comparison of these approaches in Tables 3 and 4.

Figure 7. Percentage distribution of ranks. AI: artificial intelligence.



	AI Variant Prioritizer	Variant Prioritizer	AMELIE	VarElect	Exomiser	Phenolyzer
Over 100	4	1	9	0	8	45
51-100	0	8	5	5	7	11
21-50	0	14	17	12	6	6
11-20	3	4	11	11	11	1
6-10	9	6	11	8	6	7
2-5	26	27	29	37	36	14
Top	66	48	26	35	34	24

Figure 8. Cumulative percentage distribution of ranks. AI: artificial intelligence.



	Top	2-5	6-10	11-20	21-50	51-100	Over 100
AI Variant Prioritizer	61.11%	85.19%	93.52%	96.30%	96.30%	96.30%	1
Variant Prioritizer	44.44%	69.44%	75.00%	78.70%	91.67%	99.07%	1
AMELIE	24.07%	50.93%	61.11%	71.30%	87.04%	91.67%	1
VarElect	32.41%	66.67%	74.07%	84.26%	95.37%	100.00%	1
Exomiser	31.48%	64.81%	70.37%	80.56%	86.11%	92.59%	1
Phenolyzer	22.22%	35.19%	41.67%	42.59%	48.15%	58.33%	1

Table 3. The comparison among AI Variant Prioritizer, Fabric GEM, MOON, and Exomiser.

Tool	AI[a] Variant Prioritizer	Fabric GEM	MOON	Exomiser
Variant scoring algorithm	Random forest	Bayes factor	Decision trees, Bayesian models, neural networks	Rule based
Phenotype-gene score	Variant Prioritizer	Phevor	Phevor	PhenoDigm
Phenotype input format	Plain texts	HPO[b] terms	HPO terms extracted from electronic health record	HPO terms

[a]AI: artificial intelligence.

[b]HPO: Human Phenotype Ontology.

Table 4. The comparison among Variant Prioritizer, Phevor, and PhenoDigm.

Tool	Variant Prioritizer	Phevor	PhenoDigm
Algorithm	Okapi BM25	Graph algorithm	OWLSim
Phenotype input format	Plain texts	HPO[a] terms	HPO terms
Knowledge base	OMIM[b], GeneReviews, Entrez Gene and PubTator	HPO and GO[c]	OMIM (HPO), Sanger-MGP [66], MGD [67], and ZFIN [68]

[a]HPO: Human Phenotype Ontology.

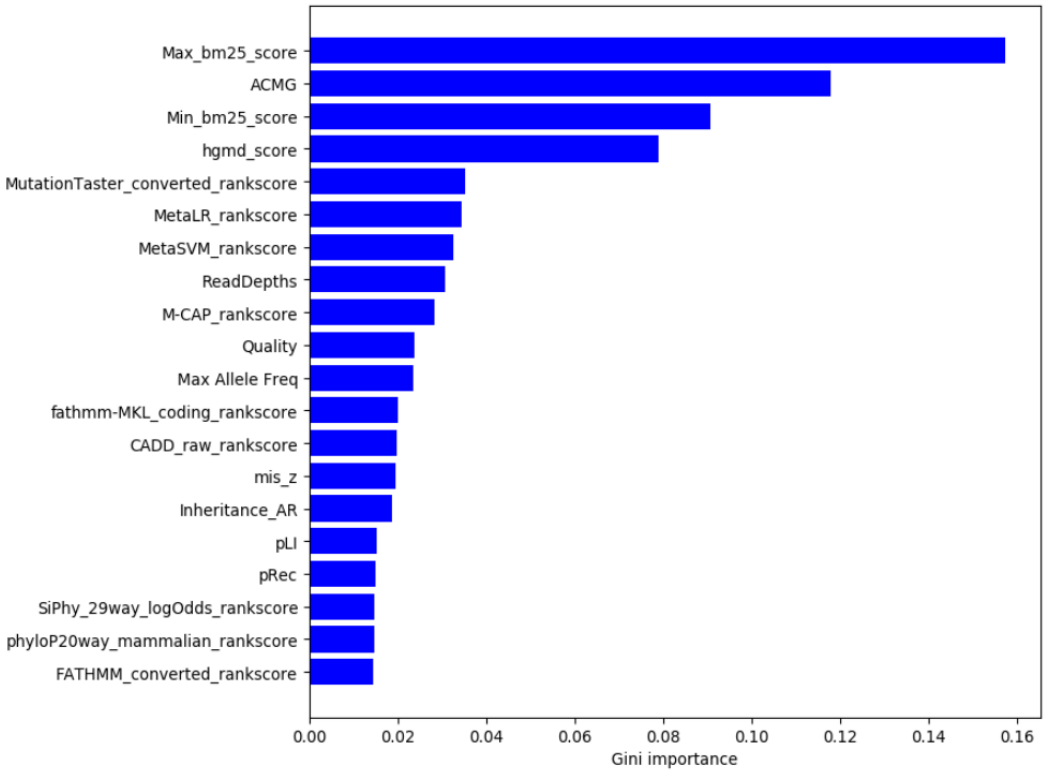
[b]OMIM: Online Mendelian Inheritance in Man.

[c]GO: Gene Ontology.

Feature Importance

For interpreting model predictions, we used the feature importance method provided by scikit-learn to identify which feature has the most predictive power. Figure 9 shows the top 20 important features. According to clinical experience, the connection between a variant and phenotype of a patient is a key factor that influences the physician to decide whether to report a variant or not. Similarly, from the figure we can see that the most important feature is the max bm25 score, which refers to the similarity score between the given variant and keywords. Another important factor that influences the reporting decision during clinical analysis is the severity of a variant. The corresponding feature we use is the ACMG score, which is in the second place of feature importance. By contrast, the result of feature importance might provide information for physicians or researchers about the features that they can consider when finding causative variant.

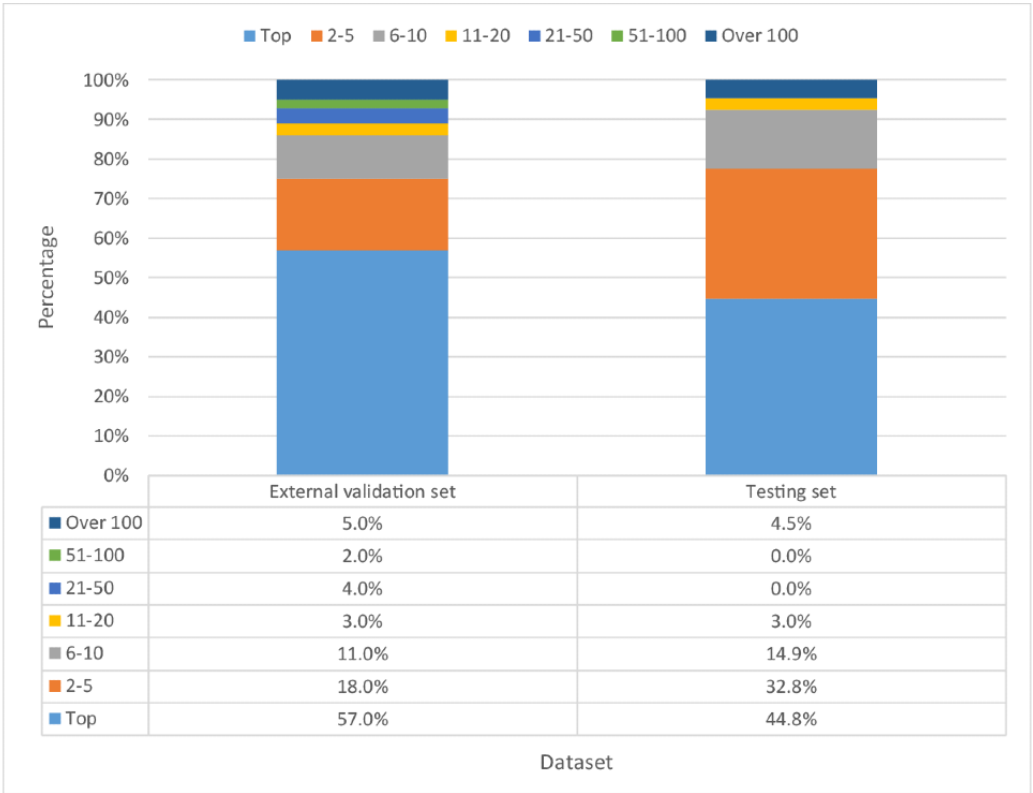
Figure 9. Feature importance.



External Validation

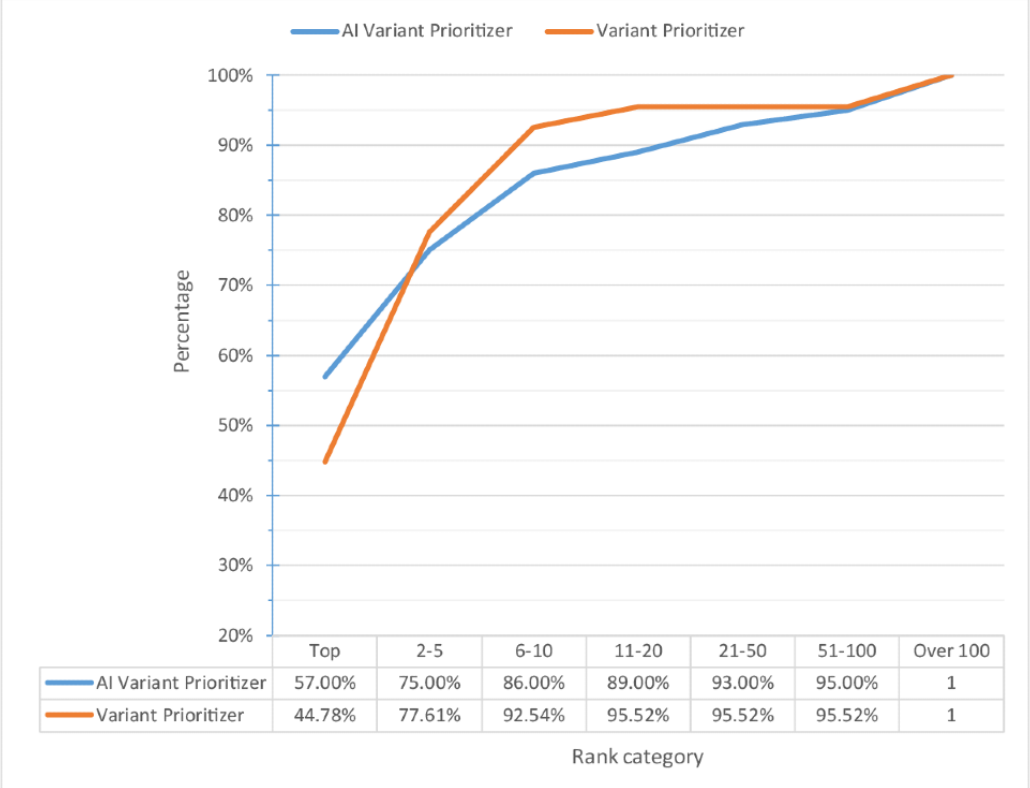
We compared the cumulative percentage distribution of ranks from the testing set and the external validation set. The result is shown in Figures 10 and 11. Their percentage values are close to each other in different regions such as top 10 and top 5. The percentage of top 1 rank of the external validation set is even higher than that of the testing set. With this result, we believe that our approach has shown its potential for robust clinical usage.

Figure 10. Percentage distribution of ranks.



	External validation set	Testing set
Over 100	5.0%	4.5%
51-100	2.0%	0.0%
21-50	4.0%	0.0%
11-20	3.0%	3.0%
6-10	11.0%	14.9%
2-5	18.0%	32.8%
Top	57.0%	44.8%

Figure 11. Cumulative percentage distribution of ranks.



	Top	2-5	6-10	11-20	21-50	51-100	Over 100
AI Variant Prioritizer	57.00%	75.00%	86.00%	89.00%	93.00%	95.00%	1
Variant Prioritizer	44.78%	77.61%	92.54%	95.52%	95.52%	95.52%	1

Limitations

The study has several potential limitations. First, we could not find massive data for training and testing. This not only restricts the amount of teaching material for the machine learning model, but also restricts available measurements to evaluate the trained model. Second, the gene-phenotype score used in this study did not have enough power to detect small or moderate associations because it relies on how frequently the gene-phenotype relationship is reported to the databases it utilizes. Finally, the study did not adjust for potential confounders, such as diet and physical activity. This could cause potential bias because the way in which genes are expressed can be impacted by lifestyle of patients.

Overall, this study could have potential bias resulting from the lack of sufficient data, lack of reported gene-phenotype relationship, and lack of observation of lifestyle. The impact from the first and the second can be reduced if there are more data and reports available in the future. On the other side, the influence of lifestyle and environment remains an issue that needs more dedicated studies.

Conclusions

In this research, we proposed a machine learning model, AI Variant Prioritizer, to predict whether a variant is disease causing for patients with rare Mendelian disorder. We have successfully applied sequencing data from WES and free-text phenotypic information of patient's disease automatically extracted by keyword extraction tools for model training and testing. By interpreting our model, we identified which features of variants are important. Besides, we achieved a satisfactory result on finding the target variant in our testing data set. After testing 108 patients' WES data, we succeeded in 93.5% (n=101) of the cases to locate the causative variant in the top 10 ranking list among an average of 741 candidate variants per person after the filtering process. The performance of the model is similar to that of manual analysis by the physicians in the Department of Medical Genetics, NTUH, and it has been used to help NTUH with a genetic diagnosis.

As the physicians are very busy almost all the time in taking care of their patients, the search time spent for an accurate genetic diagnosis is extremely important. Our AI predicting model can provide the top 10 hit list with a high probability of 93.5% (101/108), thus helping them save weeks of time if they have to do it manually to search beyond the top 10 list very often.

It is not an easy work to fully interpret the causative variations of a genetic disease. As the precision of the keywords extracted by tools influence the performance of our model, for the future work, we will adopt some NLP techniques such as Bidirectional Encoder Representations from Transformers (BERT) to extract keywords more properly. In addition, the AI Variant Prioritizer model has been built to analyze SNVs and small indels from WES data, but we have not dealt with copy number variations (CNVs) yet. CNVs have been recognized as critical genetic variations, which are associated with both common and complex diseases, and thus have a large influence on several Mendelian and somatic genetic disorders. Therefore, we will collect data on CNVs and extend the capability of our system to annotate and filter CNVs. Furthermore, we will enlarge our data set by adding CNVs as our training data to enable the AI Variant Prioritizer model to predict any kind of causative genetic variations. Before implementation of AI Variant Prioritizer, the mean turnaround time of the entire WES pipeline, from DNA

extraction to clinical diagnosis, was 5.8 (SD 1.1) days using Variant Prioritizer. However, after implementation of AI Variant Prioritizer, the mean turnaround time was reduced to 4.8 (SD 1.2) days for rapid trio exome sequencing analysis in NTUH.

Acknowledgments

This work was funded by a grant from the Ministry of Science and Technology (110-2634-F-002-032-) of Taiwan. The analyses presented in this publication are based on the use of study data downloaded from the dbGaP website, under dbGaP accession numbers phs000744.v4.p2, phs001272.v1.p1, phs000971.v2.p1, phs000711.v6.p2, and phs001232.v3.p2. W-LH applied for the data with project name “Variant prioritization for rapid exome analysis of rare genetic disease” (project ID 20911). Data were downloaded from the FTP site of dbGaP after approval.

Authors' Contributions

Y-SH investigated the model and data feasibility, performed formal analysis, developed the software, visualized data, and wrote the initial manuscript. CH conceived the idea, curated the data, reviewed the manuscript, and advised the software development team. N-CL and W-LH conceived the idea, curated the patient data, and reviewed and edited the draft. Y-CC and I-CL edited, revised, and strengthened the manuscript. HW and Y-LL tested the data performance. FPL supervised the project progress and supported the project and managed the project and reviewed the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Allele frequency.
[DOCX File , 16 KB - bioinform_v3i1e37701_app1.docx]

Multimedia Appendix 2
Description of features used in this research.
[XLSX File (Microsoft Excel File), 13 KB - bioinform_v3i1e37701_app2.xlsx]

Multimedia Appendix 3
Target variants, HPO term, and keywords of test case. HPO: Human Phenotype Ontology.
[XLSX File (Microsoft Excel File), 24 KB - bioinform_v3i1e37701_app3.xlsx]

References

1. Behjati S, Tarpey PS. What is next generation sequencing? Arch Dis Child Educ Pract Ed 2013 Dec;98(6):236-238 [FREE Full text] [doi: 10.1136/archdischild-2013-304340] [Medline: 23986538]
2. O'Brien TD, Campbell NE, Potter AB, Letaw JH, Kulkarni A, Richards CS. Artificial intelligence (AI)-assisted exome reanalysis greatly aids in the identification of new positive cases and reduces analysis time in a clinical diagnostic laboratory. Genet Med 2022 Jan;24(1):192-200. [doi: 10.1016/j.gim.2021.09.007] [Medline: 34906498]
3. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 1977 Dec;74(12):5463-5467 [FREE Full text] [doi: 10.1073/pnas.74.12.5463] [Medline: 271968]
4. Hamosh A, Scott AF, Amberger J, Valle D, McKusick VA. Online Mendelian Inheritance in Man (OMIM). Hum Mutat 2000;15(1):57-61. [doi: 10.1002/(SICI)1098-1004(200001)15:1<57::AID-HUMU12>3.0.CO;2-G] [Medline: 10612823]
5. Adam MP, Everman DB, Mirzaa GM, Pagon RA, Wallace SE, Bean LJH, et al, editors. GeneReviews. Seattle, WA: University of Washington, Seattle; 1993.
6. Faintuch J, Faintuch S. Precision Medicine for Investigators, Practitioners and Providers. New York, NY: Academic Press; Nov 16, 2019.
7. Smedley D, Jacobsen JOB, Jäger M, Köhler S, Holtgrewe M, Schubach M, et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. Nat Protoc 2015 Dec;10(12):2004-2015 [FREE Full text] [doi: 10.1038/nprot.2015.124] [Medline: 26562621]
8. Boudellioua I, Kulmanov M, Schofield PN, Gkoutos GV, Hoehndorf R. DeepPVP: phenotype-based prioritization of causative variants using deep learning. BMC Bioinformatics 2019 Feb 06;20(1):65 [FREE Full text] [doi: 10.1186/s12859-019-2633-8] [Medline: 30727941]
9. Li Q, Zhao K, Bustamante CD, Ma X, Wong WH. Xrare: a machine learning method jointly modeling phenotypes and genetic evidence for rare disease diagnosis. Genet Med 2019 Sep;21(9):2126-2134 [FREE Full text] [doi: 10.1038/s41436-019-0439-8] [Medline: 30675030]

10. Holt JM, Wilk B, Birch CL, Brown DM, Gajapathy M, Moss AC, Undiagnosed Diseases Network, et al. VarSight: prioritizing clinically reported variants with binary classification algorithms. BMC Bioinformatics 2019 Oct 15;20(1):496 [FREE Full text] [doi: 10.1186/s12859-019-3026-8] [Medline: 31615419]
11. Yang H, Robinson PN, Wang K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. Nat Methods 2015 Sep;12(9):841-843 [FREE Full text] [doi: 10.1038/nmeth.3484] [Medline: 26192085]
12. De La Vega FM, Chowdhury S, Moore B, Frise E, McCarthy J, Hernandez EJ, et al. Artificial intelligence enables comprehensive genome interpretation and nomination of candidate diagnoses for rare genetic diseases. Genome Med 2021 Oct 14;13(1):153 [FREE Full text] [doi: 10.1186/s13073-021-00965-0] [Medline: 34645491]
13. Rentzsch P, Witten D, Cooper G, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res 2019 Jan 08;47(D1):D886-D894 [FREE Full text] [doi: 10.1093/nar/gky1016] [Medline: 30371827]
14. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet 2015 May 15;24(8):2125-2137 [FREE Full text] [doi: 10.1093/hmg/ddu733] [Medline: 25552646]
15. Singleton MV, Guthery SL, Voelkerding KV, Chen K, Kennedy B, Margraf RL, et al. Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. Am J Hum Genet 2014 Apr 03;94(4):599-610 [FREE Full text] [doi: 10.1016/j.ajhg.2014.03.010] [Medline: 24702956]
16. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods 2010 Apr;7(4):248-249 [FREE Full text] [doi: 10.1038/nmeth0410-248] [Medline: 20354512]
17. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 2003 Jul 01;31(13):3812-3814 [FREE Full text] [doi: 10.1093/nar/gkg509] [Medline: 12824425]
18. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods 2014 Apr;11(4):361-362. [doi: 10.1038/nmeth.2890] [Medline: 24681721]
19. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet 2008 Nov;83(5):610-615 [FREE Full text] [doi: 10.1016/j.ajhg.2008.09.017] [Medline: 18950739]
20. Hsu C. An integrated genetic variation analysis system for gene diagnostics in precision medicine (Master's thesis). NDLTD. Taipei City, Taiwan: National Taiwan University; 2018. URL: https://hdl.handle.net/11296/v9rcd8 [accessed 2022-08-31]
21. ChenT-F. Variants Prioritizer for Exome Data Based on Text-mining. NTU Thesis and Dissertations Repository. Taipei City, Taiwan: National Taiwan University; 2018. URL: https://tdr.lib.ntu.edu.tw/handle/123456789/17687?mode=full [accessed 2022-08-31]
22. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010 Sep;38(16):e164 [FREE Full text] [doi: 10.1093/nar/gkq603] [Medline: 20601685]
23. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol 2016 Jun 06;17(1):122 [FREE Full text] [doi: 10.1186/s13059-016-0974-4] [Medline: 27268795]
24. Stromberg M, Roy R, Lajugie J, Jiang Y, Li H, Margulies E. Nirvana: Clinical Grade Variant Annotator. New York, NY: Association for Computing Machinery; 2017 Presented at: The 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics; August 20-23, 2017; Boston, MA. [doi: 10.1145/3107411.3108204]
25. Li Q, Wang K. InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. Am J Hum Genet 2017 Feb 02;100(2):267-280 [FREE Full text] [doi: 10.1016/j.ajhg.2017.01.004] [Medline: 28132688]
26. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 2014 Jan;42(Database issue):D980-D985 [FREE Full text] [doi: 10.1093/nar/gkt1113] [Medline: 24234437]
27. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NST, et al. Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat 2003 Jul;21(6):577-581. [doi: 10.1002/humu.10212] [Medline: 12754702]
28. Fan C, Lin J, Lee C. Taiwan Biobank: a project aiming to aid Taiwan's transition into a biomedical island. Pharmacogenomics 2008 Feb;9(2):235-246. [doi: 10.2217/14622416.9.2.235] [Medline: 18370851]
29. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 2015 May;17(5):405-424 [FREE Full text] [doi: 10.1038/gim.2015.30] [Medline: 25741868]
30. Glassberg EC, Gao Z, Harpak A, Lant X, Pritchard JK. Measurement of selective constraint on human gene expression. bioRxiv 2022. [doi: 10.1101/345801]
31. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001:17-21 [FREE Full text] [Medline: 11825149]
32. Aronson AR, Lang F. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010;17(3):229-236 [FREE Full text] [doi: 10.1136/jamia.2009.002733] [Medline: 20442139]

33. Lindberg DAB, Humphreys BL, McCray AT. The Unified Medical Language System. Yearb Med Inform 2018 Mar 05;02(01):41-51. [doi: 10.1055/s-0038-1637976]
34. Liu C, Peres Kury FS, Li Z, Ta C, Wang K, Weng C. Doc2Hpo: a web application for efficient and accurate HPO concept curation. Nucleic Acids Res 2019 Jul 02;47(W1):W566-W570 [FREE Full text] [doi: 10.1093/nar/gkz386] [Medline: 31106327]
35. Tchechmedjiev A, Abdaoui A, Emonet V, Melzi S, Jonnagaddala J, Jonquet C. Enhanced functionalities for annotating and indexing clinical text with the NCBO Annotator. Bioinformatics 2018 Jun 01;34(11):1962-1965 [FREE Full text] [doi: 10.1093/bioinformatics/bty009] [Medline: 29846492]
36. Lee L. IDF revisited: A simple new derivation within the Robertson-Spärck Jones probabilistic model. New York, NY: ACM; 2007 Jul Presented at: SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information; July 23-27, 2007; Amsterdam, The Netherlands p. 751-752. [doi: 10.1145/1277741.1277891]
37. Robertson S, Walker S, Beaulieu MM. Okapi at TREC-7: automatic ad hoc, filtering, VCL and interactive track. Microsoft. Gaithersburg, MD: National Institute of Standards and Technology; 1999 Jan. URL: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/okapi_trec7.pdf [accessed 2022-08-31]
38. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 2005 Jan 01;33(Database issue):D54-D58 [FREE Full text] [doi: 10.1093/nar/gki031] [Medline: 15608257]
39. Wei C, Kao H, 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]
40. Hintzsche JD, Robinson WA, Tan AC. A Survey of Computational Tools to Analyze and Interpret Whole Exome Sequencing Data. Int J Genomics 2016;2016:7983236 [FREE Full text] [doi: 10.1155/2016/7983236] [Medline: 28070503]
41. Doniger SW, Kim HS, Swain D, Corcuera D, Williams M, Yang S, et al. A catalog of neutral and deleterious polymorphism in yeast. PLoS Genet 2008 Aug 29;4(8):e1000183 [FREE Full text] [doi: 10.1371/journal.pgen.1000183] [Medline: 18769710]
42. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res 2011 Sep 01;39(17):e118 [FREE Full text] [doi: 10.1093/nar/gkr407] [Medline: 21727090]
43. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat 2013 Jan;34(1):57-65 [FREE Full text] [doi: 10.1002/humu.22225] [Medline: 23033316]
44. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS One 2012;7(10):e46688 [FREE Full text] [doi: 10.1371/journal.pone.0046688] [Medline: 23056405]
45. Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet 2016 Oct 24;48(12):1581-1586. [doi: 10.1038/ng.3703]
46. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol 2010 Dec 02;6(12):e1001025 [FREE Full text] [doi: 10.1371/journal.pcbi.1001025] [Medline: 21152010]
47. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics 2015 Mar 01;31(5):761-763 [FREE Full text] [doi: 10.1093/bioinformatics/btu703] [Medline: 25338716]
48. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day INM, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics 2015 May 15;31(10):1536-1543 [FREE Full text] [doi: 10.1093/bioinformatics/btv009] [Medline: 25583119]
49. Lu Q, Hu Y, Sun J, Cheng Y, Cheung K, Zhao H. A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. Sci Rep 2015 May 27;5(1):10576-10513 [FREE Full text] [doi: 10.1038/srep10576] [Medline: 26015273]
50. Gulko B, Hubisz MJ, Gronau I, Siepel A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. Nat Genet 2015 Mar;47(3):276-283 [FREE Full text] [doi: 10.1038/ng.3196] [Medline: 25599402]
51. Hubisz MJ, Pollard KS, Siepel A. PHAST and RPHAST: phylogenetic analysis with space/time models. Brief Bioinform 2011 Jan;12(1):41-51 [FREE Full text] [doi: 10.1093/bib/bbq072] [Medline: 21278375]
52. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 2005 Aug 15;15(8):1034-1050 [FREE Full text] [doi: 10.1101/gr.3715005] [Medline: 16024819]
53. Garber M, Guttman M, Clamp M, Zody M, Friedman N, Xie X. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics 2009 Jul 15;25(12):i54-i62 [FREE Full text] [doi: 10.1093/bioinformatics/btp190] [Medline: 19478016]
54. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Hum Mutat 2016 Mar;37(3):235-241 [FREE Full text] [doi: 10.1002/humu.22932] [Medline: 26555599]
55. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O. Scikit-learn: machine learning in python. Journal of Machine Learning Research 2011;12:2825-2830 [FREE Full text]

56. Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information. Phys. Rev. E 2004 Jun 23;69(6):066138-1-066138-16. [doi: 10.1103/physreve.69.066138]
57. Ross BC. Mutual information between discrete and continuous data sets. PLoS One 2014 Feb 19;9(2):e87357 [FREE Full text] [doi: 10.1371/journal.pone.0087357] [Medline: 24586270]
58. Breiman L. Random forests. Machine Learning 2001;45:5-32 [FREE Full text] [doi: 10.1023/A:1010933404324]
59. Breiman L. Consistency for a simple model of random forests. University of California, Berkeley. 2004. URL: https://www.stat.berkeley.edu/~breiman/RandomForests/consistencyRFA.pdf [accessed 2022-08-31]
60. Ellerman D. Logical Entropy: Introduction to Classical and Quantum Logical Information Theory. Entropy (Basel) 2018 Oct 06;20(9):679 [FREE Full text] [doi: 10.3390/e20090679] [Medline: 33265768]
61. Birgmeier J, Haeussler M, Deisseroth CA, Jagadeesh KA, Ratner AJ, Guturu H, et al. AMELIE accelerates Mendelian patient diagnosis directly from the primary literature. bioRxiv Preprint posted online on August 02, 2017. [doi: 10.1101/171322]
62. Stelzer G, Plaschkes I, Oz-Levi D, Alkelai A, Olender T, Zimmerman S, et al. VarElect: the phenotype-based variation prioritizer of the GeneCards Suite. BMC Genomics 2016 Jun 23;17 Suppl 2(S2):444-206 [FREE Full text] [doi: 10.1186/s12864-016-2722-2] [Medline: 27357693]
63. Clark MM, Hildreth A, Batalov S, Ding Y, Chowdhury S, Watkins K, et al. Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation. Sci Transl Med 2019 Apr 24;11(489):eaat6177. [doi: 10.1126/scitranslmed.aat6177] [Medline: 31019026]
64. Smedley D, Oellrich A, Köhler S, Ruef B, Sanger Mouse Genetics Project, Westerfield M, et al. PhenoDigm: analyzing curated annotations to associate animal models with human diseases. Database (Oxford) 2013;2013:bat025 [FREE Full text] [doi: 10.1093/database/bat025] [Medline: 23660285]
65. Washington NL, Haendel MA, Mungall CJ, Ashburner M, Westerfield M, Lewis SE. Linking human diseases to animal models using ontology-based phenotype annotation. PLoS Biol 2009 Dec 24;7(11):e1000247 [FREE Full text] [doi: 10.1371/journal.pbio.1000247] [Medline: 19956802]
66. Ayadi A, Birling M, Bottomley J, Bussell J, Fuchs H, Fray M, et al. Mouse large-scale phenotyping initiatives: overview of the European Mouse Disease Clinic (EUMODIC) and of the Wellcome Trust Sanger Institute Mouse Genetics Project. Mamm Genome 2012 Oct;23(9-10):600-610 [FREE Full text] [doi: 10.1007/s00335-012-9418-y] [Medline: 22961258]
67. Bult CJ, Eppig JT, Blake JA, Kadin JA, Richardson JE, Mouse Genome Database Group. The mouse genome database: genotypes, phenotypes, and models of human disease. Nucleic Acids Res 2013 Jan;41(Database issue):D885-D891 [FREE Full text] [doi: 10.1093/nar/gks1115] [Medline: 23175610]
68. Howe DG, Bradford YM, Conlin T, Eagle AE, Fashena D, Frazer K, et al. ZFIN, the Zebrafish Model Organism Database: increased support for mutants and transgenics. Nucleic Acids Res 2013 Jan;41(Database issue):D854-D860 [FREE Full text] [doi: 10.1093/nar/gks938] [Medline: 23074187]

Abbreviations

ACMG: American College of Medical Genetics and Genomics
AI: artificial intelligence
AMP: Association for Molecular Pathology
API: application programming interface
BERT: Bidirectional Encoder Representations from Transformers
CNV: copy number variation
EMR: electronic medical record
GBDT: gradient boosting decision tree
gnomAD: Genome Aggregation Database
GO: Gene Ontology
HGMD: Human Genome Mutation Database
HIPAA: Health Insurance Portability and Accountability Act
HPO: Human Phenotype Ontology
IRB: institutional review board
MViewer: Mutation Viewer
NGS: next-generation sequencing
NLP: natural language processing
NTUH: National Taiwan University Hospital
OMIM: Online Mendelian Inheritance in Man
SNV: single-nucleotide variant
SVM: support vector machine
UMLS: Unified Medical Language System
VCF: variant call format

VEP: Variant Effect Predictor

Edited by A Mavragani; submitted 04.03.22; peer-reviewed by YF Cheng, N Pontikos, C Liuu; comments to author 09.05.22; revised version received 29.07.22; accepted 22.08.22; published 15.09.22.

Please cite as:
Huang YS, Hsu C, Chune YC, Liao IC, Wang H, Lin YL, Hwu WL, Lee NC, Lai F
Diagnosis of a Single-Nucleotide Variant in Whole-Exome Sequencing Data for Patients With Inherited Diseases: Machine Learning Study Using Artificial Intelligence Variant Prioritization
JMIR Bioinform Biotech 2022;3(1):e37701
URL: https://bioinform.jmir.org/2022/1/e37701
doi:10.2196/37701
PMID:

Original Paper

A Bioinformatics Tool for Predicting Future COVID-19 Waves Based on a Retrospective Analysis of the Second Wave in India: Model Development Study

Ashutosh Kumar[1], MD; Adil Asghar[1], MD; Prakhar Dwivedi[1], MBBS; Gopichand Kumar[1], MBBS; Ravi K Narayan[2], MD; Rakesh K Jha[1], MD; Rakesh Parashar[3], MHA, PhD; Chetan Sahni[4], MD; Sada N Pandey[5], PhD

[1]Department of Anatomy, All India Institute of Medical Sciences - Patna, Patna, India

[2]Department of Anatomy, Dr B C Roy Multispeciality Medical Research Center, Indian Institute of Technology-Kharagpur, Kharagpur, India

[3]India Health Lead, Oxford Policy Management Limited, Oxford, United Kingdom

[4]Department of Anatomy, Institute of Medical Sciences, Banaras Hindu University, Varanasi, India

[5]Department of Zoology, Banaras Hindu University, Varanasi, India

Corresponding Author:
Ashutosh Kumar, MD
Department of Anatomy
All India Institute of Medical Sciences - Patna
Phulwari Sharif
Patna, 801507
India
Phone: 91 612 245 ext 1335
Email: drashutoshkumar@aiimspatna.org

Abstract

Background: Since the start of the COVID-19 pandemic, health policymakers globally have been attempting to predict an impending wave of COVID-19. India experienced a devastating second wave of COVID-19 in the late first week of May 2021. We retrospectively analyzed the viral genomic sequences and epidemiological data reflecting the emergence and spread of the second wave of COVID-19 in India to construct a prediction model.

Objective: We aimed to develop a bioinformatics tool that can predict an impending COVID-19 wave.

Methods: We analyzed the time series distribution of genomic sequence data for SARS-CoV-2 and correlated it with epidemiological data for new cases and deaths for the corresponding period of the second wave. In addition, we analyzed the phylodynamics of circulating SARS-CoV-2 variants in the Indian population during the study period.

Results: Our prediction analysis showed that the first signs of the arrival of the second wave could be seen by the end of January 2021, about 2 months before its peak in May 2021. By the end of March 2021, it was distinct. B.1.617 lineage variants powered the wave, most notably B.1.617.2 (Delta variant).

Conclusions: Based on the observations of this study, we propose that genomic surveillance of SARS-CoV-2 variants, complemented with epidemiological data, can be a promising tool to predict impending COVID-19 waves.

(JMIR Bioinform Biotech 2022;3(1):e36860) doi:10.2196/36860

KEYWORDS

COVID-19; epidemiology; genomic surveillance; second wave; SARS-CoV-2

Introduction

The year 2019 had a SARS-CoV-2–driven wave of COVID worldwide that soon turned into a pandemic, and to date, this disease has killed about 65 million people [1]. Since the pandemic's start, much policy talk has been about whether an impending COVID wave can be predicted [2]. Unfortunately, successful prediction of COVID waves has not yet been achieved. A prediction tool that can inform about an upcoming COVID wave well before time and reasonably accurately could minimize the enormous loss of life and other collateral damages.

Multiple waves at a global scale driven by SARS-CoV-2 variants, primarily Alpha, Delta [3], and, most recently, Omicron [4], have followed since the first wave. The successive

SARS-CoV-2 variants showed increased transmissibility and virulence compared with the wild-type strain [3]; however, the latest Omicron variant has shown higher transmissibility and immune escape but lesser lethality compared with the Delta variant [4]. The Delta variant–driven wave was characterized by high speed of rising cases, increased oxygen demand, vaccine breakthrough [5], a highly increased proportion of severe cases, and high mortality [6].

More comprehensive coverage of COVID vaccines in the global population is helping to create an immunity barrier against the rise of a new wave. However, an increase in the immune escape potential of emerging variants causes a grave concern for vaccine breakthroughs and reinfections [3,4,7]. With the waning of immunity derived from vaccines and previous infections [8], the risk of the emergence of a more lethal variant capable of creating a global wave remains high and therefore demands continued surveillance [9].

The Delta variant–driven wave showed a rapid peak and fall to the baseline, making it ideal for prediction studies. The Delta strain was first reported from India [10]. Of note, India witnessed a devastating second COVID wave that began toward the end of February 2021 [11]. The unexpected arrival of the second COVID wave, accompanied by an exponential increase in infections, brought the country's epidemic response system and health infrastructure to a standstill [11], and resulted in massive suffering and loss of life [12].

The Delta variant belongs to the SARS-CoV-2 lineage B.1.617, which appeared as a precursor. The first case of the B.1.617 variant was also reported from India as early as October 2020 [13]. The World Health Organization (WHO) recognized the B.1.617 lineage as a global variant of concern (VOC). The strain evolved into 3 more sublines, namely, B.1.617.1-3, of which B.1.617.1 (the Kappa variant) was declared a variant of interest (VOI) and B.1.617.2 was later declared a VOC by the WHO [14]. B.1.617 contained mutations in key spike protein regions involved in host interactions and the induction of neutralizing antibodies (S: L452R, E484Q, D614G, del681, and del1072) [15]. The sublineages contained lineage-defining spike mutations (L452R and D614G) as well as newly developed mutations as follows: B.1.617.1 (S: T95I, G142D, E154K, L452R, E484Q, D614G, P681R, and Q1071H); B.1.617.2 (S: T19R, G142D, 156del, 157del, R158G, L452R, T478K, D614G, P681R, D614G, P681R, and D950N); and B.1.617.3 (S: T19R, L452R, E484Q, D614G, and P681R) [16]. Contemporary studies suggested that B.1.617 lineage variants were more easily transmissible [13,17-21] and deadlier [18] than the B.1.1.7 lineage (Alpha variant), a globally dominant strain before the second wave [10]. Studies also showed a significant reduction in the neutralization of variants of the B.1.617 lineage by antibodies derived from natural infections and many currently used COVID-19 vaccines, and multiple monoclonal antibodies [18-21]. Notably, B.1.617.2 showed very high transmissibility and immunological escape [10,13,17,22].

Several studies worldwide have shown that predicting an impending COVID-19 wave is possible [23-28]. These studies used mathematical modeling of epidemiological data. Unfortunately, none of them could accurately anticipate a COVID-19 wave. The ability to predict an established wave from epidemiological data alone seems severely limited [12,29].

The analysis of SARS-CoV-2 genomic sequences has emerged as an efficient surveillance tool for understanding the emergence of new variants and their spread. Fortunately, millions of SARS-CoV-2 genomic sequences from regions worldwide are being made publicly available as a collaborative effort to contain the pandemic [30]. The easy availability of high-quality viral sequences with patient metadata has opened a new avenue for potential predictions of the COVID-19 pandemic [31]. However, viral genomic sequences alone may not be sufficient for efficient predictions, and their current uses for this purpose are constrained.

In this study, we propose an integrated approach using viral genome surveillance and epidemiological data for the prediction of an impending COVID-19 wave. We retrospectively analyzed viral genomic sequences and epidemiological data reflecting the emergence and spread of the second wave of COVID-19 in India to construct such a model.

Methods

Study Design, Participants, and Data Sources

We analyzed the time series (weekly and monthly) distributions of SARS-CoV-2 variants coupled with epidemiological data from December 1, 2020, to July 26, 2021 (34 weeks) for new cases and deaths from COVID-19 in India. Further, a phylodynamic analysis for individual variants was performed.

We downloaded SARS-CoV-2 genomic sequence data and epidemiological data from the EpiCoV database of the Global Initiative on Sharing All Influenza Data (GISAID) [32] and the Worldometer database [33], respectively. A total of 40,359 genomic sequences of SARS-CoV-2 were analyzed. The sequence for each SARS-CoV-2 variant was retrieved using an automated search function that entered lineage and sublineage information into the EpiCoV database. The total numbers of sequences per week and month for the variants and their relative proportions were calculated (in percentage). The data were tabulated, and each variant's weekly and monthly distributions were compared to COVID-19 epidemiological data (new cases and deaths) and statistically analyzed. The genomic sequences of SARS-CoV-2 variants in each state and union territory were also examined to check deviations from overall patterns in data.

Phylodynamics of SARS-CoV-2 Variants

A phylodynamic analysis of the variants circulating in the Indian population during the study period was performed on GISAID sequences using the bioinformatics tool available at EpiCoV.

Statistical Analysis

XLSTAT (Addinsoft) was used to perform all statistical analyses. Descriptive statistics were calculated for each variable. Levene and Anderson tests were used to determine the homogeneity or normality of the data. In addition, a correlation matrix was constructed, and a linear regression analysis was performed between contrasting variables (R values = −1 to +1). Finally, the statistical significance level for each comparison was set at $P<.05$.

Ethical Considerations

Approval from the institutional ethics committee was not required as the data used in this study were retrieved from publicly available databases.

Results

Our retrospective analysis of the epidemiological data reflected that the second COVID-19 wave started rising by the end of February 2021 and peaked by the end of the first week of May 2021. Based on the distinct epidemiological trends observed (Multimedia Appendix 1), we divided the study period (December 1, 2020, to July 26, 2021; 34 weeks) into prepeak (weeks 1-23) and postpeak (weeks 24-34) periods. The weekly average of new cases and deaths showed a strong correlation in the study period (R=0.98, $P<.001$), signifying the high statistical validity of the data for further comparisons. Further, we analyzed the distribution of SARS-CoV-2 variants circulating in the Indian population in correlation with new cases and deaths before and after the peak. For description, based on epidemiological trends, the prepeak period was further divided into the following 3 time series intervals: “very early” (weeks 1-8), “early” (weeks 9-16), and “near peak” (weeks 17-23). New cases and deaths showed a downward trend in the “very early” period and maintained a plateau in the “early” period (except toward the end when cases and deaths started increasing, indicating the start of the second wave). In the “near peak” period, a steep rise in new cases and deaths was observed (Figure 1).

The rise and fall of circulating SARS-CoV-2 variants were studied against the observed epidemiological data trends in the respective time series intervals. Observing the composite data trends of epidemiological and SARS-CoV-2 genomic data provides a glimpse of the formation of the second COVID-19 wave, with clear indications of which SARS-CoV-2 strains may have driven it (Figures 1 and 2). By December 2020, 8 SARS-CoV-2 Pango lineages and their multiple sublineages were circulating in the Indian population, including 4 VOCs (B.1.1.7, B.1.351, P1, and B.1.617.2) and 3 VOIs (B.1.617.1, B.1.127/B.1.429, and B.1.525). However, B.1.1.7 was the most dominant variant in that period. B.1.617 lineage variants collectively (B.1.617+) showed an upward trend since their emergence, and surpassed other VOCs, including B.1.1.7, by the end of January 2021 (weeks 8-9) and subsequently kept rising. In contrast, B.1.1.7 showed a downward trend by the end of March 2021 (weeks 17-18), with B.1.617 lineage variants becoming the dominant variants. By the end of April 2021, B.1.617 lineage variants were detected in 78.5% of SARS-CoV-2 sequences uploaded on the GISAID database, reaching about 83% in the week of the peak.

The phylodynamic analysis of the circulating variants in the study period strongly corroborated the trends present in the graph data, showing an exclusive increase in the cluster density of B.1.617.2 compared with other variants in the “near peak” period (Figure 3).

To know whether the rise in the B.1.617.2 variant was localized to specific geographical regions, which may have influenced the collective data trends, we compared the monthly distribution of genomic sequences of SARS-CoV-2 variants for the states and union territories of India individually. A similar increase in the detection of the B.1.617.2 variant was observed in most states and union territories (Multimedia Appendix 2), except Kerala, where different patterns were visible (Figure S15 in Multimedia Appendix 2). In Kerala, the rise of the B.1.617.2 variant was slower in comparison with the rest of the country (55.5% vs 72% of total cases by the end of April 2021), which was further confirmed in the state-wise serosurvey data from the period of the second wave (44.4% vs 67.7% of the national average) [34]. Notably, a sharp rise in B.1.617.2 cases was observed in Kerala in a later period.

Figure 1. Weekly distribution of SARS-CoV-2 variants in genomic sequence data from India and the correlation with daily new COVID-19 cases and deaths from December 1, 2020, to July 26, 2021. The data were analyzed for the period before the peak of the second wave (23rd week) and after that. SARS-CoV-2 genomic sequence data were obtained from the EpiCoV database of the Global Initiative on Sharing All Influenza Data, and epidemiological data were obtained from the Worldometer database.

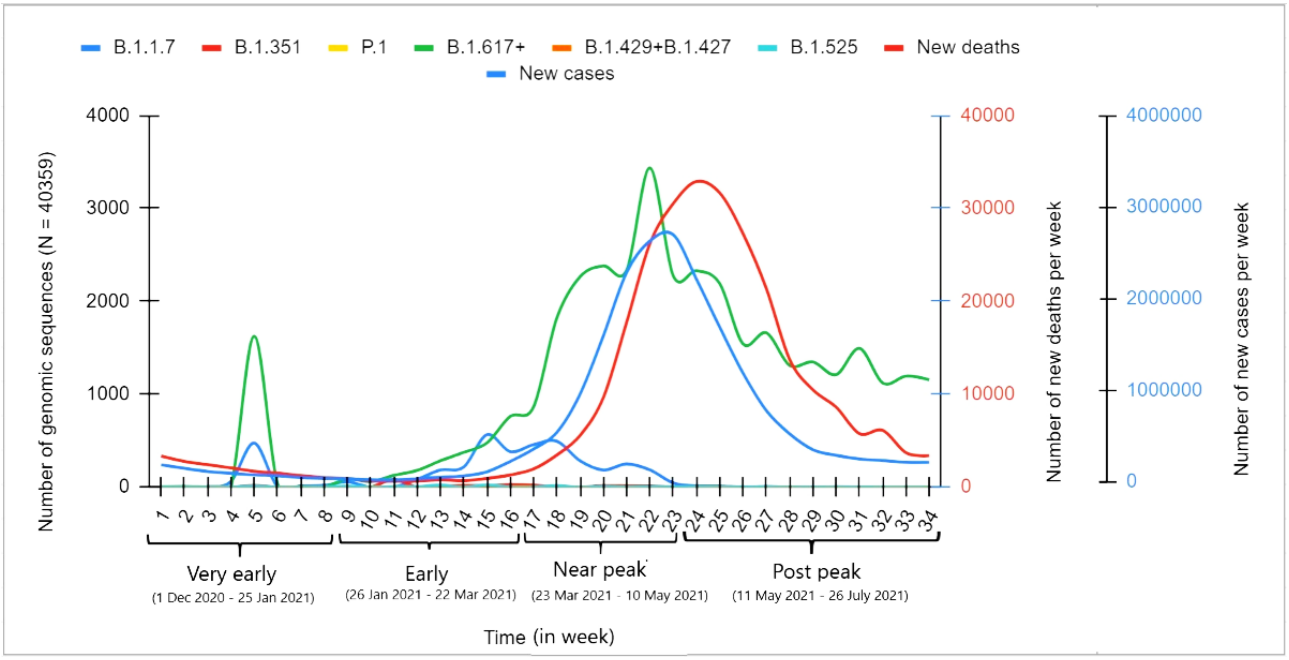


Figure 2. Origin and spread of B.1.617 lineage SARS-CoV-2 variants in the Indian population. Data were analyzed from December 1, 2020, to July 26, 2021. SARS-CoV-2 genomic sequence data were obtained from the EpiCoV database of the Global Initiative on Sharing All Influenza Data, and epidemiological data were obtained from the Worldometer database.

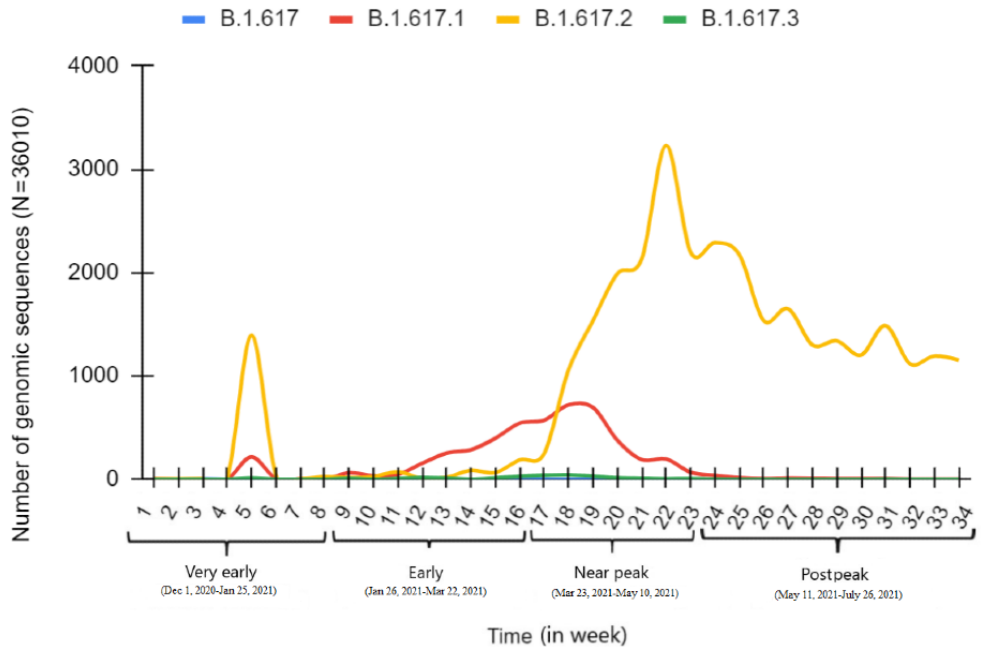
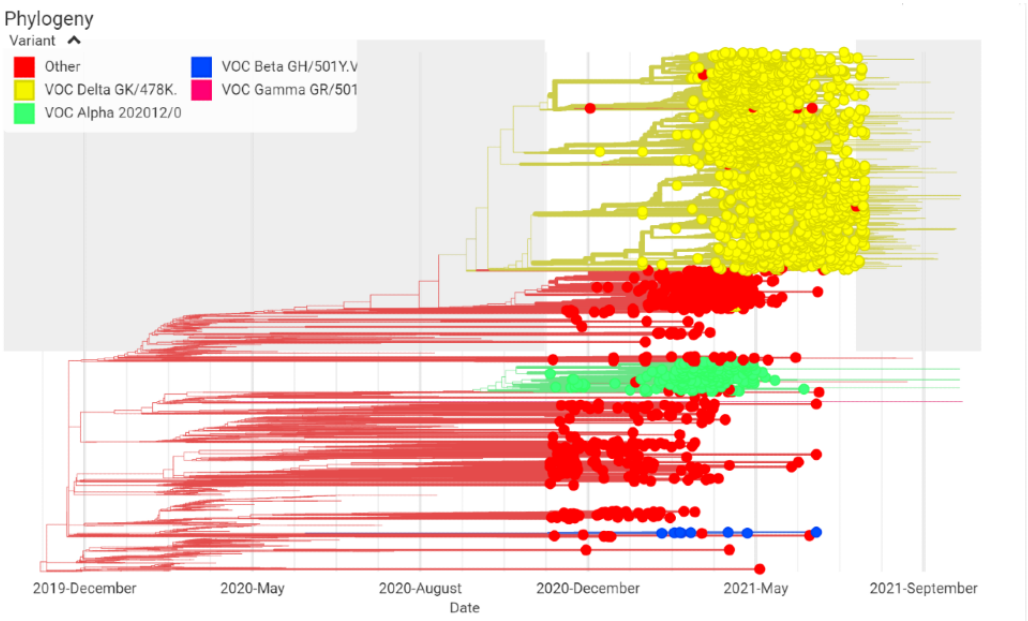


Figure 3. Phylodynamics of SARS-CoV-2 variants in the Indian population from December 1, 2020, to July 26, 2021. SARS-CoV-2 genomic sequence data were obtained from the EpiCoV database of the Global Initiative on Sharing All Influenza Data, and epidemiological data were obtained from the Worldometer database. VOC: variant of concern.



Discussion

Principal Findings

The retrospective examination of linked viral genomic sequences and epidemiological data in this study clearly showed that the occurrence of B.1.617 lineage variants, particularly the B.1.617.2 sublineage, was strongly related to the second wave of COVID-19 in India. In late January 2021, when instances of B.1.617.2 surpassed those of all other variants, the first signs of an imminent wave of COVID-19 began to appear. The rise of the wave could be observed closely until the end of March 2021, when instances of B.1.617.2 showed a sharp increase in line with the total number of new cases.

Comparison With Prior Work

Current prediction models in the COVID-19 pandemic are dominated by purely epidemiological analyses, from which hardly anyone could accurately predict an impending COVID-19 wave [23-27]. The importance of studying viral genomic sequences for the epidemiological surveillance of new SARS-CoV-2 variants is well recognized [31,35-40]. However, its application in developing a predictive model to forecast upcoming virus waves has received little appreciation in the existing literature [41]. Interestingly, strong conceptual validation for the applicability of an integrated approach to predict an impending COVID-19 wave using viral genomic surveillance and epidemiological data came from a recent study by de Hoffer et al [42]. These authors studied the temporal dynamics of emerging SARS-CoV-2 variants using a machine learning algorithm–based analysis of the spike protein sequences of viral samples from England, Scotland, and Wales reported in the GISAID database. Further, they correlated the relative percentage of each variant with the weekly and monthly epidemiological data of active cases from the studied geographical regions. They showed a strong relationship between the genesis of a new emerging variant and the onset of a new wave, with an exponential increase in the number of infections [42].

Moreover, our findings regarding the second wave of COVID-19 in India are corroborated by a previous study by Dhar et al [10]. The authors analyzed viral genomic sequences retrospectively and observed a similar pattern in the rise of the B.1.617 lineage, mainly the B.1.617.2 variant, in Delhi before the second wave [10]. A B.1.617.2-driven second wave was also reflected in the analysis of viral genomic sequences performed by Adiga and Nayak in 2021 [43]. We recently used our prediction model prospectively during the initial rise of cases caused by the Omicron strain in South Africa, which indicated an upcoming wave with very high transmissibility but limited lethality [4]. These predictions were later accurately reflected in the studies reporting the Omicron-mediated fourth wave of COVID-19 in South Africa [44,45].

The potential predictability of the second wave of COVID-19 in India in the retrospective data analysis suggests that genomic surveillance of SARS-CoV-2 variants, enriched with epidemiological data, could be a potential tool to predict upcoming COVID-19 waves. Still, the prediction accuracy is largely dependent on population-based viral genomic sequencing and consistency in data upload from all geographic regions, as well as accurate reporting of epidemiological data. The sole increase in the proportion of an emerging SARS-CoV-2 variant, coupled with an associated rise in new cases, might inform the arrival of a new wave of COVID-19. However, consideration of other epidemiological factors, such as previous exposure to related virus strains and the immunization status of the population, will be necessary to determine the magnitude of an impending wave [46]. Notably, the first wave of COVID-19 in India was limited in scope, as evidenced by the serosurvey data [47,48], and only a small part of the population was vaccinated as of early 2021 [49]. With the emergence of a new variant, both these factors may have created an ideal environment for a

massive second wave to emerge. In addition, preventive measures, such as blocking or limiting gatherings and using face masks, can also influence the prospects and magnitude of a new wave [29].

Limitations

There were some limitations in our study that may have influenced the interpretation of the results. First, the samples used in our analyses might not be representative of the population. In many geographical regions, the sample size was grossly disproportionate. Therefore, the genomic sequence data presented in this study might not reflect the exact epidemiological extent of the distribution of the variants in the reported geographical regions but only show their relative proportions in the samples for which genomic sequences were uploaded to the GISAID database. We have assumed that similar proportions exist between variants in the actual population. Second, inconsistent reports and uploads of genomic sequences made it challenging to study a daily trend in the spread of variants. Finally, the scarcity of genomic sequences and inconsistency in uploading to the databases used for some states/union territories made determining variant dominance difficult.

Conclusions

Based on the observations of this study, we propose that genomic surveillance of SARS-CoV-2 variants, complemented with epidemiological data, can be a promising tool to predict upcoming COVID-19 waves.

Acknowledgments

The study used SARS-CoV-2 genomic sequence data and epidemiological data from the EpiCoV database of the Global Initiative on Sharing All Influenza Data and the Worldometer database, respectively.

Data Availability

The primary data used for this study are publicly available on the EpiCoV database of the Global Initiative on Sharing All Influenza Data (SARS-CoV-2 genomic sequence data) and the Worldometer database (epidemiological data). The categorized data for the study period can be availed from the corresponding author on reasonable request.

Authors' Contributions

AK, PD, and GK collected the samples and analyzed the data. AK wrote the first draft. AA performed the statistical analysis. AK, RKN, RKJ, RP, CS, and SNP reviewed and edited the paper. All authors consented to submit the final draft.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Weekly new COVID-19 cases and deaths in the Indian population for the period of December 1, 2020, to July 26, 2021.
[DOCX File , 225 KB - bioinform_v3i1e36860_app1.docx]

Multimedia Appendix 2
Monthly distribution of SARS-CoV-2 variants in genomic sequence data from the states and union territories of India uploaded on the EpiCoV database for the period of December 1, 2020, to July 26, 2021.
[DOCX File , 678 KB - bioinform_v3i1e36860_app2.docx]

References

1. Coronavirus Statistics. Worldometers. URL: https://www.worldometers.info/coronavirus/ [accessed 2022-09-05]
2. O'Brien DA, Clements C. Early warning signal reliability varies with COVID-19 waves. Biol Lett 2021 Dec;17(12):20210487 [FREE Full text] [doi: 10.1098/rsbl.2021.0487] [Medline: 34875183]
3. Kumar A, Parashar R, Kumar S, Faiq MA, Kumari C, Kulandhasamy M, et al. Emerging SARS-CoV-2 variants can potentially break set epidemiological barriers in COVID-19. J Med Virol 2022 Apr;94(4):1300-1314 [FREE Full text] [doi: 10.1002/jmv.27467] [Medline: 34811761]
4. Kumar A, Asghar A, Singh H, Faiq M, Kumar S, Narayan R, et al. An in silico analysis of early SARS-CoV-2 variant B.1.1.529 (Omicron) genomic sequences and their epidemiological correlates. medRxiv. 2021. URL: https://www.medrxiv.org/content/10.1101/2021.12.18.21267908v1 [accessed 2022-09-16]
5. Tareq A, Emran TB, Dhama K, Dhawan M, Tallei T. Impact of SARS-CoV-2 delta variant (B.1.617.2) in surging second wave of COVID-19 and efficacy of vaccines in tackling the ongoing pandemic. Hum Vaccin Immunother 2021 Nov 02;17(11):4126-4127 [FREE Full text] [doi: 10.1080/21645515.2021.1963601] [Medline: 34473593]
6. Ong S, Chiew C, Ang L, Mak T, Cui L, Toh M, et al. Clinical and Virological Features of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Variants of Concern: A Retrospective Cohort Study Comparing B.1.1.7 (Alpha),

B.1.351 (Beta), and B.1.617.2 (Delta). Clin Infect Dis 2022 Aug 24;75(1):e1128-e1136 [FREE Full text] [doi: 10.1093/cid/ciab721] [Medline: 34423834]
7. Guo Y, Han J, Zhang Y, He J, Yu W, Zhang X, et al. SARS-CoV-2 Omicron Variant: Epidemiological Features, Biological Characteristics, and Clinical Significance. Front Immunol 2022;13:877101 [FREE Full text] [doi: 10.3389/fimmu.2022.877101] [Medline: 35572518]
8. Ferdinands JM, Rao S, Dixon BE, Mitchell PK, DeSilva MB, Irving SA, et al. Waning 2-Dose and 3-Dose Effectiveness of mRNA Vaccines Against COVID-19-Associated Emergency Department and Urgent Care Encounters and Hospitalizations Among Adults During Periods of Delta and Omicron Variant Predominance - VISION Network, 10 States, August 2021-January 2022. MMWR Morb Mortal Wkly Rep 2022 Feb 18;71(7):255-263 [FREE Full text] [doi: 10.15585/mmwr.mm7107e2] [Medline: 35176007]
9. Rapidly escalating COVID-19 cases amid reduced virus surveillance forecasts a challenging autumn and winter in the WHO European Region. World Health Organization. 2022. URL: https://www.who.int/europe/news/item/19-07-2022-rapidly-escalating-covid-19-cases-amid-reduced-virus-surveillance-forecasts-a-challenging-autumn-and-winter-in-the-who-european-region [accessed 2022-09-05]
10. Dhar MS, Marwal R, Vs R, Ponnusamy K, Jolly B, Bhoyar RC, Indian SARS-CoV-2 Genomics Consortium (INSACOG), et al. Genomic characterization and epidemiology of an emerging SARS-CoV-2 variant in Delhi, India. Science 2021 Nov 19;374(6570):995-999 [FREE Full text] [doi: 10.1126/science.abj9932] [Medline: 34648303]
11. Samarasekera U. India grapples with second wave of COVID-19. The Lancet Microbe 2021 Jun;2(6):e238. [doi: 10.1016/s2666-5247(21)00123-3]
12. Jha P, Deshmukh Y, Tumbe C, Suraweera W, Bhowmick A, Sharma S, et al. COVID mortality in India: National survey data and health facility deaths. Science 2022 Feb 11;375(6581):667-671. [doi: 10.1126/science.abm5154] [Medline: 34990216]
13. Planas D, Veyer D, Baidaliuk A, Staropoli I, Guivel-Benhassine F, Rajah M, et al. Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. Nature 2021 Aug;596(7871):276-280. [doi: 10.1038/s41586-021-03777-9] [Medline: 34237773]
14. Tracking SARS-CoV-2 variants. World Health Organization. URL: https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/ [accessed 2021-06-17]
15. B.1.617.2) Lineage Report. Outbreak.info. URL: https://outbreak.info/situation-reports?pango=B.1.617.2) [accessed 2022-09-09]
16. Weekly epidemiological update on COVID-19 - 18 May 2021. World Health Organization. URL: https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---18-may-2021 [accessed 2022-09-16]
17. Lopez Bernal J, Andrews N, Gower C, Gallagher E, Simmons R, Thelwall S, et al. Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant. N Engl J Med 2021 Aug 12;385(7):585-594. [doi: 10.1056/nejmoa2108891]
18. Yadav P, Mohandas S, Shete A, Nyayanit D, Gupta N, Patil D, et al. SARS CoV-2 variant B.1.617.1 is highly pathogenic in hamsters than B.1 variant. bioRxiv. 2021. URL: https://www.biorxiv.org/content/10.1101/2021.05.05.442760v1 [accessed 2022-09-16]
19. Tada T, Zhou H, Dcosta BM, Samanovic MI, Mulligan MJ, Landau NR. Partial resistance of SARS-CoV-2 Delta variants to vaccine-elicited antibodies and convalescent sera. iScience 2021 Nov 19;24(11):103341 [FREE Full text] [doi: 10.1016/j.isci.2021.103341] [Medline: 34723159]
20. Hoffmann M, Hofmann-Winkler H, Krüger N, Kempf A, Nehlmeier I, Graichen L, et al. SARS-CoV-2 variant B.1.617 is resistant to bamlanivimab and evades antibodies induced by infection and vaccination. Cell Rep 2021 Jul 20;36(3):109415 [FREE Full text] [doi: 10.1016/j.celrep.2021.109415] [Medline: 34270919]
21. Ferreira I, Datir R, Papa G, Kemp S, Meng B, Rakshit P, et al. SARS-CoV-2 B.1.617 emergence and sensitivity to vaccine-elicited antibodies. bioRxiv. 2021. URL: https://www.biorxiv.org/content/10.1101/2021.05.08.443253v1 [accessed 2022-09-16]
22. Kumar A, Asghar A, Raza K, Narayan R, Jha R, Satyam A, et al. Demographic characteristics of SARS-CoV-2 B.1.617.2 (Delta) variant infections in Indian population. medRxiv. 2021. URL: https://www.medrxiv.org/content/10.1101/2021.09.23.21263948v1 [accessed 2022-09-16]
23. Kolozsvári LR, Bérczes T, Hajdu A, Gesztelyi R, Tiba A, Varga I, et al. Predicting the epidemic curve of the coronavirus (SARS-CoV-2) disease (COVID-19) using artificial intelligence: An application on the first and second waves. Inform Med Unlocked 2021;25:100691 [FREE Full text] [doi: 10.1016/j.imu.2021.100691] [Medline: 34395821]
24. Kibria H, Jyoti O, Matin A. Forecasting the spread of the third wave of COVID-19 pandemic using time series analysis in Bangladesh. Inform Med Unlocked 2022;28:100815 [FREE Full text] [doi: 10.1016/j.imu.2021.100815] [Medline: 34961844]
25. Sharif O, Hasan M, Rahman A. Determining an effective short term COVID-19 prediction model in ASEAN countries. Sci Rep 2022 Mar 24;12(1):5083 [FREE Full text] [doi: 10.1038/s41598-022-08486-5] [Medline: 35332192]
26. Mohan S, Solanki A, Taluja H, Anuradha, Singh A. Predicting the impact of the third wave of COVID-19 in India using hybrid statistical machine learning models: A time series forecasting and sentiment analysis approach. Comput Biol Med 2022 May;144:105354 [FREE Full text] [doi: 10.1016/j.compbiomed.2022.105354] [Medline: 35240374]

27. Yang Z, Zeng Z, Wang K, Wong S, Liang W, Zanin M, et al. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. J Thorac Dis 2020 Mar;12(3):165-174 [FREE Full text] [doi: 10.21037/jtd.2020.02.64] [Medline: 32274081]
28. Thakur S, Patel D, Soni B, Raval M, Chaudhary S. Prediction for the Second Wave of COVID-19 in India. In: Bellatreche L, Goyal V, Fujita H, Mondal A, Reddy PK, editors. Big Data Analytics. BDA 2020. Lecture Notes in Computer Science, vol 12581. Cham: Springer; 2020:134-150.
29. Salvatore M, Purkayastha S, Ganapathi L, Bhattacharyya R, Kundu R, Zimmermann L, et al. Lessons from SARS-CoV-2 in India: A data-driven framework for pandemic resilience. Sci Adv 2022 Jun 17;8(24):eabp8621 [FREE Full text] [doi: 10.1126/sciadv.abp8621] [Medline: 35714183]
30. Chen Z, Azman AS, Chen X, Zou J, Tian Y, Sun R, et al. Global landscape of SARS-CoV-2 genomic surveillance and data sharing. Nat Genet 2022 Apr;54(4):499-507 [FREE Full text] [doi: 10.1038/s41588-022-01033-y] [Medline: 35347305]
31. Oude Munnink BB, Worp N, Nieuwenhuijse D, Sikkema R, Haagmans B, Fouchier R, et al. The next phase of SARS-CoV-2 surveillance: real-time molecular epidemiology. Nat Med 2021 Sep;27(9):1518-1524. [doi: 10.1038/s41591-021-01472-w] [Medline: 34504335]
32. Global Initiative on Sharing All Influenza Data (GISAID). URL: https://gisaid.org/ [accessed 2022-09-16]
33. India coronavirus data. Worldometer. URL: https://www.worldometers.info/coronavirus/coronavirus/country/india/ [accessed 2022-09-16]
34. Dasgupta R. After India's brutal coronavirus wave, two-thirds of population has been exposed to SARS-CoV2. The Conversation. URL: https://theconversation.com/after-indias-brutal-coronavirus-wave-two-thirds-of-population-has-been-exposed-to-sars-cov2-165050 [accessed 2022-09-16]
35. Smith M, Trofimova M, Weber A, Duport Y, Kühnert D, von Kleist M. Rapid incidence estimation from SARS-CoV-2 genomes reveals decreased case detection in Europe during summer 2020. Nat Commun 2021 Oct 14;12(1):6009 [FREE Full text] [doi: 10.1038/s41467-021-26267-y] [Medline: 34650062]
36. Yadav PD, Nyayanit DA, Majumdar T, Patil S, Kaur H, Gupta N, et al. An Epidemiological Analysis of SARS-CoV-2 Genomic Sequences from Different Regions of India. Viruses 2021 May 17;13(5):925 [FREE Full text] [doi: 10.3390/v13050925] [Medline: 34067745]
37. Long S, Olsen R, Christensen P, Bernard D, Davis J, Shukla M, et al. Molecular Architecture of Early Dissemination and Massive Second Wave of the SARS-CoV-2 Virus in a Major Metropolitan Area. mBio 2020 Oct 30;11(6):30 [FREE Full text] [doi: 10.1128/mBio.02707-20] [Medline: 33127862]
38. Ahammad I, Hossain M, Rahman A, Chowdhury Z, Bhattacharjee A, Das K, et al. Wave-wise comparative genomic study for revealing the complete scenario and dynamic nature of COVID-19 pandemic in Bangladesh. PLoS One 2021;16(9):e0258019 [FREE Full text] [doi: 10.1371/journal.pone.0258019] [Medline: 34587212]
39. Maher M, Bartha I, Weaver S, di Iulio J, Ferri E, Soriaga L, et al. Predicting the mutational drivers of future SARS-CoV-2 variants of concern. Sci Transl Med 2022 Feb 23;14(633):eabk3445 [FREE Full text] [doi: 10.1126/scitranslmed.abk3445] [Medline: 35014856]
40. Nagpal S, Pal R, Ashima, Tyagi A, Tripathi S, Nagori A, et al. Genomic Surveillance of COVID-19 Variants With Language Models and Machine Learning. Front Genet 2022;13:858252 [FREE Full text] [doi: 10.3389/fgene.2022.858252] [Medline: 35464852]
41. Hill V, Ruis C, Bajaj S, Pybus O, Kraemer M. Progress and challenges in virus genomic epidemiology. Trends Parasitol 2021 Dec;37(12):1038-1049 [FREE Full text] [doi: 10.1016/j.pt.2021.08.007] [Medline: 34620561]
42. de Hoffer A, Vatani S, Cot C, Cacciapaglia G, Chiusano M, Cimarelli A, et al. Variant-driven early warning via unsupervised machine learning analysis of spike protein mutations for COVID-19. Sci Rep 2022 Jun 03;12(1):9275 [FREE Full text] [doi: 10.1038/s41598-022-12442-8] [Medline: 35661750]
43. Adiga R, Nayak V. Emergence of Novel SARS-CoV-2 variants in India: second wave. J Infect Dev Ctries 2021 Nov 30;15(11):1578-1583 [FREE Full text] [doi: 10.3855/jidc.15484] [Medline: 34898481]
44. Maslo C, Friedland R, Toubkin M, Laubscher A, Akaloo T, Kama B. Characteristics and Outcomes of Hospitalized Patients in South Africa During the COVID-19 Omicron Wave Compared With Previous Waves. JAMA 2022 Feb 08;327(6):583-584 [FREE Full text] [doi: 10.1001/jama.2021.24868] [Medline: 34967859]
45. Jassat W, Abdool Karim SS, Mudara C, Welch R, Ozougwu L, Groome MJ, et al. Clinical severity of COVID-19 in patients admitted to hospital during the omicron wave in South Africa: a retrospective observational study. The Lancet Global Health 2022 Jul;10(7):e961-e969. [doi: 10.1016/s2214-109x(22)00114-0]
46. Dyson L, Hill E, Moore S, Curran-Sebastian J, Tildesley M, Lythgoe K, et al. Possible future waves of SARS-CoV-2 infection generated by variants of concern with a range of characteristics. Nat Commun 2021 Sep 30;12(1):5730 [FREE Full text] [doi: 10.1038/s41467-021-25915-7] [Medline: 34593807]
47. Murhekar MV, Bhatnagar T, Thangaraj JWV, Saravanakumar V, Kumar MS, Selvaraju S, ICMR Serosurveillance Group. SARS-CoV-2 seroprevalence among the general population and healthcare workers in India, December 2020-January 2021. Int J Infect Dis 2021 Jul;108:145-155 [FREE Full text] [doi: 10.1016/j.ijid.2021.05.040] [Medline: 34022338]

48. Jahan N, Brahma A, Kumar M, Bagepally B, Ponnaiah M, Bhatnagar T, et al. Seroprevalence of IgG antibodies against SARS-CoV-2 in India, March 2020 to August 2021: a systematic review and meta-analysis. Int J Infect Dis 2022 Mar;116:59-67 [FREE Full text] [doi: 10.1016/j.ijid.2021.12.353] [Medline: 34968773]
49. Chakraborty C, Sharma A, Bhattacharya M, Agoramoorthy G, Lee S. The current second wave and COVID-19 vaccination status in India. Brain Behav Immun 2021 Aug;96:1-4 [FREE Full text] [doi: 10.1016/j.bbi.2021.05.018] [Medline: 34022371]

Abbreviations

GISAID: Global Initiative on Sharing All Influenza Data
VOC: variant of concern
VOI: variant of interest
WHO: World Health Organization

Edited by A Mavragani; submitted 28.01.22; peer-reviewed by R Rastmanesh, M Nali; comments to author 20.07.22; revised version received 26.08.22; accepted 12.09.22; published 22.09.22.

Please cite as:
Kumar A, Asghar A, Dwivedi P, Kumar G, Narayan RK, Jha RK, Parashar R, Sahni C, Pandey SN
A Bioinformatics Tool for Predicting Future COVID-19 Waves Based on a Retrospective Analysis of the Second Wave in India: Model Development Study
JMIR Bioinform Biotech 2022;3(1):e36860
URL: https://bioinform.jmir.org/2022/1/e36860
doi:10.2196/36860
PMID:36193192

Original Paper

Prediction of Antibody-Antigen Binding via Machine Learning: Development of Data Sets and Evaluation of Methods

Chao Ye[1], MIS; Wenxing Hu[2], MIT; Bruno Gaeta[1], PhD

[1]School of Computer Science and Engineering, The University of New South Wales, Sydney, Australia

[2]Department of Computer Science, School of Information Science and Technology, Tokyo Institute of Technology, Tokyo, Japan

Corresponding Author:
Bruno Gaeta, PhD
School of Computer Science and Engineering
The University of New South Wales
Computer Science Building (K17)
Engineering Rd, UNSW
Sydney, 2052
Australia
Phone: 61 293857213
Email: bgaeta@unsw.edu.au

Abstract

Background: The mammalian immune system is able to generate antibodies against a huge variety of antigens, including bacteria, viruses, and toxins. The ultradeep DNA sequencing of rearranged immunoglobulin genes has considerable potential in furthering our understanding of the immune response, but it is limited by the lack of a high-throughput, sequence-based method for predicting the antigen(s) that a given immunoglobulin recognizes.

Objective: As a step toward the prediction of antibody-antigen binding from sequence data alone, we aimed to compare a range of machine learning approaches that were applied to a collated data set of antibody-antigen pairs in order to predict antibody-antigen binding from sequence data.

Methods: Data for training and testing were extracted from the Protein Data Bank and the Coronavirus Antibody Database, and additional antibody-antigen pair data were generated by using a molecular docking protocol. Several machine learning methods, including the weighted nearest neighbor method, the nearest neighbor method with the BLOSUM62 matrix, and the random forest method, were applied to the problem.

Results: The final data set contained 1157 antibodies and 57 antigens that were combined in 5041 antibody-antigen pairs. The best performance for the prediction of interactions was obtained by using the nearest neighbor method with the BLOSUM62 matrix, which resulted in around 82% accuracy on the full data set. These results provide a useful frame of reference, as well as protocols and considerations, for machine learning and data set creation in the prediction of antibody-antigen binding.

Conclusions: Several machine learning approaches were compared to predict antibody-antigen interaction from protein sequences. Both the data set (in CSV format) and the machine learning program (coded in Python) are freely available for download on GitHub.

(JMIR Bioinform Biotech 2022;3(1):e29404) doi:10.2196/29404

KEYWORDS

DNA sequencing; DNA; DNA sequence; sequence data; molecular biology; genomic; random forest; nearest neighbor; immunoglobulin; genetics; antibody-antigen binding; antigen; antibody; structural biology; machine learning; protein modeling; protein; proteomic

Introduction

DNA sequencing technologies are providing new insights into the immune response by allowing for the large-scale sequencing of rearranged immunoglobulin genes that are present in an individual [1,2]. However, the applications of this approach are limited by the lack of methods for determining the antigen(s) to which a specific immunoglobulin (ie, one encoded by a given sequence) binds. Individual immunoglobulins can be tested experimentally at significant cost; however, the large-scale

characterization of binding properties based on sequence data is currently impossible.

Antigen binding is mediated by the complementarity-determining regions (CDRs) of an antibody, which are shared between heavy and light immunoglobulin chains. Computational methods for predicting antibody-antigen interactions that leverage structure prediction and docking have been proposed [3]. However, the use of these methods requires knowledge of the 3D structures of antibodies and antigens. The direct prediction of antibody-antigen interactions from protein sequences remains an open problem.

Machine learning–based tools, such as mCSM-AB [4] and ADAPT (Assisted Design of Antibody and Protein Therapeutics) [5], have had some success in predicting antibody interactions in other contexts. mCSM-AB is a web server for predicting changes in antibody-antigen affinity upon mutation, using graph-based signatures. ADAPT is an affinity maturation platform that interleaves predictions and testing, and it has been previously validated on monoclonal antibodies.

A more general method for predicting whether an antibody will bind to a protein antigen based on the antibody and antigen sequences remains elusive, in part due to the lack of comprehensive training data for the development of machine learning models. This study is intended as a first step toward this goal and aims to assemble a training data set from a range of sources and evaluate the feasibility of applying machine learning algorithms to identify the binding of antibody-antigen pairs in this data set.

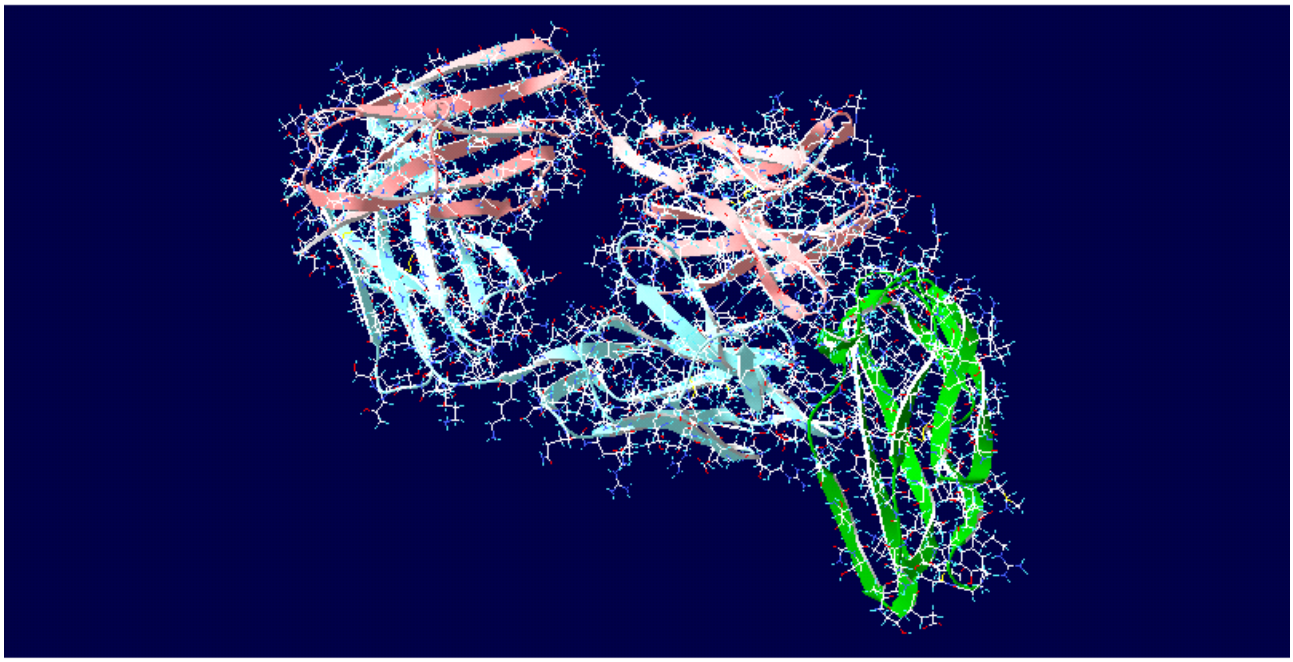
Methods

Data Set

Due to the scarcity of suitable antibody-antigen pairs, computational docking was used to generate some of the data in the training and testing data set. The ClusPro (Boston University) [6-9] and Rosetta (RosettaCommons) [10-12] web servers were used to create a data set of paired antibody-antigen complexes for machine learning. Both ClusPro and Rosetta were used for protein-protein molecular docking. Rosetta uses the SnugDock (RosettaCommons) algorithm [10]. The Swiss-PdbViewer (Swiss Institute of Bioinformatics) [13] was used to examine the resulting protein complex structures.

A total of 50 antibody-antigen complexes were selected randomly from the Protein Data Bank (PDB) [14]. The antibody-antigen complexes were separated by using a Perl script to produce PDB-formatted files as well as sequences for antibodies and antigens. CDRs were located by using the Rosetta antibody modeling web server. Antigens were docked with a range of antibodies by using ClusPro (used only to determine orientation), followed by Rosetta's antibody docking program, SnugDock. In order to keep computation times manageable, not all antibodies were docked. Instead, 10 to 14 antibodies were randomly selected to be docked with each antigen in order to find the best orientation. The resulting complexes were submitted to the Rosetta SnugDock web server in order to calculate the best interface score. This produced structures for between 10 and 14 complexes per antigen, which, when added together with the original antibody-antigen complex, totaled 11 to 15 complexes per antigen. Altogether, 50 antigens were docked with 600 antibodies. An example of a resulting complex is shown in Figure 1.

Figure 1. Example of a docking output. The 3s35 complex was generated by using the ClusPro server (docking results: "YES"; best docking interface score: −0.876).



The Rosetta interface scores were used as estimates of binding affinity in order to identify cognate antibody-antigen pairs to be used as input for machine learning. Complexes with interface scores of higher than −8.0 were classified outright as complexes with poor binding, and those with interface scores of lower than −9.0 were classified outright as complexes with good binding. For complexes with scores that ranged between −8.0 and −9.0, the docking clusters and positions were examined visually by using SwissDock (Swiss Institute of Bioinformatics). If the top 10 models had their antibodies and antigens in similar relative positions and the structures showed sensible interaction patterns, the pairs were classified as having a good binding affinity.

Rosetta interface scores have been used previously as classifiers to determine binding affinity based on docking results (eg, in an antibody-antigen cross reactivity study [15]).

Additional data were extracted from the Coronavirus Antibody Database (CoV-AbDab) [16]—a database of antibodies against coronaviruses, including SARS-CoV-2, SARS-CoV-1, and MERS-CoV (Middle East respiratory syndrome–related coronavirus). Data (2674 rows) were extracted from the CoV-AbDab on February 14, 2021. After filtering out incomplete data, 2031 rows remained, with each row corresponding to an antibody. The information extracted comprised the antibody names, their binding antigens, and their heavy and light variable region sequences, including the locations of the third CDRs (CDR3s). Each of the variable region sequences were searched against the international ImMunoGeneTics information system database [17] in order to identify the locations of the first CDRs (CDR1s) and second CDRs (CDR2s) from the heavy and light chains. Since a row may contain information about an antibody's interactions with multiple antigens, the data were further split into multiple rows, with each row containing information about the interaction between 1 antibody and 1 antigen.

Additional features were calculated for the sequences, as follows. The isoelectric point for each CDR was calculated by using the Bachem peptide calculator analysis tool (Bachem Holding AG) [18]. The average hydrophilicity of each CDR was also calculated by using the Bachem peptide calculator.

B cell epitopes were predicted by using the IEDB (Immune Epitope Database) antibody epitope prediction analysis tool [19].

The resulting data set can be downloaded from GitHub [20] and is structured with the following column headings: *H chain CDR1 sequence*, *H chain CDR2 sequence*, *H chain CDR3 sequence*, *L chain CDR1 sequence*, *L chain CDR2 sequence*, *L chain CDR3 sequence*, *Hydrophilicity of L CDR1*, *pI of L CDR1*, *Hydrophilicity of L CDR2*, *pI of L CDR2*, *Hydrophilicity of L CDR3*, *pI of L CDR3*, *Hydrophilicity of H CDR1*, *pI of H CDR1*, *Hydrophilicity of H CDR2*, *pI of H CDR2*, *Hydrophilicity of H CDR3*, *pI of H CDR3*, *Antigen Epitope*, *Rosetta Docking score*, *Antigen*, and *Docking result*.

Machine Learning

A weighted K-nearest neighbor (K-NN) classification algorithm [21] for predicting antibody-antigen binding affinity was implemented in Python. The program can be downloaded from GitHub [20].

For each antigen, the 11 to 15 antibodies that were docked were labeled as “good affinity” or “low affinity,” on the basis of the docking results. Machine learning was then performed, using the sequences of both antigens and antibodies.

Neighbors were determined by using the string distances between the CDR1, CDR2, and CDR3 amino acid sequences of different antibodies. Weights were calculated from distances, so that nearer neighbors were considered to have more weight, as detailed below.

For every antigen, the class (good affinity or low affinity) was learned by using the K-NN method, using a training subset (N − 1) of the labeled antigen-antibody sequence pairs and using the CDR string distances as features. The model performance was then evaluated on the remaining antigen-antibody sequence pair that was not used for training (leave-one-out cross-validation).

In order to ensure that the K-NN pairs only included pairs with the same antigen, a fixed penalty of 1000 was added to the distances between antibody-antigen pairs involving different antigens.

The similarity between antibodies was measured via a comparison of their CDRs. Each antibody has a heavy chain and a light chain, and each chain contains 3 CDRs. The distance between 2 antibodies was calculated as the Euclidean distance between their CDR distance vectors, as shown in the following equation (equation 1):

where $(q_i - p_i)$ represents the string distance between the CDR_i of antibody q and the CDR_i of antibody p.

The Python code is given in Multimedia Appendix 1.

Two different CDR distance calculation methods were tested and compared; one was based on sequence identity, and the other used the BLOSUM62 matrix, as detailed below.

For the identity-based distance measure, pairs of equivalent CDRs were compared with each other based on their Levenshtein string distances [22], as shown in the following equation (equation 2):

$Cost=0$ for $a_i=b_i$, $Cost=1$ for $a_i \neq b_i$

The Levenshtein distance only accounts for amino acid identity when it is used for comparing sequences. A more biologically significant distance measure needs to take into account the different properties of amino acids, which means that some amino acid substitutions are more likely to be accepted in an interaction than others. The BLOSUM62 substitution matrix [23] was used as a proxy for amino acid similarity in the Levenshtein distance calculation. Although the BLOSUM matrices were designed to reflect evolutionary conservation,

they can provide an estimate of similarity in interaction potential [24].

The Levenshtein distance was calculated as per equation 2, using the following cost function:

For $a_i=b_i$, $Cost=0$

where S_{ij}, S_{ii}, and S_{jj} are obtained from the BLOSUM62 matrix.

The following columns from the data set were used to train the model for leave-one-out cross-validation: *H chain CDR1 sequence*, *H chain CDR2 sequence*, *H chain CDR3 sequence*, *L chain CDR1 sequence*, *L chain CDR2 sequence*, *L chain CDR3 sequence*, *Antigen*, and *Docking result*. The trained model was then evaluated on its ability to predict the docking results from the other columns.

A random forest machine learning algorithm incorporating the previous K-NN results was also used for predicting antibody-antigen binding classification. The isoelectric point and net charge at neutral pH (7.0) for each CDR were used as additional features, in addition to the BLOSUM62-derived CDR distances, for training the random forest. Binding was predicted by combining the votes from each of the features, and each individual feature contributed 1 vote, according to the nearest neighbor predictions based on each feature.

The following columns from the data set were used for training the random forest: *String distance (calculate by KNN method)*, *Hydrophilicity of L CDR1*, *pI of L CDR1*, *Hydrophilicity of L CDR2*, *pI of L CDR2*, *Hydrophilicity of L CDR3*, *pI of L CDR3*, *Hydrophilicity of H CDR1*, *pI of H CDR1*, *Hydrophilicity of H CDR2*, *pI of H CDR2*, *Hydrophilicity of H CDR3*, *pI of H CDR3*, *Antigen*, and *Docking result*. The trained model was then evaluated on its ability to predict the docking results from the other columns.

Each feature was considered as an individual decision tree and contributed 1 vote. For example, the isoelectric point of the CDR1 of an antibody's heavy chain was considered as 1 feature, and the K-NN method was used, as previously described, to find the results of this decision tree. Altogether, there were 13 decision trees, and each tree used the K-NN method to determine its vote, for a total of 13 votes. The final decision was determined based on a simple majority vote. The best results were obtained when the whole forest (all 13 decision trees) took part in the vote.

The performance of the K-NN and random forest learners was evaluated by using leave-one-out cross-validation on an antigen basis. For each of the 57 antigens, a training data set was constructed by removing 1 row, that is, 1 antibody-antigen pair, from the data set. After training with the remaining antibodies that bound to this antigen, model performance was evaluated based on the removed antibody. The process was repeated until all 5041 antibody-antigen pairs were tested. Model accuracy was calculated as the ratio of the number of correctly predicted antibody-antigen pairs over the total number of pairs in the data set.

Results

Data Set

A total of 600 antibody-antigen complexes were generated via the computational docking of 50 antibody structures with 50 antigen structures. In addition, a total of 4441 antibody-antigen pairs were extracted from the Cov-AbDab. The composition of this section of the data set is shown in Table 1.

In total, the data set contained 5041 antibody-antigen pairs comprising 1157 antibodies and 57 antigens.

Table 1. Number of antibodies and positive and negative antibody-antigen pairs extracted from the Coronavirus Antibody Database.

Antigen	Number of antibodies	Positive samples, n	Negative samples, n
SARS-CoV-2	1943	1912	31
SARS-CoV-1	1241	597	644
MERS-CoV[a]	264	119	145
HCoV-OC43[b]	257	21	236
HCoV-HKU1[c]	254	84	170
HCoV-NL63[d]	258	51	207
HCoV-229E[e]	207	49	158

[a]MERS-CoV: Middle East respiratory syndrome–related coronavirus.

[b]HCoV-OC43: human coronavirus OC43.

[c]HCoV-HKU1: human coronavirus HKU1.

[d]HCoV-NL63: human coronavirus NL63.

[e]HCoV-229E: human coronavirus 229E.

Machine Learning

The antigen-antibody binding classification methods were evaluated by using leave-one-out cross-validation. For a K value of 2 nearest neighbors, the K-NN method, when the Levenshtein distance was calculated based on sequence identity, achieved an accuracy of 81%. A slight improvement (accuracy of 82%) was observed when using the BLOSUM62 matrix to calculate the Levenshtein string distance.

Different K values were also evaluated when the Levenshtein distance was calculated based on the BLOSUM62 matrix. A K value of 2 provided the best accuracy. For a K value of 1 nearest neighbor, the accuracy was 80%. For a K value of 3, classification accuracy dropped to 79%.

For the random forest predictions, votes were used as the classification prediction results. The accuracy was highest when the whole forest was considered, in which case each feature contributed to the classification results. The performance of the random forest method was best (accuracy of 80%) when all 13 features—the Levenshtein string distance and the isoelectric point and net charge at neutral pH (7.0) for each CDR—took part in the final votes.

Discussion

We created a training and test data set of 5041 antibody-antigen complexes by using a combination of structure modeling and computational docking via Rosetta, together with antibody-antigen pairs extracted from the CoV-AbDab.

We also developed weighted nearest neighbor and random forest approaches to predict antibody-antigen binding based on sequence data. These machine learning procedures can perform classifications to identify antigens that are likely to bind to a given antibody.

Leave-one-out cross-validation testing yielded an accuracy of 82% for classification results that were based on 2 nearest neighbors. The prediction accuracy ranged from around 77% to 82% when varying the number of nearest neighbors. The best prediction results (accuracy of 82%) were obtained with 2 nearest neighbors, using string distance and BLOSUM62 matrices.

This study demonstrates that the interaction between an antibody and a protein antigen can be predicted from the amino acid sequences of both the antibody's variable regions and the antigen by using a relatively simple machine learning approach. Compared to the docking prediction method, which is based on the spatial protein structure, the method proposed in this project does not require a 3D structure and is more suitable for antibodies for which a 3D structure is unavailable.

In the absence of large amounts of experimental data on antibody-antigen binding affinities, the Rosetta interface scores, along with the top 10 binding positions, were used to determine the classification for binding affinity. Although this method was unlikely to provide a full representation of the problem, it provided a data set suitable for comparing a range of approaches. This method will certainly improve as larger data sets become available. The docking data set contained 600 rows of antibody-antigen pairs. Subsets of this data set (200, 300, 400, and 500 rows) were tested during the data collection process. Classification accuracy was quite consistent across all of these subsets. This indicates that while the data set is limited, it provides a good starting point for the development of our approach for the prediction of antibody-antigen binding affinity, which can be further validated as more data become available. The K-NN method was chosen as the initial machine learning method. The best prediction results were obtained with 2 nearest neighbors (K=2). Random forests were also used that incorporated sequence distance as well as the chemical properties of CDRs (isoelectric point and hydrophobicity). The best prediction results (accuracy of 82%) were obtained with the nearest neighbor method when the Levenshtein distance was calculated based on BLOSUM62 matrices. The additional features included in the random forest did not improve classification accuracy, and this was probably due to these features' dependence on the amino acid sequences.

Around 20% (907/5041, 18%) of our method's predictions were inaccurate. These errors mostly occurred with some large antigens. The docking results for these antigens were further examined. The decreased accuracy was likely the result of conformational flexibility in the larger antigens, the presence of multiple epitopes, and the higher number of discontinuous epitopes in larger antigens relative to the number of such epitopes in smaller antigens.

As a step toward the development of a machine learning method suitable for predicting antibody-antigen binding affinities from sequence data, the weighted nearest neighbor and random forest machine learning approaches were applied to the problem. The basic hypothesis was that antibodies with similar sequences may be similar in terms of their ability to bind to a given antigen. A prediction program was coded in Python and evaluated via cross-validation on a data set containing 1157 antibodies and 57 antigens that were combined in 5041 antibody-antigen pairs. The best classification prediction accuracy was around 82% for this data set.

These results provide a useful frame of reference, as well as protocols and considerations, for machine learning and data set creation in the prediction of antibody-antigen binding. Our method is still limited due to the scarcity of training data, but its usefulness for large-scale prediction should increase as more antibody-antigen binding data become available. The ability to predict antibody-antigen binding will allow for a more informed use of data from large-scale immune receptor sequencing. This, in turn, will increase our understanding of the variation in antigen recognition in an organism over time, under a range of conditions and between individuals and populations.

Both the data set (in CSV format) and the machine learning program (coded in Python) are freely available for download on GitHub [20].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Python code for Euclidean distance calculation.
[DOCX File , 12 KB - bioinform_v3i1e29404_app1.docx]

References

1. Dunn-Walters D, Townsend C, Sinclair E, Stewart A. Immunoglobulin gene analysis as a tool for investigating human immune responses. Immunol Rev 2018 Jul;284(1):132-147 [FREE Full text] [doi: 10.1111/imr.12659] [Medline: 29944755]
2. Boyd SD, Crowe JEJ. Deep sequencing and human antibody repertoire analysis. Curr Opin Immunol 2016 Jun;40:103-109 [FREE Full text] [doi: 10.1016/j.coi.2016.03.008] [Medline: 27065089]
3. Weitzner BD, Jeliazkov JR, Lyskov S, Marze N, Kuroda D, Frick R, et al. Modeling and docking of antibody structures with Rosetta. Nat Protoc 2017 Feb;12(2):401-416 [FREE Full text] [doi: 10.1038/nprot.2016.180] [Medline: 28125104]
4. Pires DEV, Ascher DB. mCSM-AB: a web server for predicting antibody-antigen affinity changes upon mutation with graph-based signatures. Nucleic Acids Res 2016 Jul 08;44(W1):W469-W473 [FREE Full text] [doi: 10.1093/nar/gkw458] [Medline: 27216816]
5. Vivcharuk V, Baardsnes J, Deprez C, Sulea T, Jaramillo M, Corbeil CR, et al. Assisted Design of Antibody and Protein Therapeutics (ADAPT). PLoS One 2017 Jul 27;12(7):e0181490 [FREE Full text] [doi: 10.1371/journal.pone.0181490] [Medline: 28750054]
6. Kozakov D, Hall DR, Xia B, Porter KA, Padhorny D, Yueh C, et al. The ClusPro web server for protein-protein docking. Nat Protoc 2017 Feb;12(2):255-278 [FREE Full text] [doi: 10.1038/nprot.2016.169] [Medline: 28079879]
7. Comeau SR, Gatchell DW, Vajda S, Camacho CJ. ClusPro: an automated docking and discrimination method for the prediction of protein complexes. Bioinformatics 2004 Jan 01;20(1):45-50. [doi: 10.1093/bioinformatics/btg371] [Medline: 14693807]
8. Kozakov D, Brenke R, Comeau SR, Vajda S. PIPER: an FFT-based protein docking program with pairwise potentials. Proteins 2006 Nov 01;65(2):392-406. [doi: 10.1002/prot.21117] [Medline: 16933295]
9. Comeau SR, Gatchell DW, Vajda S, Camacho CJ. ClusPro: a fully automated algorithm for protein-protein docking. Nucleic Acids Res 2004 Jul 01;32(Web Server issue):W96-W99 [FREE Full text] [doi: 10.1093/nar/gkh354] [Medline: 15215358]
10. Sircar A, Gray JJ. SnugDock: paratope structural optimization during antibody-antigen docking compensates for errors in antibody homology models. PLoS Comput Biol 2010 Jan 22;6(1):e1000644 [FREE Full text] [doi: 10.1371/journal.pcbi.1000644] [Medline: 20098500]
11. Lyskov S, Gray JJ. The RosettaDock server for local protein-protein docking. Nucleic Acids Res 2008 Jul 01;36(Web Server issue):W233-W238 [FREE Full text] [doi: 10.1093/nar/gkn216] [Medline: 18442991]
12. Lyskov S, Chou FC, Conchúir S, Der BS, Drew K, Kuroda D, et al. Serverification of molecular modeling applications: the Rosetta Online Server that Includes Everyone (ROSIE). PLoS One 2013 May 22;8(5):e63906 [FREE Full text] [doi: 10.1371/journal.pone.0063906] [Medline: 23717507]
13. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 1997 Dec;18(15):2714-2723. [doi: 10.1002/elps.1150181505] [Medline: 9504803]
14. RCSB PDB: Homepage. RCSB Protein Data Bank. URL: https://www.rcsb.org/ [accessed 2019-07-12]
15. Kilambi KP, Gray JJ. Structure-based cross-docking analysis of antibody-antigen interactions. Sci Rep 2017 Aug 15;7(1):8145 [FREE Full text] [doi: 10.1038/s41598-017-08414-y] [Medline: 28811664]
16. Raybould MIJ, Kovaltsuk A, Marks C, Deane CM. CoV-AbDab: The Coronavirus Antibody Database. Bioinformatics 2021 May 05;37(5):734-735 [FREE Full text] [doi: 10.1093/bioinformatics/btaa739] [Medline: 32805021]
17. IMGT home page. The international ImMunoGeneTics information system. URL: http://www.imgt.org [accessed 2020-01-23]
18. Peptide calculator. Bachem. URL: https://www.bachem.com/knowledge-center/peptide-calculator/ [accessed 2020-03-10]
19. Antibody epitope prediction. Immune Epitope Database. URL: http://tools.iedb.org/bcell/ [accessed 2020-07-12]
20. Chao Ye. jessye123/ab-ag-seq-machine-learning. GitHub. URL: https://github.com/jessye123/ab-ag-seq-machine-learning [accessed 2022-10-19]
21. Taunk K, De S, Verma S, Swetapadma A. A brief review of nearest neighbor algorithm for learning and classification. 2019 Presented at: 2019 International Conference on Intelligent Computing and Control Systems (ICCS); May 15-17, 2019; Madurai, India p. 1255-1260. [doi: 10.1109/iccs45141.2019.9065747]
22. Levenshtein VI. Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Doklady 1966 Feb;10(8):707-710 [FREE Full text]
23. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 1992 Nov 15;89(22):10915-10919 [FREE Full text] [doi: 10.1073/pnas.89.22.10915] [Medline: 1438297]

24. Huang YA, You ZH, Gao X, Wong L, Wang L. Using Weighted Sparse Representation Model Combined with Discrete Cosine Transformation to Predict Protein-Protein Interactions from Protein Sequence. Biomed Res Int 2015;2015:902198 [FREE Full text] [doi: 10.1155/2015/902198] [Medline: 26634213]

Abbreviations

ADAPT: Assisted Design of Antibody and Protein Therapeutics
CDR: complementarity-determining region
CDR1: first complementarity-determining region
CDR2: second complementarity-determining region
CDR3: third complementarity-determining region
CoV-AbDab: Coronavirus Antibody Database
IEDB: Immune Epitope Database
K-NN: K-nearest neighbor
MERS-CoV: Middle East respiratory syndrome–related coronavirus
PDB: Protein Data Bank

Edited by A Mavragani; submitted 06.04.21; peer-reviewed by Z Qiu, Y Xiao, ME Ackerman, H Sundaramoorthi; comments to author 20.05.21; revised version received 23.09.21; accepted 18.10.22; published 28.10.22.

Please cite as:
Ye C, Hu W, Gaeta B
Prediction of Antibody-Antigen Binding via Machine Learning: Development of Data Sets and Evaluation of Methods
JMIR Bioinform Biotech 2022;3(1):e29404
URL: https://bioinform.jmir.org/2022/1/e29404
doi:10.2196/29404
PMID:

Original Paper

Differential Expression of Long Noncoding RNAs in Murine Myoblasts After Short Hairpin RNA-Mediated Dysferlin Silencing In Vitro: Microarray Profiling

Richa Singhal[1*], MS, PhD; Rachel Lukose[2*], BS, MPAS; Gwenyth Carr[3], BS; Afsoon Moktar[2], BSN, CT, EMBA, MS, PhD; Ana Lucia Gonzales-Urday[4], BS, MS; Eric C Rouchka[1], DSc; Bathri N Vajravelu[2], MPH, MBBS, PhD

[1]Department of Biochemistry and Molecular Genetics, KY IDeA Networks of Biomedical Research Excellence Bioinformatics Core, University of Louisville, Louisville, KY, United States

[2]Department of Physician Assistant Studies, Massachusetts College of Pharmacy and Health Sciences, Boston, MA, United States

[3]Department of Medical and Molecular Biology, School of Arts and Sciences, Massachusetts College of Pharmacy and Health Sciences, Boston, MA, United States

[4]Department of Pharmaceutical Sciences, Massachusetts College of Pharmacy and Health Sciences, Boston, MA, United States

[*]these authors contributed equally

Corresponding Author:
Bathri N Vajravelu, MPH, MBBS, PhD
Department of Physician Assistant Studies
Massachusetts College of Pharmacy and Health Sciences
179 Longwood Avenue
Boston, MA, 02115
United States
Phone: 1 617 732 2961
Fax: 1 617 732 1027
Email: bathri.vajravelu@mcphs.edu

Abstract

Background: Long noncoding RNAs (lncRNAs) are noncoding RNA transcripts greater than 200 nucleotides in length and are known to play a role in regulating the transcription of genes involved in vital cellular functions. We hypothesized the disease process in dysferlinopathy is linked to an aberrant expression of lncRNAs and messenger RNAs (mRNAs).

Objective: In this study, we compared the lncRNA and mRNA expression profiles between wild-type and dysferlin-deficient murine myoblasts (C2C12 cells).

Methods: LncRNA and mRNA expression profiling were performed using a microarray. Several lncRNAs with differential expression were validated using quantitative real-time polymerase chain reaction. Gene Ontology (GO) analysis was performed to understand the functional role of the differentially expressed mRNAs. Further bioinformatics analysis was used to explore the potential function, lncRNA-mRNA correlation, and potential targets of the differentially expressed lncRNAs.

Results: We found 3195 lncRNAs and 1966 mRNAs that were differentially expressed. The chromosomal distribution of the differentially expressed lncRNAs and mRNAs was unequal, with chromosome 2 having the highest number of lncRNAs and chromosome 7 having the highest number of mRNAs that were differentially expressed. Pathway analysis of the differentially expressed genes indicated the involvement of several signaling pathways including PI3K-Akt, Hippo, and pathways regulating the pluripotency of stem cells. The differentially expressed genes were also enriched for the GO terms, *developmental process* and *muscle system process*. Network analysis identified 8 statistically significant (P<.05) network objects from the upregulated lncRNAs and 3 statistically significant network objects from the downregulated lncRNAs.

Conclusions: Our results thus far imply that dysferlinopathy is associated with an aberrant expression of multiple lncRNAs, many of which may have a specific function in the disease process. GO terms and network analysis suggest a muscle-specific role for these lncRNAs. To elucidate the specific roles of these abnormally expressed noncoding RNAs, further studies engineering their expression are required.

(JMIR Bioinform Biotech 2022;3(1):e33186) doi:10.2196/33186

KEYWORDS

dysferlinopathy; long noncoding RNAs; lncRNA; abnormal expression; muscular dystrophy; limb-girdle muscular dystrophy 2B; LGMD-2B; messenger RNA; mRNA; quantitative real-time polymerase chain reaction; qRT-PCR; gene ontology; bioinformatics; transcription; noncoding RNA; protein expression

Introduction

Dysferlinopathy is a type of muscular dystrophy, which are a group of inherited muscle degenerative disorders characterized by progressive muscle weakness. It is caused by the deficiency of dysferlin, a type II transmembrane protein that is highly expressed in skeletal muscle and the heart. Dysferlin interacts with several proteins by acting as a scaffold and plays a crucial role in calcium-dependent membrane repair in skeletal muscles [1]. Dysferlin deficiency in the muscles is characterized by vesicular accumulations, sarcolemmal disruptions, defective myogenesis, and increased inflammation [2], both in mouse models and in humans [3]. Dysferlinopathy has 2 subtypes—limb-girdle muscular dystrophy (LGMD) 2B and Miyoshi myopathy. LGMD-2B involves the proximal muscles of the limb and trunk, whereas Miyoshi myopathy involves the posterior compartment muscles of the lower limb [4,5]. The progressive muscle degeneration caused by this disease results in mobility impairment and disability that increases in severity during advanced stages. However, most have an approximately normal life span. The age of onset, rate of progression, and severity of this disease are highly variable and unpredictable in patients with dysferlinopathy, indicating a role for environmental and epigenetic factors. Although the exact prevalence of dysferlinopathy is not known, most LGMDs are rare with an estimated prevalence ranging from 0.07-0.43 per 100,000 people [6]. Currently, there is no cure for this disease, and existing treatment strategies are aimed at managing complications and prolonging life span.

In recent years, scientists have discovered a group of heterogeneous RNA molecules called noncoding RNAs (ncRNAs) that participates in many physiological functions and disease processes. Interestingly, only 2% of the eukaryotic genome is transcribed to functional protein and the remaining 98% is considered as nonprotein coding RNAs or ncRNAs [7]. NcRNAs are broadly classified into (1) structural ncRNAs that include ribosomal RNAs, transfer RNAs, small nuclear RNAs, and small nucleolar RNAs and (2) regulatory ncRNAs that include small ncRNAs (shorter than 200 nucleotides) and long noncoding RNAs (lncRNAs). LncRNAs are transcripts longer than 200 nucleotides and participate in regulating gene expression through a variety of mechanisms [8]. For example, they participate in chromatin remodeling through the recruitment of polycomb repressive complex 2, causing gene repression [9,10]. In addition, they regulate the binding of transcription factors to the DNA loci by forming hybridization structures with the DNA. Some of the lncRNAs are known to function as competing endogenous RNAs and sponge-specific microRNAs to regulate gene expression [11-13]. Interestingly, they can reduce the stability of messenger RNAs (mRNAs) or increase their translation through different mechanisms depending on their subcellular localization, interacting partners, and local environments in the cells [14]. Some recent studies have described the function and mechanism of selected lncRNAs in the pathogenesis of specific diseases, including cardiac hypertrophy [15], osteoarthritis [16], and fascioscapulohumeral muscular dystrophy [17]. Further, some studies have revealed a few muscle-specific lncRNAs that are involved in regulating muscle cell growth and differentiation [13,18]. Nevertheless, their expression signature, function, and contribution to the disease process of dysferlinopathy are not well studied.

This paper presents the results of the analysis on the lncRNAs expression profile between wild-type and dysferlin-deficient murine myoblasts using a microarray. To confirm our microarray results, we validated some of the lncRNAs that were differentially expressed with the help of quantitative real-time polymerase chain reaction (qRT-PCR). Additionally, using bioinformatics, this paper also annotates the possible cellular function and interactions for these lncRNAs.

Methods

Cell Culture

The C2C12 cell line was a generous gift from Dr Robert H. Brown, Department of Neurology, University of Massachusetts [19]. A fixed density of wild-type and dysferlin-deficient C2C12 cells were cultured in T75 flasks in growth media containing DMEM supplemented with 20% bovine growth serum (HyClone) and 1% penicillin-streptomycin. For the dysferlin-deficient cells, 1.5 µg/mL puromycin was added to the growth media. Media were changed every 2 days and cells were split before they reach confluence in order to avoid differentiation and cell death.

RNA Extraction

Total RNA from the wild-type and dysferlin-deficient C2C12 cells was extracted using the RNeasy plus kit (Qiagen) following the manufacturer's instructions. A NanoDrop ND-1000 spectrophotometer was used to measure the quality and quantify the RNA samples. The integrity of the RNA samples was assessed by agarose gel electrophoresis.

Microarray

LncRNA and mRNA expression profiling were performed using Arraystar Mouse LncRNA Microarray V3.0 containing 21,486 lncRNA and 18,921 mRNA probes. The lncRNA probes were created based on the information derived from reputable transcriptome public databases, such as Refseq, UCSC known genes, and Ensembl and landmark publications. The percentage of probes made for each category of lncRNAs is given in Table 1. For accurate identification of individual transcripts, probes that will bind to specific exons or splice junctions were designed. For the purpose of hybridization quality control, the array also contained positive probes (for housekeeping genes) and negative probes.

Table 1. Percentage of lncRNA probes designed for each category.

LncRNA category	Probes designed, n (%)
Bidirectional	993 (4.6)
Exon sense-overlapping	7451 (34.7)
Intergenic	9183 (42.7)
Intron sense-overlapping	497 (2.3)
Intronic antisense	1425 (6.6)
Natural antisense	1937 (9)
Total	21,486 (100)

RNA Labeling and Array Hybridization

Sample labeling and array hybridization were performed according to the Agilent One-Color Microarray-Based Gene Expression Analysis protocol (Agilent Technology) with minor modifications. Briefly, ribosomal RNA was first removed from the total RNA sample, and then mRNA was purified using the mRNA-ONLY Eukaryotic mRNA Isolation Kit (Epicentre Biotechnologies). The purified samples were then amplified and transcribed into fluorescent complementary RNA (cRNA) along the entire length of the transcripts without 3' bias using a mixture of oligo(dT) and random priming method (Arraystar Flash RNA Labeling Kit). This was followed by cRNA purification using the RNeasy Mini Kit (Qiagen) following the manufacturer's instructions. A NanoDrop ND-1000 spectrophotometer was used to measure the concentration and specific activity of the labeled cRNAs (pmol Cy3/µg cRNA). 5 µL of 10 × Blocking Agent and 1 µL of 25 × Fragmentation Buffer were added to 1 µg of the labeled cRNA to fragment, and then the mixture was heated at 60 °C for 30 minutes. To achieve the desired dilution, 25 µL of 2 × GE Hybridization buffer was added. 50 µL of hybridization solution was used to assemble the probes on the microarray slides. The slides were incubated for 17 hours at 65 °C in an Agilent Hybridization Oven. The hybridized arrays were washed, fixed, and scanned using the Agilent DNA Microarray Scanner (part number G2505C).

Data Analysis

The array images were acquired and analyzed using the Agilent Feature Extraction software (version 11.0.1.1). GeneSpring GX software package (version 12.1; Agilent Technologies) was used for further data processing and quantile normalization. Further analysis was done on samples that had flags in Present or Marginal ("All Targets Value") values. Differentially expressed lncRNAs and mRNAs were identified through fold change (FC) filtering between the samples. The cutoff values were $|FC| \geq 2$, where FC values in linear scale (not in $\log_2$ scale) were calculated based on the normalized intensities.

GO Analysis and Pathway Analysis

GO analysis was derived from Gene Ontology [20], which has 3 structured networks of defined terms describing gene product attributes. The *P* value denotes the significance of GO term enrichment in the differentially expressed mRNA list. Pathway analysis for differentially expressed mRNAs was based on the latest KEGG (Kyoto Encyclopedia of Genes and Genomes) [21] database. For both the GO and pathway analysis, a *P* value of <.05 was considered to be statistically significant.

qRT-PCR Analyses

Total RNA was extracted from the wild-type and dysferlin-deficient myoblast cells using RNeasy Mini Kit (Qiagen) following the manufacturer's instructions. They were then reverse transcribed using a reverse transcription kit (QuantaBio), and qRT-PCR analyses was completed using SYBR Green FastMix (QuantaBio) and StepOnePlus RT-PCR instrument. The sequences of the primers used in this study are given in Table 2. GAPDH was used for normalization, and the FC was calculated using the $2^{-\Delta\Delta Ct}$ method. The results presented are an average from 3 biological replicates.

Table 2. List of primers used for quantitative real-time polymerase chain reaction.

Long noncoding RNA	Primer sequence (5' to 3')
TnnT3	Fwd - AGCTCCAAGCCCTCATTGAC Rev - CTCCTCCTCTCTTCTGGCCT
H19	Fwd - ATCCTGGAGCCAAGCCTCTA Rev - TCACGGGTGCTTTGAGTCTC
XLOC_011052	Fwd - CCAGGAAGTTGAAGCAGGAG Rev - CGGAGAACAATGTGGTGGTA
XLOC_008220	Fwd - TTTCACTTGCTGGCTTTTGA Rev - ACTCCCAGCCAGCTGTGTC
AK085239	Fwd - CCATCCCTACACTGCAGCAA Rev - GTTGGGAAAGCATGGCTGTG
AK032137	Fwd - CCTTGGAGTGAAGTGGCCAT Rev - CTCTCTCCTCCCTTGCCTCT
Trak2	Fwd - CCTAGCTCCGGTTTCCCATC Rev - CGGTTTGTGATGGATTCGCC

Network Analysis

MetaCore software (2021 version; GeneGo Inc) was used for network analysis. Based on the FCs between the knockout (KO) versus wild-type lncRNAs, the top 25 upregulated and top 25 downregulated lncRNAs were selected from the data set (Multimedia Appendix 1 shows the gene symbols [lncRNA identifiers] and FCs). The selected data were uploaded to the software's build network tab under the selection *Build Network for Single Gene/Protein/Compound or a List*. Biological networks were built using the Analyze Network Algorithm with default MetaCore settings. The key parameters used for generating the output are the relative enrichment obtained from the uploaded data and the relative saturation of the networks with canonical pathways.

Results

Differentially Expressed LncRNAs

The microarray revealed that of the 19,744 lncRNAs tested, 3195 were differentially expressed in the dysferlin-deficient cells compared to the wild-type cells. Among these lncRNAs, 1035 were upregulated and 2160 were downregulated (Figure 1). Interestingly, 59 lncRNAs were found to have an FC of greater than 10. The most upregulated lncRNA was humanlincRNA0955 (FC=3077.01), and the most downregulated was AK085239 (FC=110.64). Analyzing the chromosomal distribution of the differentially expressed lncRNAs revealed that they are mostly unequally distributed (Figure 2). There were 2 lncRNAs assigned to the mitochondrial genome. Chromosome 2 contained the largest number of dysregulated lncRNAs (295/3195, 9.2%) and included 80 upregulated and 215 downregulated lncRNAs in the dysferlin-deficient cells compared to wild-type cells.

Figure 1. Differentially expressed lncRNAs based on microarray data. Scatter plot of differentially expressed lncRNAs in dysferlin-deficient murine myoblasts compared to normal controls. Red points represent upregulated lncRNAs and green points represent downregulated lncRNAs in dysferlin-deficiency myoblasts with a fold change greater than 2.0. KO: knockout; LncRNA: long noncoding RNA; WT: wild type.

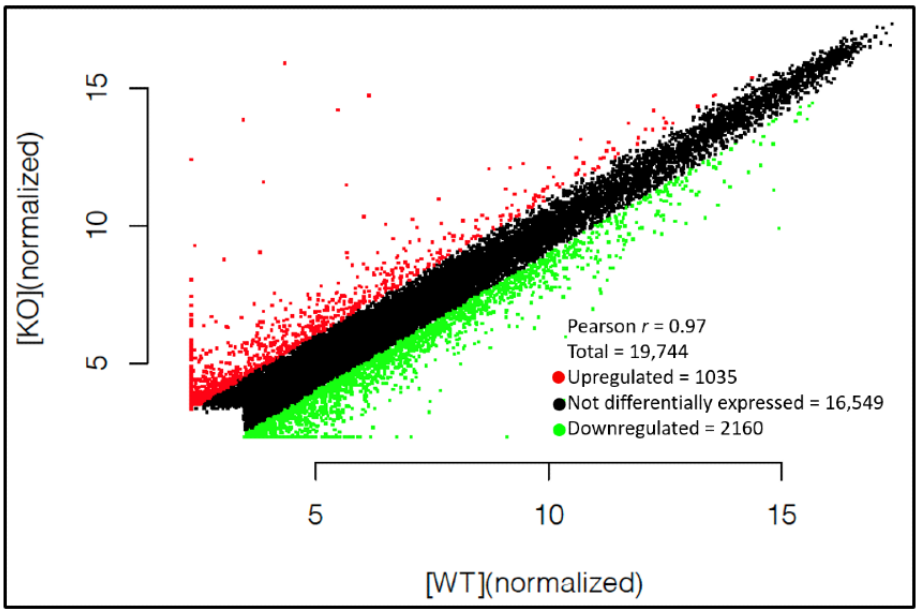
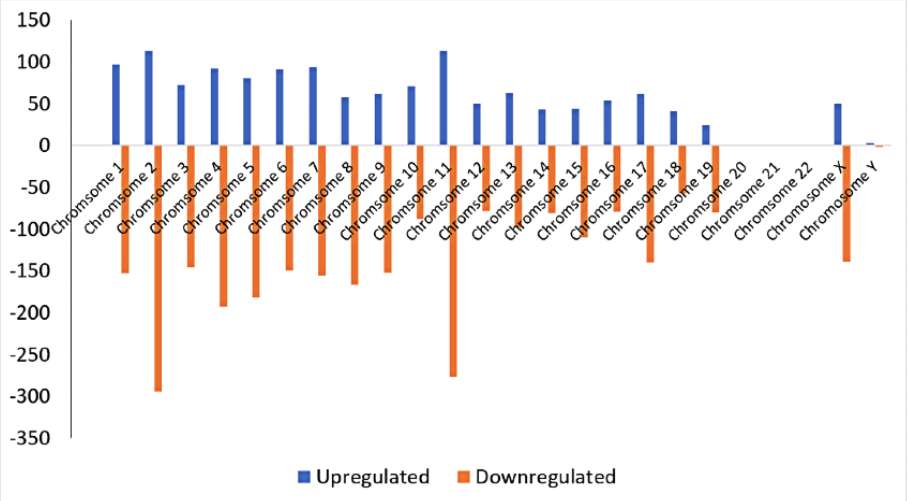


Figure 2. Chromosomal distribution of differentially expressed lncRNAs in dysferlin-deficient myoblasts, showing upregulated (blue) and downregulated (orange) lncRNAs in each chromosome. lncRNA: long noncoding RNA.



LncRNA Classification

LncRNAs are classified based on their relative position to the nearby protein-coding genes. In this study, we identified differentially expressed lncRNAs that were distributed among 4 different categories: sense, intergenic, antisense, and bidirectional. The sense and antisense lncRNAs are transcribed from the sense and antisense strands of the DNA, respectively. Intergenic lncRNAs are derived from DNA sequences between genes, and bidirectional lncRNAs use the same promoter as the protein-coding genes but are transcribed in the opposite direction [22]. The majority of the differentially expressed lncRNAs were sense (total: 1385/3195, 43.3%; upregulated: 272/1035; downregulated: 1113/2160) and intergenic (total: 1224/3195, 38.3%; upregulated: 513/1035; and downregulated: 711/2160) lncRNAs.

Differentially Expressed mRNAs

Of the 15,633 mRNAs screened by the microarray, 1966 were differentially expressed in the dysferlin-deficient cells compared to the wild-type cells. Among the 1966 mRNAs, 1233 were upregulated and 733 were downregulated (Figure 3). A total of 126 mRNAs had an FC greater than 10. The most upregulated mRNA was Gm11565 (FC=422.77), and the most downregulated mRNA was Lce1h (FC=33.92). Around 9.2% (181/1966; 119 upregulated and 62 downregulated) of the probed mRNAs were found on chromosome 7, followed by 8.7% (171/1966) on chromosome 2 (Figure 4).

Figure 3. Differentially expressed mRNAs based on microarray data. Scatter plot of differentially expressed mRNAs in dysferlin-deficient murine myoblasts compared to normal controls. Red points represent upregulated mRNAs and green points represent downregulated lncRNAs in dysferlin-deficiency myoblasts with a fold change greater than 2.0. KO: knockout; mRNA: messenger RNA; WT: wild type.

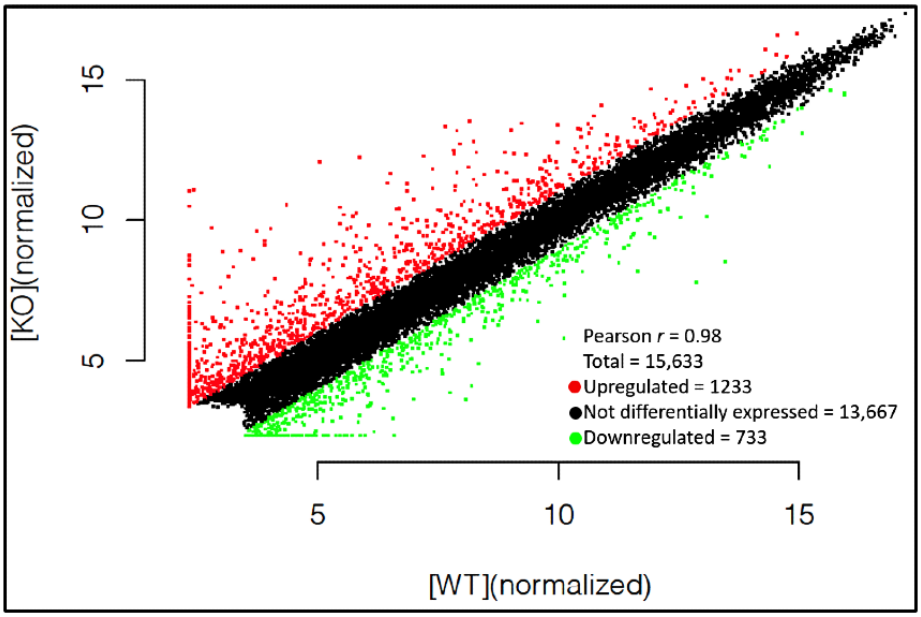
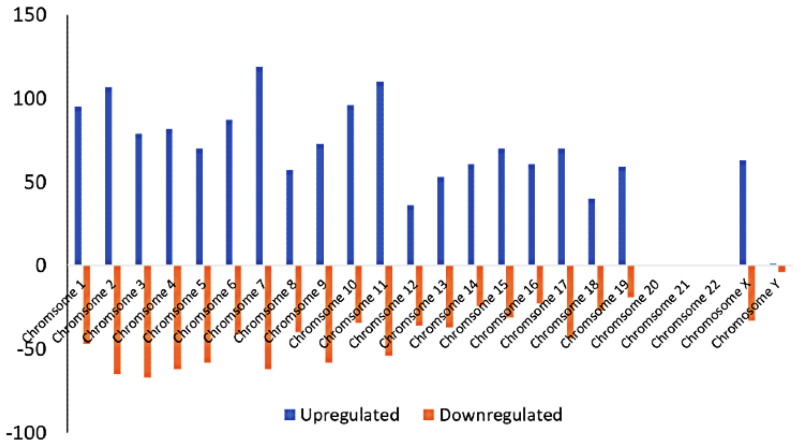


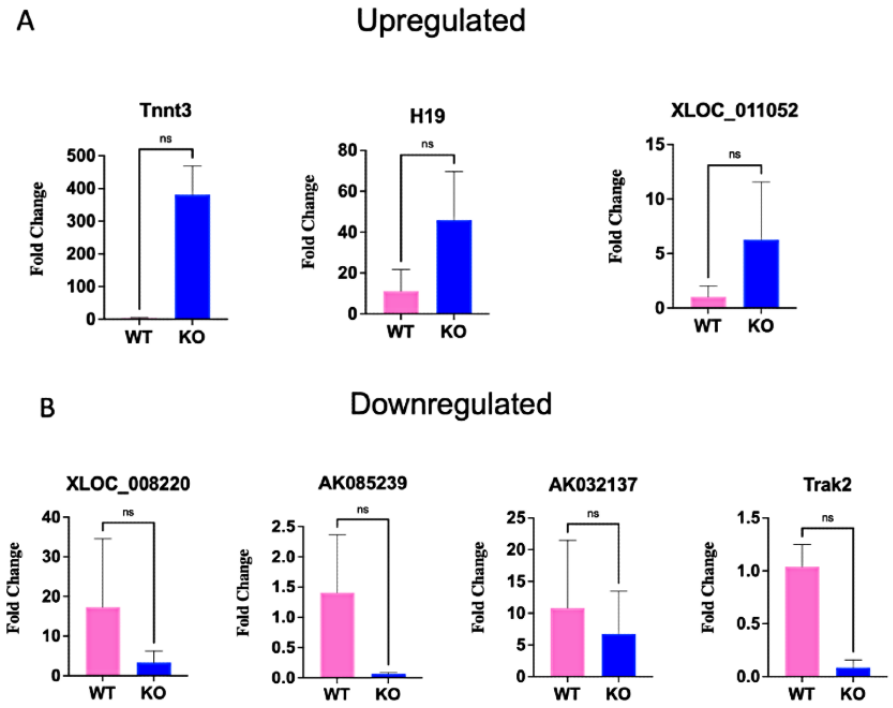
Figure 4. Chromosomal distribution of differentially expressed mRNAs in dysferlin-deficient myoblasts, showing upregulated (blue) and downregulated (orange) mRNAs in each chromosome. mRNA: messenger RNA.



qRT-PCR Validation of Microarray Results

To confirm the reliability of the microarray results, we randomly selected 6 differentially expressed lncRNAs—3 upregulated (TnnT3, H19, and XLOC_011052) and 4 downregulated (XLOC_008220, AK085239, AK032137, and Trak2)—from the microarray results. The analyses' results were consistent with the direction of change of the microarray results (Figure 5).

Figure 5. Quantitative real-time polymerase chain reaction (qRT-PCR) validation of microarray results. The bar graphs show the fold changes in the knockout samples compared to the wild type. Error bars represent SD (N=3). KO: knockout; WT: wild type.



GO and KEGG Pathway Analysis of the Differentially Expressed mRNAs

GO analysis was performed to understand the functional role of the differentially expressed mRNAs (Figures 6-7). The analysis covered 3 domains: biological process, cellular component, and molecular function. The most significantly enriched terms in this study were *protein binding* and *binding*. Additionally, *proteinaceous extracellular matrix*, *extracellular region*, and *cell part* were enriched at the cellular component level, and *single-multicellular organism process*, *muscle system process*, and *developmental process* were significantly enriched at the biological process level. We used enrichment scores to rank the pathways involved in dysferlin deficiency. The top 10 enriched pathways associated with the upregulated and downregulated genes are shown in Figures 6 and 7, respectively.

Figure 6. Gene Ontology (GO) analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis of upregulated genes. mRNA: messenger RNA.

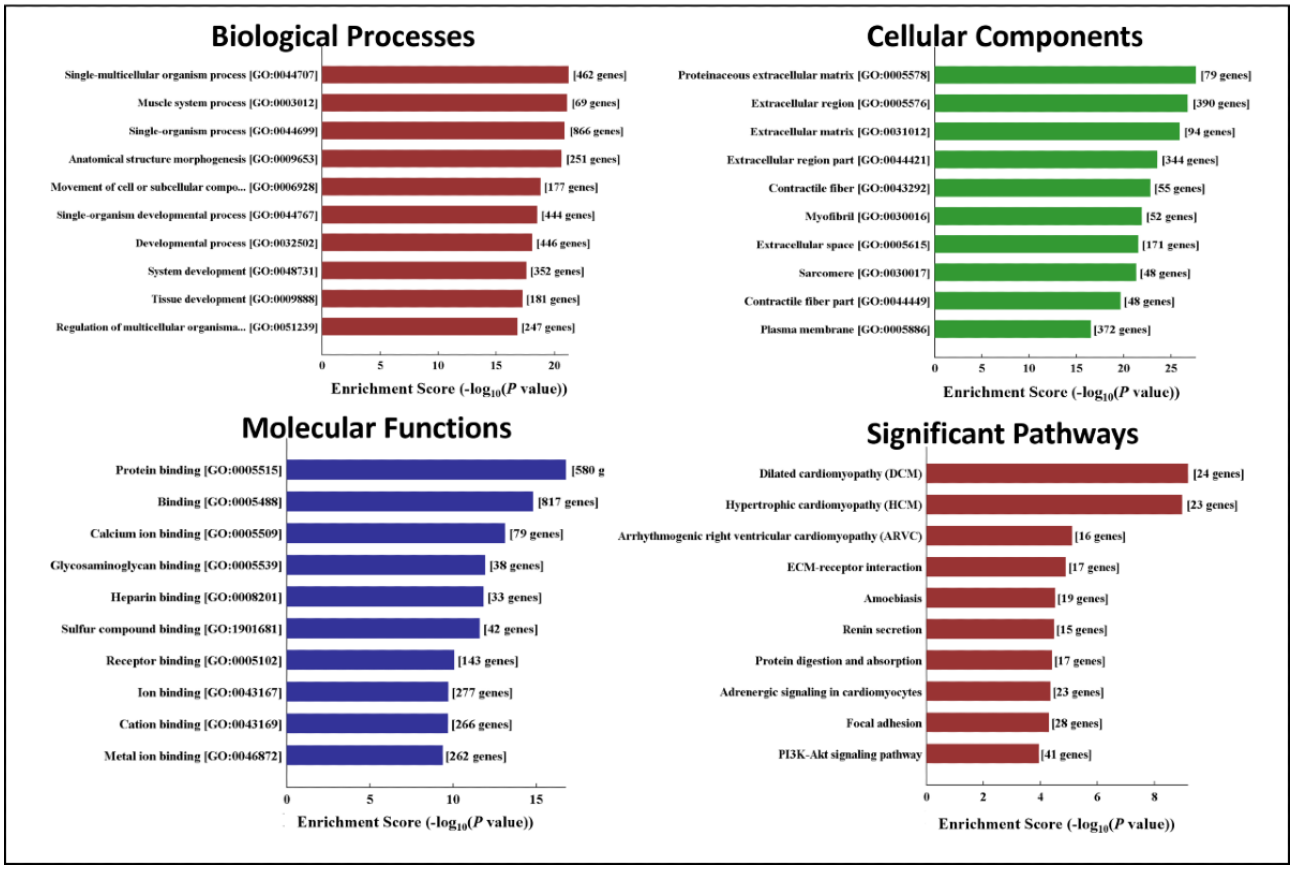
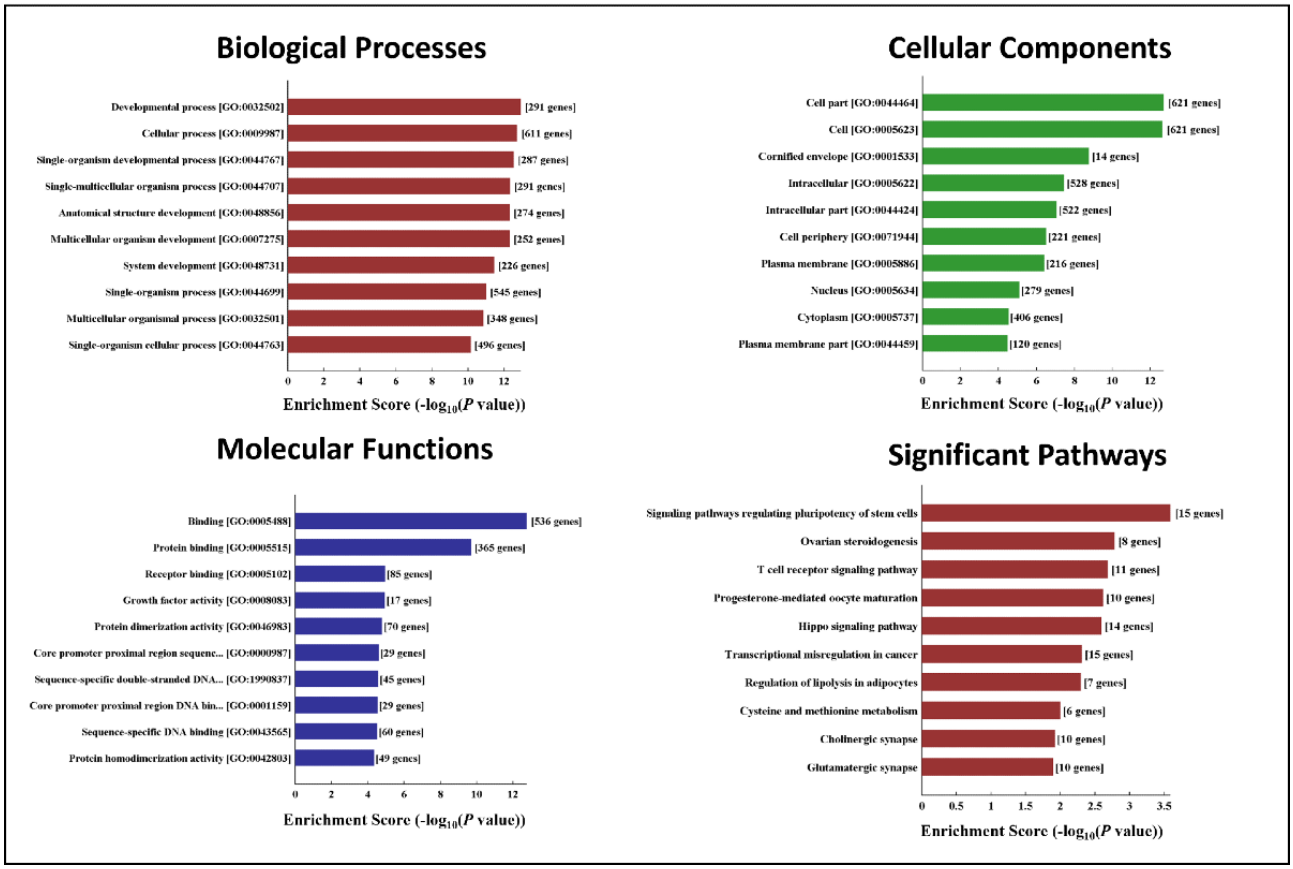


Figure 7. Gene Ontology (GO) analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis of downregulated genes. mRNA: messenger RNA.



Network Analysis

Network analysis on lncRNAs was performed to evaluate their associations with transcription factors, receptors, protein complex, small molecules, and biochemical pathways. The top 25 upregulated lncRNAs (humanlincRNA0955, Trak2, 4930480G23Rik, AK078726, AK037210, AK085419, mouselincRNA0086, Tnnt3, AK038305, Filip1, Gm20485, Fam189a1, AK135257, Myh6, Coro2b, H19, AK045129, Enpep, AK144424, AK143389, Gm5401, AK035065, AK009210, humanlincRNA1720, and AK034241) were selected for network analysis. Of these 25 lncRNAs, 13 lncRNAs (Trak2, 4930480G23Rik, AK085419, Tnnt3, Filip1, Fam189a1, AK135257, Myh6, Coro2b, H19, Enpep, AK009210, and AK034241) were recognized by MetaCore, and these specific lncRNAs were used for network analysis. A total of 8 statistically significant network objects were identified from upregulated lncRNAs (Table 3). The top scored network consisted of input lncRNAs Tnnt3, listed as Beta TnTF (Tnnt3), and MyH6, listed as alpha MHC. The top network also included Troponin cardiac (Tnnt2), Troponin T (Tnnt2), and p300 (Table 3 and Figure 8). Interestingly, the top upregulated network was associated 56.2% with muscle system process and 50% with muscle contraction functions. The top 25 downregulated lncRNAs (AK085239, AK135501, AK032137, mouselincRNA1640, AK005833, XLOC_011793, 5830416P10Rik, Gm15389, AK078320, AK017917, Prl2c5, AK144783, AK155441, AK076675, Vmn2r-ps67, Dnajc2, Atrn, Vwc2l, AK032666, DQ687127, humanlincRNA2050, Gm14879, AK019774, Reps2, and AK041109) were selected for network analysis. Of these 25 lncRNAs, 16 lncRNAs (AK085239, AK135501, AK032137, AK005833, AK078320, AK017917, Prl2c5, AK076675, Vmn2r-ps67, Dnajc2, Atrn, Vwc2l, AK032666, AK019774, Reps2, and AK041109) were recognized by MetaCore, and these specific lncRNAs were used for network analysis. A total of 3 statistically significant network objects were identified from downregulated lncRNAs (Table 4 and Figure 9). The top scored network consisted of input lncRNA Dnajc2, listed as Dnajc25. The top network also included LGR6, HNF4-alpha, MRPL43, and Emi2.

Table 3. Statistically significant networks obtained from 13 upregulated long noncoding RNAs. The sequence of network objects is prioritized based on the number of fragments of canonical pathways on the network.

No., name, Gene Ontology processes (%; *P* value)	Total nodes	Seed nodes	*P* value
1. Troponin T, cardiac, Beta TnTF, Troponin cardiac, alpha-MHC, p300	50	8	7.97×10^{-26}
muscle filament sliding (35.4; 9.832×10^{-35})			
actin-myosin filament sliding (35.4; 1.459×10^{-34})			
muscle system process (56.2; 1.447×10^{-33})			
muscle contraction (50; 6.408×10^{-31})			
actin-mediated cell contraction (35.4; 9.680×10^{-27})			
2. H19, FILIP, AKT1, Tip60, miR-29a-3p	50	2	4.52×10^{-06}
histone H4 acetylation (25; 1.131×10^{-16})			
peptidyl–amino acid modification (50; 1.613×10^{-16})			
histone acetylation (25; 1.170×10^{-13})			
internal peptidyl-lysine acetylation (25; 2.149×10^{-13})			
peptidyl-lysine acetylation (25; 2.712×10^{-13})			
3. NckAP1, c-Myc, PCNT1, CD133, SHB	50	2	5.62×10^{-06}
viral transcription (34; 5.579×10^{-27})			
viral gene expression (34; 7.051×10^{-26})			
peptidyl-lysine modification (34; 2.567×10^{-18})			
viral process (46; 1.024 × 10-17)			
biological process involved in symbiotic interaction (46; 1.493×10^{-16})			
4. c-Myc, POM121, NUP54, FZD9, NUP37	50	1	3.32×10^{-03}
viral transcription (40.4; 8.401×10^{-32})			
viral gene expression (40.4; 1.476×10^{-30})			
intracellular transport of virus (31.9; 1.691×10^{-27})			
transport of virus (31.9; 1.250×10^{-26})			
multiorganism localization (31.9; 2.301×10^{-26})			
5. H19, Cyclin D1, ITGB4, LKB1, GLUT4	50	1	3.39×10^{-03}
response to organic cyclic compound (61.4; 6.074×10^{-19})			
tissue development (68.2; 1.553×10^{-18})			
response to abiotic stimulus (61.4; 3.091×10^{-18})			
gland development (45.5; 6.551×10^{-18})			
positive regulation of cellular metabolic process (79.5; 1.774×10^{-17})			
6. WRC, c-Myc, OTX2, APEX, Folliculin	50	1	3.46×10^{-03}
positive regulation of nitrogen compound metabolic process (77.1; 1.738×10^{-18})			
positive regulation of macromolecule metabolic process (79.2; 5.340×10^{-18})			
positive regulation of cellular process (91.7; 1.152×10^{-17})			
positive regulation of biological process (93.8; 2.262×10^{-17})			

No., name, Gene Ontology processes (%; *P* value)	Total nodes	Seed nodes	*P* value
positive regulation of protein metabolic process (60.4; 2.337×10^{-17})			
7. Y549 (GRIF1), PPARGC1 (PGC1-alpha), OGT (GlcNAc transferase), MMP-9, mTOR	50	1	3.54×10^{-03}
positive regulation of macromolecule metabolic process (87; 7.785×10^{-22})			
positive regulation of DNA-templated transcription (67.4; 1.717×10^{-21})			
positive regulation of nucleic acid–templated transcription (67.4; 4.954×10^{-21})			
positive regulation of RNA biosynthetic process (67.4; 5.018×10^{-21})			
negative regulation of nitrogen compound metabolic process (76.1; 6.518×10^{-21})			
8. Y549 (GRIF1), c-Myc, mRNA intracellular, NUP35, RAE1	50	1	3.54×10^{-03}
viral transcription (44.9; 5.321×10^{-38})			
viral gene expression (44.9; 1.541×10^{-36})			
Signal Recognition Protein (SRP)–dependent cotranslational protein targeting to membrane (28.6; 2.873×10^{-22})			
cotranslational protein targeting to membrane (28.6; 4.992×10^{-22})			
viral process (53.1; 6.767×10^{-22})			

Figure 8. The top scored (by the number of pathways) network obtained from the top 13 upregulated long noncoding RNAs (lncRNAs). Green arrows indicate activation effect, red arrows indicate inhibition effect, and gray arrows indicate unspecified effects. The lncRNAs beta TnTF (Tnnt3) and alpha-MHC (MyH6) are circled in black.

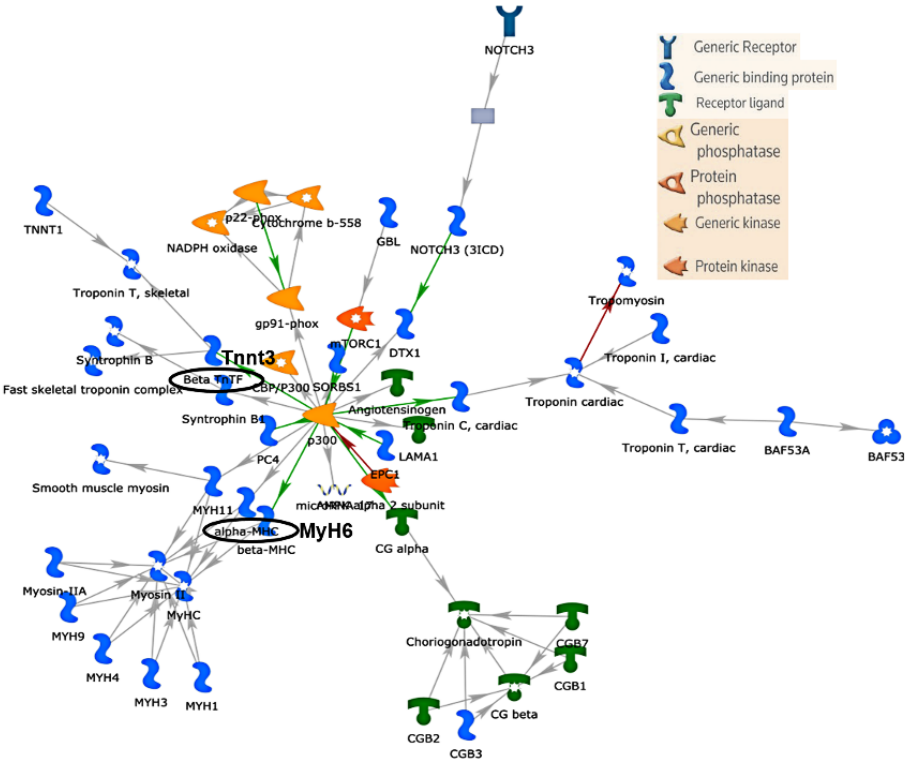
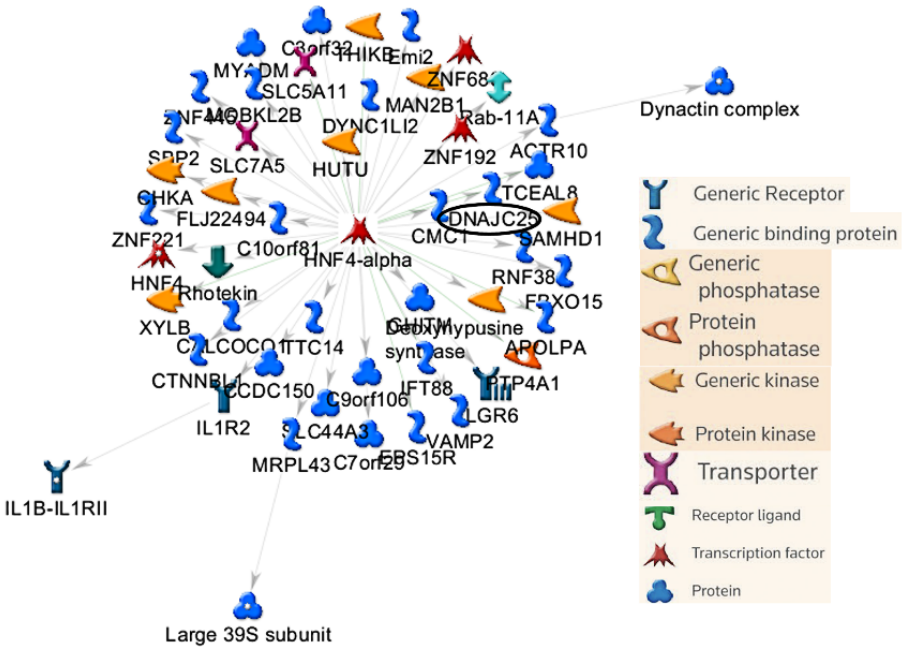


Table 4. Statistically significant networks obtained from 16 downregulated long noncoding RNAs. The sequence of network objects is prioritized based on the number of fragments of canonical pathways on the network.

No., name, Gene Ontology processes (%; *P* value)	Total nodes	Seed nodes	*P* value
1. DNAJC25, LGR6, HNF4-alpha, MRPL43, Emi2	50	2	1.98×10^{-06}
negative regulation of activation of Janus kinase activity (4.7; 9.716×10^{-06})			
retrograde axonal transport of mitochondrion (4.7; 1.941×10^{-05})			
negative regulation of interleukin-1 alpha production (4.7; 3.231×10^{-05})			
purine deoxyribonucleotide catabolic process (4.7; 4.841×10^{-05})			
negative regulation of interleukin-1-mediated signaling pathway (4.7; 6.770×10^{-05})			
2. DNAJC27, GATA-4, ERK1/2, Decanoyl-CoA + Acetyl-CoA = 3-Oxo-dodecanoyl-CoA + CoA, BAF250A	50	1	1.84×10^{-03}
cellular response to organic cyclic compound (50; 4.858×10^{-18})			
cellular response to lipid (47.6; 7.417×10^{-17})			
response to lipid (57.1; 8.792×10^{-17})			
cardiocyte differentiation (31; 1.763×10^{-16})			
response to organic cyclic compound (57.1; 5.534×10^{-16})			
3. MPP11, NANOG, HOXB1, PECAM1, VISA	50	1	2.09×10^{-03}
muscle system process (34; 6.752×10^{-17})			
platelet aggregation (18; 1.312×10^{-13})			
homotypic cell-cell adhesion (18; 1.131×10^{-12})			
regulation of muscle system process (26; 2.361×10^{-12})			
muscle hypertrophy (16; 4.725×10^{-12})			

Figure 9. The top scored (by the number of pathways) network from the top 16 downregulated long noncoding RNAs (lncRNAs). Gray arrows indicate unspecified effects. The lncRNA DNAJC25 is circled in black.



Discussion

Abnormal expression of specific lncRNAs have been described in various diseases including certain types of muscular dystrophies, such as Duchenne, myotonic, and facioscapulohumeral muscular dystrophies [17,23,24]. However, to date, there is no information on the lncRNAs associated with dysferlinopathy or any type of LGMD. This study is the first to screen and report the difference in the expression patterns of lncRNAs and mRNAs in wild-type and dysferlin-deficient myoblasts. Thereafter, we analyzed the microarray results and performed bioinformatics analysis to understand their possible interactions and functions.

Our results show that there were a high number of lncRNAs and mRNAs that were differentially expressed due to dysferlin deficiency. There were more downregulated than upregulated lncRNAs, but more upregulated than downregulated mRNAs. We validated the microarray results by testing the expression of several lncRNAs, which were consistent for the direction of change, although there were small inconsistencies in the magnitude of FCs. This is expected due to the differences between the 2 methods, the different normalization strategies used, and their inherent pitfalls.

The chromosomal distribution of the differentially expressed lncRNAs and mRNAs were not equal. Chromosomes 2 and 11 had a greater percentage of lncRNAs, whereas chromosomes 2 and 7 had a greater percentage of mRNAs that were differentially expressed. It is interesting to note that the dysferlin gene is located in chromosome 2 in humans and in chromosome 6 in mice. It is tempting to posit that at least some of the differentially expressed lncRNAs and mRNA gene products may directly regulate the expression or function of the dysferlin protein.

According to their position and directionality of transcription in relation to other genes, lncRNAs can be classified into multiple subgroups such as sense lncRNAs, antisense lncRNAs, bidirectional lncRNAs, and long-intergenic noncoding RNAs (lincRNAs). Interestingly, a majority of the differentially expressed lncRNAs identified in this study are sense lncRNAs (43.3%) or intergenic lncRNAs (38.3%). Together, they contribute to more than three-quarters of the total lncRNAs that were differentially expressed. LincRNAs are ncRNAs that are transcribed from regions nearby the protein-coding genes without overlapping them. They are known to be highly tissue-specific and can regulate the nearby protein-coding genes and genes far away from them [25]. Sense lncRNAs are those that are transcribed from the sense strand of DNA, and they may overlap or contain an entire protein-coding gene sequence within them. The differential expression of these 2 types of lncRNAs suggests that these lncRNAs may be involved in regulating the protein-coding genes that are involved in the progression of dysferlinopathy. Since the lincRNAs are more tissue-specific, future studies may focus on evaluating any correlation between their expression and tissue involvement or disease severity in dysferlinopathy.

As lncRNAs are regulatory molecules, the differentially expressed lincRNAs and sense lncRNAs could possibly control the expression of the nearby or overlapping genes through multiple mechanisms. Hence, it is tempting to predict that the function of many of the novel lncRNAs identified in this study could be related to the function of the associated genes. From our GO and pathway analysis, some of the differentially expressed mRNAs were functionally related to skeletal muscle contraction (16 genes), skeletal muscle relaxation (12 genes), regulation of muscle contraction (5 genes), and actin-myosin filament sliding (5 genes). The related signaling pathways included the PI3K-Akt signaling pathway (41 genes), Hippo signaling pathway (14 genes), and pathways that control the pluripotency of stem cells (15 genes).

This study has several limitations. The dysferlin-deficient murine myoblasts used in this study had dysferlin silenced using a short hairpin RNA (shRNA). Hence, these cells could have low-level dysferlin expression that may have influenced the differential expression. Since there is very minimal prior research in this area, the significance and interactions of the differentially expressed lncRNAs are only predicted and not confirmed. Additionally, the microarray was performed using pooled RNA from 3 biological replicates, and therefore, the statistical significance for the FCs could not be calculated. To note, puromycin is known to cause changes in gene expression [26,27] and cellular stress [28,29] in mammalian cell lines. Hence, its addition to the KO cells may have contributed to the differential lncRNA and mRNA expression in the myoblasts tested. Although the shRNA construct that was used to establish the KO cell line carried the puromycin acetyltransferase (pac) gene and is expected to confer resistance to puromycin, the efficiency can vary depending on the cell lines [30]. Finally, since the study has been conducted using murine myoblasts, the differentially expressed lncRNA signature may differ in human-derived, dysferlin-deficient myoblasts.

In conclusion, this is the first report illustrating the lncRNA signature in dysferlinopathy. Our results highlight that lncRNAs are involved in regulating the gene expression and critical biological functions such as muscle contraction and relaxation in dysferlinopathy. Future studies focusing on deciphering the exact mechanisms that contribute to the regulatory function of these lncRNAs will be interesting and add more to our understanding of this incurable disease.

Acknowledgments

The study was funded by grants from the Faculty Development Grant Committee and the Summer Undergraduate Research Program at the Massachusetts College of Pharmacy and Health Sciences University in Boston. We would like to thank Dr Robert H Brown, Chair of Neurology at the University of Massachusetts Medical School, for kindly providing us the dysferlin-deficient and wild-type murine myoblasts; Dr Yanggu Shi, Senior Scientist at Arraystar Inc, for his support with the bioinformatics analysis;

and the Department of Physician Assistant Studies at the Massachusetts College of Pharmacy and Health Sciences University in Boston and the Jain Foundation for their encouragement and support of dysferlinopathy research.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Gene symbols and fold changes.
[PDF File (Adobe PDF File), 34 KB - bioinform_v3i1e33186_app1.pdf]

References

1. Glover L, Brown RH. Dysferlin in membrane trafficking and patch repair. Traffic 2007 Jul;8(7):785-794 [FREE Full text] [doi: 10.1111/j.1600-0854.2007.00573.x] [Medline: 17547707]
2. Gallardo E, Rojas-García R, de Luna N, Pou A, Brown R, Illa I. Inflammation in dysferlin myopathy: immunohistochemical characterization of 13 patients. Neurology 2001 Dec 11;57(11):2136-2138. [doi: 10.1212/wnl.57.11.2136] [Medline: 11739845]
3. Matsuda C, Kameyama K, Tagawa K, Ogawa M, Suzuki A, Yamaji S, et al. Dysferlin interacts with affixin (beta-parvin) at the sarcolemma. J Neuropathol Exp Neurol 2005 Apr 01;64(4):334-340. [doi: 10.1093/jnen/64.4.334] [Medline: 15835269]
4. Bashir R, Britton S, Strachan T, Keers S, Vafiadaki E, Lako M, et al. A gene related to Caenorhabditis elegans spermatogenesis factor fer-1 is mutated in limb-girdle muscular dystrophy type 2B. Nat Genet 1998 Sep;20(1):37-42. [doi: 10.1038/1689] [Medline: 9731527]
5. Liu J, Aoki M, Illa I, Wu C, Fardeau M, Angelini C, et al. Dysferlin, a novel skeletal muscle gene, is mutated in Miyoshi myopathy and limb girdle muscular dystrophy. Nat Genet 1998 Sep;20(1):31-36. [doi: 10.1038/1682] [Medline: 9731526]
6. Narayanaswami P, Weiss M, Selcen D, David W, Raynor E, Carter G, Guideline Development Subcommittee of the American Academy of Neurology. Neurology 2014 Oct 14;83(16):1453-1463 [FREE Full text] [doi: 10.1212/WNL.0000000000000892] [Medline: 25313375]
7. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. Nature 2012 Sep 06;489(7414):101-108 [FREE Full text] [doi: 10.1038/nature11233] [Medline: 22955620]
8. Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. Annu Rev Biochem 2012 Jul 07;81(1):145-166 [FREE Full text] [doi: 10.1146/annurev-biochem-051410-092902] [Medline: 22663078]
9. Tu S, Yuan G, Shao Z. The PRC2-binding long non-coding RNAs in human and mouse genomes are associated with predictive sequence features. Sci Rep 2017 Jan 31;7(1):41669 [FREE Full text] [doi: 10.1038/srep41669] [Medline: 28139710]
10. Cabianca D, Casa V, Bodega B, Xynos A, Ginelli E, Tanaka Y, et al. A long ncRNA links copy number variation to a polycomb/trithorax epigenetic switch in FSHD muscular dystrophy. Cell 2012 May 11;149(4):819-831 [FREE Full text] [doi: 10.1016/j.cell.2012.03.035] [Medline: 22541069]
11. Wang KC, Chang HY. Molecular mechanisms of long noncoding RNAs. Mol Cell 2011 Sep 16;43(6):904-914 [FREE Full text] [doi: 10.1016/j.molcel.2011.08.018] [Medline: 21925379]
12. Paci P, Colombo T, Farina L. Computational analysis identifies a sponge interaction network between long non-coding RNAs and messenger RNAs in human breast cancer. BMC Syst Biol 2014 Jul 17;8(1):83 [FREE Full text] [doi: 10.1186/1752-0509-8-83] [Medline: 25033876]
13. Cesana M, Cacchiarelli D, Legnini I, Santini T, Sthandier O, Chinappi M, et al. A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. Cell 2011 Oct 14;147(2):358-369 [FREE Full text] [doi: 10.1016/j.cell.2011.09.028] [Medline: 22000014]
14. Yoon JH, Abdelmohsen K, Gorospe M. Posttranscriptional gene regulation by long noncoding RNA. J Mol Biol 2013 Oct 09;425(19):3723-3730 [FREE Full text] [doi: 10.1016/j.jmb.2012.11.024] [Medline: 23178169]
15. Zhang M, Jiang Y, Guo X, Zhang B, Wu J, Sun J, et al. Long non-coding RNA cardiac hypertrophy-associated regulator governs cardiac hypertrophy via regulating miR-20b and the downstream PTEN/AKT pathway. J Cell Mol Med 2019 Nov 29;23(11):7685-7698 [FREE Full text] [doi: 10.1111/jcmm.14641] [Medline: 31465630]
16. Wang C, Peng J, Chen X. LncRNA-CIR promotes articular cartilage degeneration in osteoarthritis by regulating autophagy. Biochem Biophys Res Commun 2018 Nov 02;505(3):692-698. [doi: 10.1016/j.bbrc.2018.09.163] [Medline: 30292414]
17. Vizoso M, Esteller M. The activatory long non-coding RNA DBE-T reveals the epigenetic etiology of facioscapulohumeral muscular dystrophy. Cell Res 2012 Oct 19;22(10):1413-1415 [FREE Full text] [doi: 10.1038/cr.2012.93] [Medline: 22710800]
18. Watts R, Johnsen VL, Shearer J, Hittel DS. Myostatin-induced inhibition of the long noncoding RNA Malat1 is associated with decreased myogenesis. Am J Physiol Cell Physiol 2013 May 15;304(10):C995-1001 [FREE Full text] [doi: 10.1152/ajpcell.00392.2012] [Medline: 23485710]

19. Belanto JJ, Diaz-Perez SV, Magyar CE, Maxwell MM, Yilmaz Y, Topp K, et al. Dexamethasone induces dysferlin in myoblasts and enhances their myogenic differentiation. Neuromuscul Disord 2010 Feb;20(2):111-121. [doi: 10.1016/j.nmd.2009.12.003] [Medline: 20080405]
20. The gene ontology resource. The Gene Ontology Consortium. URL: http://www.geneontology.org/ [accessed 2022-05-17]
21. KEGG: Kyoto Encyclopedia of Genes and Genomes. Kanehisa Laboratories. URL: http://www.genome.jp/kegg [accessed 2021-01-08]
22. Lanzafame M, Bianco G, Terracciano L, Ng C, Piscuoglio S. The role of long non-coding RNAs in hepatocarcinogenesis. Int J Mol Sci 2018 Feb 28;19(3):682 [FREE Full text] [doi: 10.3390/ijms19030682] [Medline: 29495592]
23. Zhang Y, Li Y, Hu Q, Xi Y, Xing Z, Zhang Z, et al. The lncRNA H19 alleviates muscular dystrophy by stabilizing dystrophin. Nat Cell Biol 2020 Nov 26;22(11):1332-1345 [FREE Full text] [doi: 10.1038/s41556-020-00595-5] [Medline: 33106653]
24. Wheeler TM, Leger AJ, Pandey SK, MacLeod AR, Nakamori M, Cheng SH, et al. Targeting nuclear RNA for in vivo correction of myotonic dystrophy. Nature 2012 Aug 02;488(7409):111-115 [FREE Full text] [doi: 10.1038/nature11362] [Medline: 22859208]
25. Ulitsky I, Bartel DP. LincRNAs: genomics, evolution, and mechanisms. Cell 2013 Jul 03;154(1):26-46 [FREE Full text] [doi: 10.1016/j.cell.2013.06.020] [Medline: 23827673]
26. Miyata S, Saku N, Akiyama S, Javaregowda PK, Ite K, Takashima N, et al. Puromycin-based purification of cells with high expression of the cytochrome P450 CYP3A4 gene from a patient with drug-induced liver injury (DILI). Stem Cell Res Ther 2022 Jan 10;13(1):6 [FREE Full text] [doi: 10.1186/s13287-021-02680-4] [Medline: 35012658]
27. Burke JF, Mogg AE. Suppression of a nonsense mutation in mammalian cells in vivo by the aminoglycoside antibiotics G-418 and paromomycin. Nucleic Acids Res 1985 Sep 11;13(17):6265-6272 [FREE Full text] [doi: 10.1093/nar/13.17.6265] [Medline: 2995924]
28. Rincon J, Romero M, Viera N, Pedreañez A, Mosquera J. Increased oxidative stress and apoptosis in acute puromycin aminonucleoside nephrosis. Int J Exp Pathol 2004 Feb;85(1):25-33. [doi: 10.1111/j.0959-9673.2004.0368.x] [Medline: 15113391]
29. Min S, Ha D, Ha T. Puromycin aminonucleoside triggers apoptosis in podocytes by inducing endoplasmic reticulum stress. Kidney Res Clin Pract 2018 Sep 30;37(3):210-221 [FREE Full text] [doi: 10.23876/j.krcp.2018.37.3.210] [Medline: 30254845]
30. de la Luna S, Ortín J. pac gene as efficient dominant marker and reporter gene in mammalian cells. Methods Enzymol 1992;216:376-385. [doi: 10.1016/0076-6879(92)16035-i] [Medline: 1479910]

Abbreviations

cRNA: complementary RNA
FC: fold change
GO: Gene Ontology
KEGG: Kyoto Encyclopedia of Genes and Genomes
KO: knockout
LGMD: limb-girdle muscular dystrophy
lincRNA: long-intergenic noncoding RNA
lncRNA: long noncoding RNA
mRNA: messenger RNAs
ncRNA: noncoding RNAs
qRT-PCR: quantitative real-time polymerase chain reaction

Edited by A Mavragani; submitted 27.08.21; peer-reviewed by K Rathi, Anonymous; comments to author 06.10.21; revised version received 02.01.22; accepted 10.05.22; published 17.06.22.

Please cite as:
Singhal R, Lukose R, Carr G, Moktar A, Gonzales-Urday AL, Rouchka EC, Vajravelu BN
Differential Expression of Long Noncoding RNAs in Murine Myoblasts After Short Hairpin RNA-Mediated Dysferlin Silencing In Vitro: Microarray Profiling
JMIR Bioinform Biotech 2022;3(1):e33186
URL: https://bioinform.jmir.org/2022/1/e33186
doi:10.2196/33186
PMID:

Original Paper

Identification of a Novel c.3080delC JAG1 Gene Mutation Associated With Alagille Syndrome: Whole Exome Sequencing

Deepak Panwar[1], PhD; Vandana Lal[1], MD; Atul Thatai[1,2], PhD

[1]Molecular Diagnostic Division, National Reference Laboratory, Dr Lal Pathlabs, New Delhi, India

[2]Molecular and Cytogenomics Division, Max Specialty Hospital, New Delhi, India

Corresponding Author:
Atul Thatai, PhD
Molecular Diagnostic Division
National Reference Laboratory
Dr Lal Pathlabs
Block E, Sector 18, Rohini
New Delhi, 110085
India
Phone: 91 1139885050
Email: atul_thatai@hotmail.com

Abstract

Background: Alagille syndrome is an autosomal dominant disorder associated with variable clinical phenotypic features including cholestasis, congenital heart defects, vertebral defects, and dysmorphic facies.

Objective: Whole exome sequencing (WES) has become technically feasible due to the recent advances in next-generation sequencing technologies, therefore offering new possibilities for mutations or genes identification.

Methods: WES was used to identify pathogenic variants, which may have significant prognostic implications for patients' clinical presentation of the proband. In this paper, we have uncovered a novel *JAGGED1* gene (*JAG1*) mutation associated with Alagille syndrome in a 5-year-old girl presented with conjugated hyperbilirubinemia and infantile cholestasis.

Results: The exome sequencing analysis revealed the presence of a novel *JAG1* heterozygous c.3080delC variant in exon 25. The detected variant introduced a stop codon (p.P1027RfsTer9) in the gene sequence, encoding a truncated protein. Our exome observations were confirmed through Sanger sequencing as well.

Conclusions: Here, we report a case of a patient diagnosed with Alagille syndrome, conjugated hyperbilirubinemia, and infantile cholestasis, with emphasis on its association with the detection of the novel *JAG1* mutation, thereby establishing the genetic diagnosis of the disease.

(JMIR Bioinform Biotech 2022;3(1):e33946) doi:10.2196/33946

KEYWORDS

Alagille syndrome; JAG1; stop codon mutation; whole exome sequencing; gene database; clinical database; genome; genetics; sequence technology; diagnostic tool; genetic diagnosis; gene sequencing

Introduction

Alagille syndrome (ALGS) is an autosomal dominant and multisystemic congenital disorder causing pediatric chronic liver disease with the prevalence of 1: 70,000-100,000 in infants [1,2]. Alagille syndrome (ALGS; MIM: 118450) is characterized by intrahepatic bile ducts, highly variable clinical features, including cholestasis, skeletal malformations, cardiac, ocular abnormalities, and dysmorphic facial features [1]. The classical diagnosis of ALGS includes low numbers of hepatic bile ducts, which results in chronic cholestasis leading to cirrhosis and end-stage liver disease. Hepatic manifestations are variable, and some patients present with jaundice, though it does not progress to a more serious disease [3].

ALGS is caused by mutations in either the *JAGGED1* (*JAG1*) or *NOTCH2* gene. Both these genes are involved in the Notch signaling pathway and play an important role in transcription regulation and cell fate determination [4-7]. The majority of ALGS cases (~97%) are caused by mutations in *JAG1* (MIM: 601920) gene, while less than 1% of patients have a heterozygous mutation in the *NOTCH2* gene (1p13) [8,9]. The *JAG1* gene is located on chromosome 20 (20p12.2),

encompassing 26 exons that encode a 1218 amino acid protein that participates in the Notch signaling pathway as a ligand [10]. Phenotypic effects of JAG1 mutation in ALGS are highly variable with reduced penetrance. However approximately 94% of patients with a clinically confirmed diagnosis of ALGS carry JAG1 mutations [11]. There is a high rate of de novo mutations, with approximately 60%-70% of mutations in probands not found in either parent [12-15].

It has been reported that the pathogenic mutations of in *JAG1* include missense mutations (11%), nonsense and frame-shift mutations (69%), splice site mutations (16%), and deletion of the entire *JAG1* gene (4%) [12,14,16-23]. *JAG1* mutations in ALGS clinical presentation have been reported in various populations, such as American, European, Australian, and Japanese [12,14,16-23], whereas there are only few clinical studies on ALGS from India [24-26]. Because of the wide range of clinical manifestations, early genetic testing is required to establish the condition and to take preventative steps to avoid consequences in numerous organs. The advent of molecular diagnostic testing has led to a revision of diagnostic criteria for ALGS [1]. Next-generation sequencing analysis, including either genome or exome sequences, have been recommended for the molecular diagnosis of neonatal or infantile intrahepatic cholestasis [22]. Whole exome sequencing (WES) allows sequencing of all expressed genes in the genome, which is substantial, considering the protein-coding regions cover approximately 85% of human disease-causing mutations [27]. In this study, using WES, we identified a novel *JAG1* mutation associated with early onset of Alagille syndrome.

Methods

Case Presentation

The proband is a 5-year-old girl presented with conjugated hyperbilirubinemia and infantile cholestasis with the onset of clinical manifestation of features at 10 months of age.

Ethics Approval

The study design and protocol were conducted in accordance with the guidelines of the American College of Medical Genetics and Genomics and was approved by the Ethical Review Committee of Dr Lal Pathlabs. Written informed consent has been taken from parents of the proband included in the study, and the parents have provided consent to publish the data.

Library Preparation and WES

The DNA was extracted from 2 ml of the peripheral blood using Qiagen DNA mini kit, as per the manufacturer's instructions. The quantity and quality of the extracted genomic DNA were measured by NanoDrop-2000 Spectrophotometer. Approximately 100 ng of genomic DNA was used to construct exome library using Ion Ampliseq Exome RDY Panel kit (Thermo Fisher Scientific). The resulting DNA library was quantified with Qubit dsDNA HS (High Sensitivity) Assay Kit on Qubit 3.0 Fluorometer. Approximately 25 pm of the library was used with the Ion Chef Instrument (Thermo Fisher Scientific) for template generation followed by enrichment of the templated ion sphere particles. Sequencing was performed using Hi-Q chemistry on Ion Proton system (Thermo Fisher Scientific).

Data Processing and Variant Analysis

The sequences were aligned against the reference genome (GRCh37/hg19) in Torrent Suite v.5.12.0 and Torrent Suite Variant Caller v.5.2.1 software (Thermo Fisher Scientific) with default parameters. The coverage analysis plugin and variant caller plugin from Life Technologies (Thermo Fisher) were used to analyze the Ion Proton sequencing run. Variant discovery, genotype calling of multiallelic substitutions, and indels were performed on each individual sample using the Torrent Variant Caller (TVC, version 4.6.0.7; Thermo Fisher). Statistics and graphs describing the level of sequence coverage produced for targeted genomic regions were provided by the Torrent Coverage Analysis (version 4.6.0.3). The variants were annotated by the Annotate variants 5.0 of Ion Reporter (Thermo Fisher).

Variant Prioritization and Bioinformatics Analysis

Variants that were detected in the exome sequencing were filtered based on coverage (≥15x), minor allele frequency (≤0.01), and deleterious potential. All resulting variants were contrasted with the Human Gene Mutation Database [28] and Uniprot [29]. Furthermore, intronic, up- or downstream, and synonymous variants were removed. The pathogenicity of the detected variant was evaluated using Mutation Taster [30] and MutPredLOF [31]. Additional factors that were considered include the following: (1) absence in the general population; (2) novel appearance and disease phenotype from the family pedigree; (3) absence of any other mutation in *JAG1* that could be responsible for the clinical phenotype; and (4) previous independent occurrence in an unrelated patient. An Integrative Genome Viewer [32] was used to visualize sequencing data. Variant frequencies were obtained from various databases such as the 1000 Genomes Project, dbSNP142, Exome Aggregation Consortium (ExAC) and gnomAD. Finally, for the interpretation of variant, American College of Medical Genetics and Genomics 2015 guidelines were used [33].

Mutation Confirmation: Sanger Sequencing

Confirmation of the mutation was performed by conventional Sanger sequencing using the BigDye Terminator v3.1 Cycle Sequencing kit (Applied Biosystems, Thermo Fisher Scientific) and was loaded on an ABI 3500Dx automated Genetic Analyzer (Applied Biosystems, Thermo Fisher Scientific). Primer sequences for the identified variant were designed using Primer 3.0 online as follows: Frd 5′-CCTCATTATTCGATGGCAAGGC -3′ and Rev 5′-GTTCTGTTCTTCAGAGGCCG-3′.

Results

Whole Exome Sequencing Analysis

We detected a total of 39,679 variants comprising 55% (n=21,823) synonymous, 43% (n=17,062) missense, and 2% (n=794) frameshift or indel variants. For WES data filtering procedures, a filtering tree illustrated the step-by-step narrowing down of candidate gene or variants detected during

next-generation sequencing data analysis (Figure 1). The first phase consisted of benign and synonymous variant filtering, and the second phase was based on variant impact (nonsynonymous and truncating), allele frequency (<0.1%), and pathogenicity prediction tools for missense variants (score >3). Since there were still a high number of candidate variants and genes, a second round of prioritization based on manual curation of biological function was performed, and variants in genes unrelated to ALGS were filtered out.

WES results indicated a novel heterozygous frameshift variant (c.3080delC;p.P1027RfsTer9) in the *JAG1* gene responsible for ALGS (Figure 2A). The sequence alignment of heterozygous deletion (c.3080delC) at position Chr20:10621549 in *JAG1* gene was viewed using the Integrative Genomics Viewer (Figure 2B). Sanger sequencing analysis confirmed that the proband carried the mutation in a heterozygous state (Figure 2C). This mutation is conserved across different species and can greatly affect the amino acid sequence of *JAG1* gene that might change the protein function. Different web-based bioinformatics tools were used to analyze the pathogenicity of the variant and predicted this mutation to have potential damaging effects (Table 1).

Figure 1. Illustration for the variant filtering process in whole exome sequencing . ExAC: Exome Aggregation Consortium; MAF: mutation annotation format. VCF: variant call format.

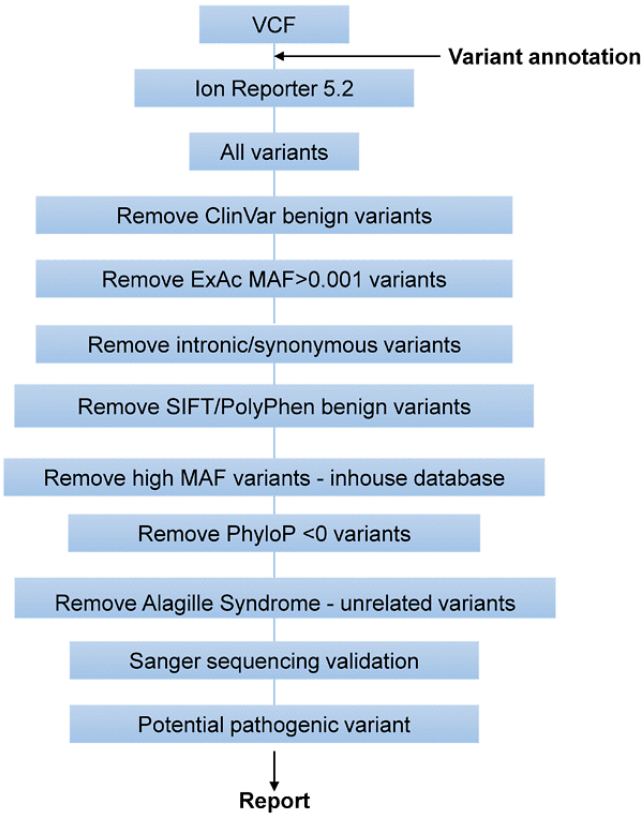


Figure 2. Proband's pedigree and electropherogram of identified disease associated variant. (A) Family pedigree of patient diagnosed with Alagille Syndrome. (B) IGV plot showing the mutation region in WES data in the proband. Track comprises two parts: a histogram of the read depth and the reads as aligned to the reference sequence. Reads are colored according to the aligned strand (red=forward strand; blue=reverse strand). (C) Sanger sequencing confirmation of heterozygous *JAG1* variant c.3080delC in the patient. A: adenine; C: cytosine; DEL: deletion; G: guanine; JAGF: *JAG1* forward primer; JAGR: *JAG1* reverse primer; T: thymine.

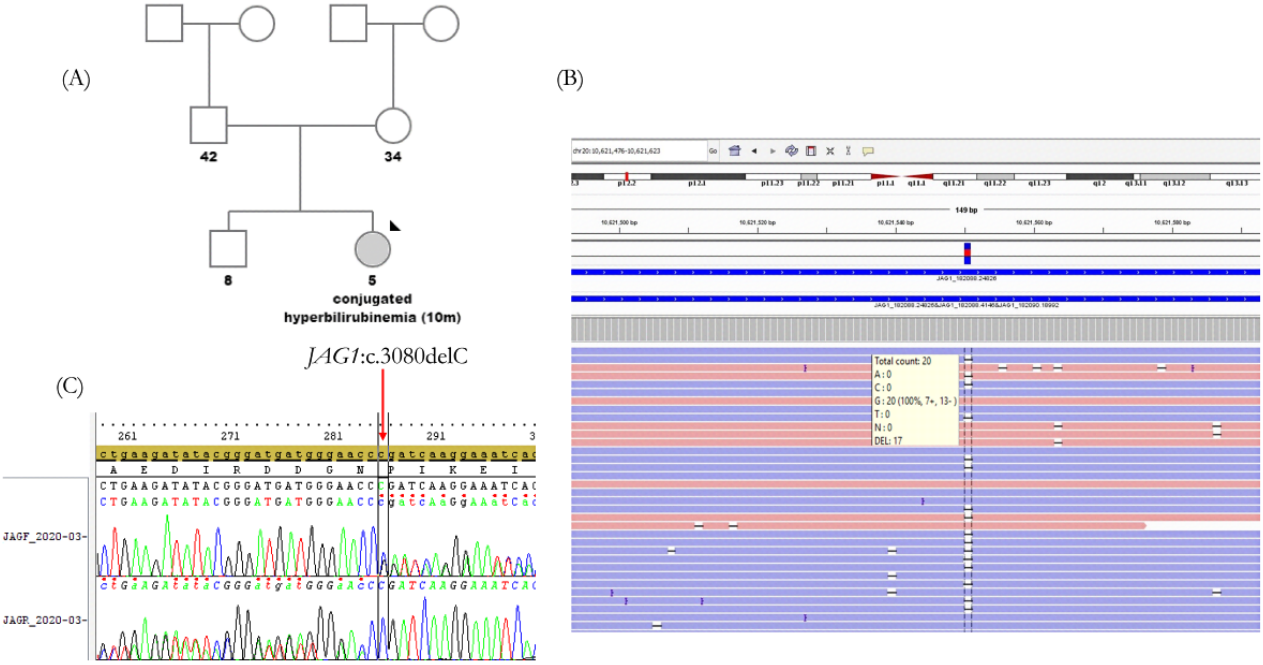


Table 1. Whole exome sequencing analysis identified the *JAGGED1* gene (*JAG1*) mutation in the proband.

Locus	Gene	Exon	Protein	Coding	Mutation Taster[a]	MutPredLOF[a]
Chr20:10621549	*JAG1*	25	p.Pro1027fs	c.3080delC	D[b]	D

[a]Mutation Taster and MutPredLOF are functional prediction scores in which increasing values indicate a more damaging effect.

[b]D: damaging or deleterious.

Discussion

Principal Findings

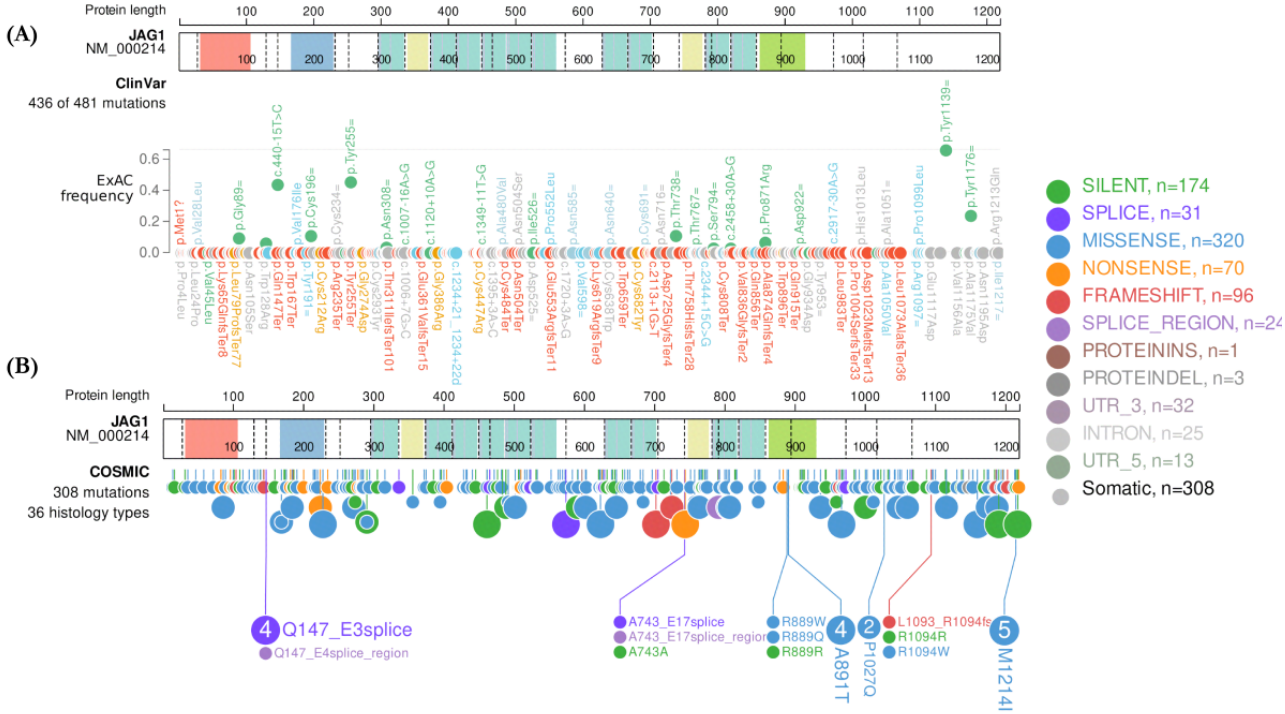
ALGS is a highly variable autosomal dominant disorder, which involves multiple organ systems, and it requires a multidisciplinary team of medical specialists for its management [33-35]. The spectrum of mutations in *JAG1* gene associated with ALGS includes full gene deletion and other protein-truncating mutations including nonsense, frameshift, and splice site as well as missense mutations, suggesting that the clinical phenotype is caused by haploinsufficiency for the *JAG1* protein [9,15]. Phenotypic effects of *JAG1* mutations are highly penetrant but with variable expressivity [36]. There is no strong correlation between the type and location of the *JAG1* mutation and the severity of the disease, suggesting that other genomic modifiers beyond the known *JAG1* mutation may be the cause of the variable expressivity that characterizes this disorder [9]. Unfortunately, no genotype-phenotype correlation exists between clinical manifestations and the specific *JAG1* pathogenic variant or the location of the mutation within gene [37]. Although genetics of ALGS is well-defined, there is variable expressivity of the disease. Individuals with the same mutations, including patients belonging to the same family, show discordance in the phenotype [38]. In support of this concept, the genotype-phenotype correlation studies did not identify a link between the mutation type and clinical manifestation or severity [39].

To the best of our knowledge, we report here a novel variant underlining a frameshift mutation in the *JAG1* gene using the Ion Torrent platform. In the patient currently under study, the onset of ALGS was at an early age with hyperbilirubinemia and infantile cholestasis. WES analysis revealed the previously unreported heterozygous c.3080delC variant in exon 25, which produces a truncated *JAGGED1* protein due to a stop codon (p.P1027RfsTer9) and probably a diminished function of the Notch signaling pathway. A study reported that some symptoms associated with ALGS are deemed indicators of a bad prognosis, such as high total bilirubin levels between 12 and 24 months of age, liver fibrosis, and xanthomata [40]. The present patient had one of these predictors, high total bilirubin. However, further studies could clarify the effect of this mutation on the protein and signaling level. This *JAG1* (c.3080delC) mutation has never been reported in public human databases, including the following: ClinVar [41], COSMIC [42], the 1000 Genomes Project [43], gnomAD [44], ExAC [45], dbSNP [46], and the Human Gene Mutation Database [28]. It was also not found in our in-house database of 1000 exomes (personal data). In Figure 3A and B, we illustrated all *JAG1* pathogenic, or likely pathogenic, mutations reported in the public version of ClinVar

and COSMIC databases according to ExAC frequency. These variants encompass frameshift (nucleotide - level deletions, insertions, and insertion-deletions), nonsense (substitutions, start loss, and stop gain), missense, splice site, and in-frame deletions. We observed that the effect of mutations on phenotype and its severity differs between patients regardless of their mode of inheritance, which makes the understanding of the physiopathological mechanisms so far unknown. Interestingly, in the same exon 25, a missense mutation c.3080C>A;p.Pro1027Gln was previously reported in the COSMIC database (Figure 3B) which was disease-causing according to MutationTaster and damaging according to other in silico tools; FATHMM [47], MutPred [31], LRT [48], and EIGEN PC [49]. A 2019 study by Gilbert et al [50] showed that 94.3% of individuals with clinically diagnosed ALGS have a pathogenic variant in the *JAG1* gene, 2.5% have a pathogenic variant in the NOTCH2 gene, and 3.2% are molecularly uncharacterized [50]. The spectrum of *JAG1* mutations includes more frequently protein-truncating mutations (75%) and nonprotein truncating mutations (25%) [38,51].

In summary, our report broadens the spectrum of mutations in the *JAG1* gene. It thus aids the geneticist for better identification of the etiological mutations leading to ALGS, thereby providing further insights to detect the associated hot spots, exons or mutations.

Figure 3. Schematic of JAG1 protein with all listed variants: (A) ClinVar and (B) COSMIC databases. Dashed lines within the protein indicate exon boundaries, and numbers indicate amino acid coordinates. Protein domains include (*JAG1*): N terminus signal peptide (Salmon), DSL domain (Skyblue), EGF - like domain (teal), EGF domain (yellow), and VWC out domain (light green). All mutations listed are shown in different colors. *JAG1*: JAGGED 1 gene; ExAC: Exome Aggregation Consortium; DSL: Delta serrate ligand; EGF: EGF-like domain; UTR: Untranslated Region; VWC: von Willebrand factor (vWF) type C domain.



Conclusions

WES has revolutionized molecular genetic research and has become an essential genetic tool for molecular diagnosis of heterogeneous disorders. This study is an attempt to improve our understanding of the origin of ALGS caused by the identification of a variant in the *JAG1* gene. The novel mutation identified here provides an appropriate course of management to the patient to offer genetic counseling to the family; it also offers to expand the genetic spectrum of *JAG1*-related ALGS, raise awareness among pediatricians on the morbidity of this severe form of ALGS, which is thus far underdiagnosed, and show them the added value of next-generation sequencing technology in reducing diagnostic wandering.

Data Availability Statement

The data used to support the findings of this study are included within the article.

Authors' Contributions

DP and AT conceived and designed the experiments. DP processed the data and conceptualized and conceived the analytical methods. DP drafted the manuscript. AT and VL supervised the study and were in charge of the overall professional scientific

direction and planning. All authors discussed the results, provided critical feedback, helped shape the research and analysis, and finalized the manuscript.

Conflicts of Interest

None declared.

References

1. Turnpenny PD, Ellard S. Alagille syndrome: pathogenesis, diagnosis and management. Eur J Hum Genet 2012 Mar 21;20(3):251-257 [FREE Full text] [doi: 10.1038/ejhg.2011.181] [Medline: 21934706]
2. Danks DM, Campbell PE, Jack I, Rogers J, Smith AL. Studies of the aetiology of neonatal hepatitis and biliary atresia. Arch Dis Child 1977 May 01;52(5):360-367 [FREE Full text] [doi: 10.1136/adc.52.5.360] [Medline: 559475]
3. Alagille D, Estrada A, Hadchouel M, Gautler M, Odièvre M, Dommergues J. Syndromic paucity of interlobular bile ducts (Alagille syndrome or arteriohepatic dysplasia): Review of 80 cases. The Journal of Pediatrics 1987 Feb;110(2):195-200. [doi: 10.1016/s0022-3476(87)80153-1]
4. Lin HC, Le Hoang P, Hutchinson A, Chao G, Gerfen J, Loomes KM, et al. Alagille syndrome in a Vietnamese cohort: mutation analysis and assessment of facial features. Am J Med Genet A 2012 May 09;158A(5):1005-1013 [FREE Full text] [doi: 10.1002/ajmg.a.35255] [Medline: 22488849]
5. Gilbert MA, Spinner NB. Alagille syndrome: Genetics and Functional Models. Curr Pathobiol Rep 2017 Sep 1;5(3):233-241 [FREE Full text] [doi: 10.1007/s40139-017-0144-8] [Medline: 29270332]
6. Penton AL, Leonard LD, Spinner NB. Notch signaling in human development and disease. Semin Cell Dev Biol 2012 Jun;23(4):450-457 [FREE Full text] [doi: 10.1016/j.semcdb.2012.01.010] [Medline: 22306179]
7. Li L, Krantz ID, Deng Y, Genin A, Banta AB, Collins CC, et al. Alagille syndrome is caused by mutations in human Jagged1, which encodes a ligand for Notch1. Nat Genet 1997 Jul;16(3):243-251. [doi: 10.1038/ng0797-243] [Medline: 9207788]
8. Guegan K, Stals K, Day M, Turnpenny P, Ellard S. JAG1 mutations are found in approximately one third of patients presenting with only one or two clinical features of Alagille syndrome. Clin Genet 2012 Jul;82(1):33-40. [doi: 10.1111/j.1399-0004.2011.01749.x] [Medline: 21752016]
9. Grochowski CM, Loomes KM, Spinner NB. Jagged1 (JAG1): Structure, expression, and disease associations. Gene 2016 Jan 15;576(1 Pt 3):381-384 [FREE Full text] [doi: 10.1016/j.gene.2015.10.065] [Medline: 26548814]
10. Oda T, Elkahloun AG, Pike BL, Okajima K, Krantz ID, Genin A, et al. Mutations in the human Jagged1 gene are responsible for Alagille syndrome. Nat Genet 1997 Jul;16(3):235-242. [doi: 10.1038/ng0797-235] [Medline: 9207787]
11. Kamath BM, Bauer RC, Loomes KM, Chao G, Gerfen J, Hutchinson A, et al. NOTCH2 mutations in Alagille syndrome. J Med Genet 2012 Feb 29;49(2):138-144 [FREE Full text] [doi: 10.1136/jmedgenet-2011-100544] [Medline: 22209762]
12. Krantz ID, Colliton RP, Genin A, Rand EB, Li L, Piccoli DA, et al. Spectrum and frequency of jagged1 (JAG1) mutations in Alagille syndrome patients and their families. Am J Hum Genet 1998 Jun;62(6):1361-1369 [FREE Full text] [doi: 10.1086/301875] [Medline: 9585603]
13. Krantz ID, Smith R, Colliton RP, Tinkel H, Zackai EH, Piccoli DA, et al. Jagged1 mutations in patients ascertained with isolated congenital heart defects. Am. J. Med. Genet 1999 May 07;84(1):56-60. [doi: 10.1002/(sici)1096-8628(19990507)84:1<56::aid-ajmg11>3.0.co;2-w]
14. Crosnier C, Driancourt C, Raynaud N, Dhorne–Pollet S, Pollet N, Bernard O, et al. Mutations in JAGGED1 gene are predominantly sporadic in Alagille syndrome. Gastroenterology 1999 May;116(5):1141-1148. [doi: 10.1016/s0016-5085(99)70017-x]
15. Warthen D, Moore E, Kamath B, Morrissette J, Sanchez-Lara PA, Sanchez P, et al. Jagged1 (JAG1) mutations in Alagille syndrome: increasing the mutation detection rate. Hum Mutat 2006 May;27(5):436-443. [doi: 10.1002/humu.20310] [Medline: 16575836]
16. Yuan Z, Kohsaka T, Ikegaya T, Suzuki T, Okano S, Abe J, et al. Mutational analysis of the Jagged 1 gene in Alagille syndrome families. Hum Mol Genet 1998 Sep;7(9):1363-1369. [doi: 10.1093/hmg/7.9.1363] [Medline: 9700188]
17. Onouchi Y, Kurahashi H, Tajiri H, Ida S, Okada S, Nakamura Y. Genetic alterations in the JAG1 gene in Japanese patients with Alagille syndrome. J Hum Genet 1999 Jul;44(4):235-239. [doi: 10.1007/s100380050150] [Medline: 10429362]
18. Pilia G, Uda M, Macis D, Frau F, Crisponi L, Balli F, et al. Jagged-1 mutation analysis in Italian Alagille syndrome patients. Hum. Mutat 1999 Nov;14(5):394-400. [doi: 10.1002/(sici)1098-1004(199911)14:5<394::aid-humu5>3.0.co;2-1]
19. Heritage ML, MacMillan JC, Colliton RP, Genin A, Spinner NB, Anderson GJ. Jagged1 (JAG1) mutation detection in an Australian Alagille syndrome population. Hum. Mutat 2000 Nov;16(5):408-416. [doi: 10.1002/1098-1004(200011)16:5<408::aid-humu5>3.0.co;2-9]
20. Colliton RP, Bason L, Lu F, Piccoli DA, Krantz ID, Spinner NB. Mutation analysis of Jagged1 (JAG1) in Alagille syndrome patients. Hum. Mutat 2001 Feb;17(2):151-152. [doi: 10.1002/1098-1004(200102)17:2<151::aid-humu8>3.0.co;2-t]
21. Giannakudis J, Röpke A, Kujat A, Krajewska-Walasek M, Hughes H, Fryns J, et al. Parental mosaicism of JAG1 mutations in families with Alagille syndrome. Eur J Hum Genet 2001 Mar 11;9(3):209-216. [doi: 10.1038/sj.ejhg.5200613] [Medline: 11313761]

22. Röpke A, Kujat A, Gräber M, Giannakudis J, Hansmann I. Identification of 36 novel Jagged1 (JAG1) mutations in patients with Alagille syndrome. Hum Mutat 2003 Jan 20;21(1):100. [doi: 10.1002/humu.9102] [Medline: 12497640]
23. Jurkiewicz D, Popowska E, Gläser C, Hansmann I, Krajewska-Walasek M. Twelve novel JAG1 gene mutations in Polish Alagille syndrome patients. Hum Mutat 2005 Mar 14;25(3):321-321. [doi: 10.1002/humu.9313] [Medline: 15712272]
24. Sengupta S, Das J, Gangopadhyay A. Alagille syndrome with prominent skin manifestations. Indian J Dermatol Venereol Leprol 2005;71(2):119-121. [doi: 10.4103/0378-6323.13999] [Medline: 16394388]
25. Hadchouel M. Alagille syndrome. Indian J Pediatr 2002 Sep;69(9):815-818. [doi: 10.1007/bf02723697]
26. Shendge H, Tullu MS, Shenoy A, Chaturvedi R, Kamat JR, Khare M, et al. Alagille syndrome. Indian J Pediatr 2002 Sep;69(9):825-827. [doi: 10.1007/bf02723701]
27. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci U S A 2009 Nov 10;106(45):19096-19101 [FREE Full text] [doi: 10.1073/pnas.0910672106] [Medline: 19861545]
28. The Human Gene Mutation Database. URL: http://www.hgmd.cf.ac.uk/ac/index.php/ [accessed 2022-05-12]
29. UniProt. URL: http://www.uniprot.org/ [accessed 2022-05-12]
30. mutation t@sting. MutationTaster. URL: http://www.mutationtaster.org/ [accessed 2022-05-12]
31. MutPred2. URL: http://mutpred.mutdb.org/ [accessed 2022-05-12]
32. Integrative Genomics Viewer. Broad Institute. URL: https://software.broadinstitute.org/software/igv/ [accessed 2022-05-12]
33. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 2015 May;17(5):405-424 [FREE Full text] [doi: 10.1038/gim.2015.30] [Medline: 25741868]
34. Lee HP, Kang B, Choi SY, Lee S, Lee S, Choe YH. Outcome of Alagille Syndrome Patients Who Had Previously Received Kasai Operation during Infancy: A Single Center Study. Pediatr Gastroenterol Hepatol Nutr 2015 Sep;18(3):175-179 [FREE Full text] [doi: 10.5223/pghn.2015.18.3.175] [Medline: 26473137]
35. Chitayat D, Kamath B, Saleh M. Alagille syndrome: clinical perspectives. TACG 2016 Jun;Volume 9:75-82. [doi: 10.2147/tacg.s86420]
36. Togawa T, Sugiura T, Ito K, Endo T, Aoyama K, Ohashi K, et al. Molecular Genetic Dissection and Neonatal/Infantile Intrahepatic Cholestasis Using Targeted Next-Generation Sequencing. J Pediatr 2016 Apr;171:171-7.e1. [doi: 10.1016/j.jpeds.2016.01.006] [Medline: 26858187]
37. Kamath B, Baker A, Houwen R, Todorova L, Kerkar N. Systematic Review: The Epidemiology, Natural History, and Burden of Alagille Syndrome. J Pediatr Gastroenterol Nutr 2018 Aug;67(2):148-156 [FREE Full text] [doi: 10.1097/MPG.0000000000001958] [Medline: 29543694]
38. Mitchell E, Gilbert M, Loomes KM. Alagille Syndrome. Clin Liver Dis 2018 Nov;22(4):625-641. [doi: 10.1016/j.cld.2018.06.001] [Medline: 30266153]
39. Spinner NB, Colliton RP, Crosnier C, Krantz ID, Hadchouel M, Meunier-Rotival M. Jagged1 mutations in Alagille syndrome. Hum. Mutat 2000;17(1):18-33. [doi: 10.1002/1098-1004(2001)17:1<18::aid-humu3>3.0.co;2-t]
40. Mouzaki M, Bass LM, Sokol RJ, Piccoli DA, Quammie C, Loomes KM, et al. Early life predictive markers of liver disease outcome in an International, Multicentre Cohort of children with Alagille syndrome. Liver Int 2016 May 18;36(5):755-760 [FREE Full text] [doi: 10.1111/liv.12920] [Medline: 26201540]
41. ClinVar. National Library of Medicine. URL: https://www.ncbi.nlm.nih.gov/clinvar/ [accessed 2022-05-12]
42. COSMIC v95, released 24-NOV-21. COSMIC: Catalogue of Somatic Mutations in Cancer. URL: https://cancer.sanger.ac.uk/cosmic [accessed 2022-05-12]
43. 1000 Genomes. The International Genome Sample Resource. URL: https://www.internationalgenome.org/ [accessed 2022-05-12]
44. Genome Aggregation Database. gnomAD. URL: https://gnomad.broadinstitute.org/ [accessed 2022-05-12]
45. The ExAC browser: displaying reference data information from over 60 000 exomes. Broad Institute. URL: https://www.broadinstitute.org/publications/broad14291 [accessed 2022-05-12]
46. dbSNP. National Library of Medicine. URL: https://www.ncbi.nlm.nih.gov/snp/ [accessed 2022-05-12]
47. fathmm: Functional Analysis through Hidden Markov Models (v2.3). University of Bristol. URL: http://fathmm.biocompute.org.uk [accessed 2022-05-12]
48. LRT. Washington University School of Medicine in St. Louis. URL: https://sites.google.com/site/jpopgen/dbNSFP [accessed 2022-05-12]
49. EIGEN: A spectral approach integrating functional genomic annotations for coding and noncoding variants. Columbia University. URL: http://www.columbia.edu/~ii2135/eigen.html [accessed 2022-05-12]
50. Gilbert MA, Bauer RC, Rajagopalan R, Grochowski CM, Chao G, McEldrew D, et al. Alagille syndrome mutation update: Comprehensive overview of JAG1 and NOTCH2 mutation frequencies and insight into missense variant classification. Hum Mutat 2019 Dec 26;40(12):2197-2220 [FREE Full text] [doi: 10.1002/humu.23879] [Medline: 31343788]

51. Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, et al. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet 2017 Jun 27;136(6):665-677 [FREE Full text] [doi: 10.1007/s00439-017-1779-6] [Medline: 28349240]

Abbreviations

ALGS: Alagille syndrome
ExAC: Exome Aggregation Consortium
***JAG1*:** *JAGGED1* gene
NOTCH2: notch receptor 2
WES: whole exome sequencing

Edited by G Eysenbach; submitted 30.09.21; peer-reviewed by DJ Yadav, AJ Nagarajan; comments to author 12.01.22; accepted 27.04.22; published 08.07.22.

Please cite as:
Panwar D, Lal V, Thatai A
Identification of a Novel c.3080delC JAG1 Gene Mutation Associated With Alagille Syndrome: Whole Exome Sequencing
JMIR Bioinform Biotech 2022;3(1):e33946
URL: https://bioinform.jmir.org/2022/1/e33946
doi:10.2196/33946
PMID:27683065

Original Paper

Monitoring Risk Factors and Improving Adherence to Therapy in Patients With Chronic Kidney Disease (Smit-CKD Project): Pilot Observational Study

Antonio Vilasi[1]; Vincenzo Antonio Panuccio[2], MD; Salvatore Morante[3], MSc; Antonino Villa[3], BSc; Maria Carmela Versace[1], MA; Sabrina Mezzatesta[1], MSc; Sergio Mercuri[4], MSc; Rosalinda Inguanta[5], MSc, PhD; Giuseppe Aiello[5], MSc, PhD; Demetrio Cutrupi[1], MSc; Rossella Puglisi[3]; Salvatore Capria[1], MSc; Maurizio Li Vigni[3]; Giovanni Tripepi[1], MSc, PhD; Claudia Torino[1], MSc, PhD

[1]Institute of Clinical Physiology, National Research Council, Reggio Calabria, Italy

[2]Nephrology Unit, Grande Ospedale Metropolitano Bianchi Melacrino Morelli, Reggio Calabria, Italy

[3]Immedia Società per Azioni, Reggio Calabria, Italy

[4]Mercuri Informatica, Reggio Calabria, Italy

[5]Department of Engineering, University of Palermo, Palermo, Italy

Corresponding Author:
Claudia Torino, MSc, PhD
Institute of Clinical Physiology
National Research Council
via Vallone Petrara, snc
Reggio Calabria, 89124
Italy
Phone: 39 09 55393252
Email: ctorino@ifc.cnr.it

Abstract

Background: Chronic kidney disease is a major public health issue, with about 13% of the general adult population and 30% of the elderly affected. Patients in the last stage of this disease have an almost uniquely high risk of death and cardiovascular events, with reduced adherence to therapy representing an additional risk factor for cardiovascular morbidity and mortality. Considering the increased penetration of mobile phones, a mobile app could educate patients to autonomously monitor cardiorenal risk factors.

Objective: With this background in mind, we developed an integrated system of a server and app with the aim of improving self-monitoring of cardiovascular and renal risk factors and adherence to therapy.

Methods: The software infrastructure for both the Smit-CKD server and Smit-CKD app was developed using standard web-oriented development methodologies preferring open source tools when available. To make the Smit-CKD app suitable for Android and iOS, platforms that allow the development of a multiplatform app starting from a single source code were used. The integrated system was field tested with the help of 22 participants. User satisfaction and adherence to therapy were measured by questionnaires specifically designed for this study; regular use of the app was measured using the daily reports available on the platform.

Results: The Smit-CKD app allows the monitoring of cardiorenal risk factors, such as blood pressure, weight, and blood glucose. Collected data are transmitted in real time to the referring general practitioner. In addition, special reminders improve adherence to the medication regimen. Via the Smit-CKD server, general practitioners can monitor the clinical status of their patients and their adherence to therapy. During the test phase, 73% (16/22) of subjects entered all the required data regularly and sent feedback on drug intake. After 6 months of use, the percentage of regular intake of medications rose from 64% (14/22) to 82% (18/22). Analysis of the evaluation questionnaires showed that both the app and server components were well accepted by the users.

Conclusions: Our study demonstrated that a simple mobile app, created to self-monitor modifiable cardiorenal risk factors and adherence to therapy, is well tolerated by patients affected by chronic kidney disease. Further studies are required to clarify if the use of this integrated system will have long-term effects on therapy adherence and if self-monitoring of risk factors will improve clinical outcomes in this population.

(JMIR Bioinform Biotech 2022;3(1):e36766) doi:10.2196/36766

KEYWORDS

SMIT-CKD; mHealth; eHealth; CKD; therapy adherence; risk factor; kidney; adherence; integrated system; health app; monitoring; cardiology; cardiac; renal; chronic kidney disease; cardiovascular; mobile health; mobile app

Introduction

Chronic kidney disease (CKD) is a recognized major public health problem, with about 13% of the general adult population falling into one of the 5 stages identified by the Kidney Disease Outcome Quality Initiative classification [1]. Its prevalence increases to 15% to 30% in older persons, and exceeds 50% in patients with cardiovascular and metabolic comorbidities [1,2].

Additionally, these patients have an almost uniquely high risk of death and cardiovascular disease, with a rate of cardiovascular events strongly associated with the level of renal function [3]. The pathogenic mechanisms underlying the close relationship between kidneys and the cardiovascular system are not fully elucidated; however, the direct involvement of diabetes mellitus, arterial hypertension, and excess body weight in the high frequency of cardiovascular events both in renal failure and ischemic heart disease has been clarified [4-6]. More recently, reduced adherence to therapy has been recognized as an additional risk factor for cardiovascular morbidity and mortality in renal patients [7,8]. It is estimated that patients affected by CKD, especially those in the late stages, take on average 10 different drugs [9]. In these patients, low adherence may be due to the difficulty in reminding days and time of medicine intake, rather than the unwillingness to take medications [10-13], and this may explain the number of apps for medication monitoring in the major app stores [14,15].

The use of mobile apps for self-management in long-term conditions is not novel [16], and their number is increasing exponentially with the increasing use of mobile phones. A review published by Timmers et al [17] in 2020 showed that the use of smartphones for patient education improves medication or treatment adherence and clinical outcomes. Several mobile apps for the management of chronic conditions, such as diabetes and high blood pressure, are available in Google Play and the App Store [18-20]. Some of them are designed to monitor and correct patient behavior in order to reduce modifiable cardiovascular risk factors [21]; other solutions implement home-based rehabilitation programs for critically ill cardiovascular patients [22]. However, the large majority of these apps do not have medication monitoring as the main purpose [23], are not specifically focused on CKD patients [14,24,25], or are dedicated to patients in the late stages (eg, dialysis patients) [26].

With this in mind, the Smit-CKD project aimed at developing an integrated system designed for general practitioners (GPs) and patients consisting of a web-based platform (Smit-CKD server) and an app (Smit-CKD app), with the aim of improving medication regimen compliance and educating patients in self-monitoring of the most common risk factors for CKD and cardiovascular disease.

Methods

Design of the Smit-CKD Server and Smit-CKD App

As previously described, the growing coverage of the mobile cellular network and increased interest in mobile health (mHealth; the use of mobile and wireless technologies to support the achievement of health objectives) led us to conclude that an app specifically designed for the CKD population could be a good solution to improve adherence to therapy and decrease modifiable cardiorenal risk factors. In order to define the general architecture of the system, an in-depth bibliographic search was conducted with the aim of identifying the main risk factors for the progression of renal disease and clinical outcomes. Renal function is closely related to cardiovascular risk [3], so we started with the Framingham Heart Study [27], deriving a minimum set of predictors of mortality and cardiovascular events in the general population (age, sex, smoking, diabetes, previous cardiovascular events, cholesterol, high blood pressure) [28-30]. Other risk factors, such as BMI, an indicator of overweight and obesity [31,32], education level [33], and marital status [34] were subsequently added because of their association with mortality in the general population. Combining bibliographic research with good clinical practice, we defined the set of variables to be collected in the platform during the baseline and follow-up visits. This set includes personal data, anamnesis, education level, marital status, work activity, laboratory tests, somatometric data, blood pressure, and medications.

The next step consisted in the choice, among all the variables included the platform, of a minimum set of easily monitorable factors to be included in the app (ie, blood pressure, weight, diabetes [if any], and adherence to therapy) [35].

Regarding the frequency of blood pressure, blood glucose, and weight measurements, current guidelines for the management of blood pressure and diabetes in patients with chronic kidney disease were used.

Development of the Smit-CKD Server and Smit-CKD App

The software infrastructure for both the Smit-CKD server and Smit-CKD app was developed using standard web-oriented development methodologies, choosing open source tools when available, and MySQL as a database server. To ensure compatibility with the two major app stores, the choice was oriented toward platforms allowing the development of a multiplatform mobile app starting from a single source code.

To make the app suitable for the target audience (ie, patients with chronic kidney disease, in most cases elderly), we chose a user friendly interface, limiting the amount of information collected and user/interface interactions.

Design and Validation of Questionnaire to Measure Adherence to Therapy

The questionnaire to measure adherence to therapy used in the SMIT-CKD project was created after a literature search using *therapy* AND *adherence* AND *questionnaire* as keywords. All questionnaires resulting from this research were analyzed in order to create a new questionnaire that suited the needs of this study. Specifically, we examined the 4- and 8-question versions of the Morisky Medication Adherence Scale [36] and the Renal Treatment Satisfaction Questionnaire [37], the latest specially designed for CKD patients. Furthermore, we included questions on marital status, education level, and income since they are known risk factors for the progression of various chronic diseases [33,34,38] and may also play an important role with regard to adherence to therapy. The final questionnaire consisted of 13 questions (4 concerning demographic information, 9 concerning satisfaction with treatment and adherence to therapy) (Textbox 1) and was validated for language clarity, completeness, and relevance with the help of 10 volunteers having the same characteristics of the final users. Volunteers were asked to answer a series of questions, such as: Was it difficult to understand? Did you understand what this text means? Could this sentence be made better? Were you able to answer the question spontaneously? Is there anything you would like to delete? Is there anything you would like to add? All the volunteers evaluated the questionnaire without difficulty in interpreting the questions and their respective answers, so it was used in the original form during the pilot phase.

Textbox 1. Questionnaire items for adherence to therapy.

Demographic information

- What is your marital status?
- Whom do you live with?
- What is your highest educational qualification?
- What is or was your source of income?

Satisfaction with treatment and adherence to therapy

- How long have you been on medications for your kidney disease, hypertension, and diabetes (expressed in years)?
- How satisfied are you with your current treatment?
- How well do you think your kidney disease is controlled?
- How often do you experience side effects from medications?
- Does your medication regimen satisfy you in terms of side effects?
- How easy or comfortable did you find your therapy in the past 2 weeks?
- How satisfied are you with the knowledge of your state of health?
- Have you taken the prescribed doses of the medications in the past 2 weeks?
- Why did you not take the prescribed doses?

Testing of the Integrated Server-App System

The alpha version of the integrated system was tested by internal staff; multiple phases of debugging and implementation of new features were performed to obtain a final product with the planned features and ease of use. The resulting beta version was field tested in Reggio Calabria, Italy, from September 2019 to April 2020 with the help of local GPs. An invitation with the synopsis of the study protocol was sent to 10 GPs, randomly distributed in the urban area of Reggio Calabria, and 4 accepted the invitation and participated in the testing under the supervision of VAP. In order to participate in the testing phase, GPs were asked to randomly recruit, from the entire list of their patients, a minimum of 5 volunteers with the same characteristics of the final users of the system, thus fulfilling the following inclusion criteria: older than 18 years, creatinine 1.5 to 4.0 mg/dL (men) or 1.3 to 3.5 mg/dL (women), taking antihypertensive medications, own a smartphone with Android or iOS operating system (or assisted by family or caregivers), and written informed consent. For patients assisted by family or caregivers, the management of the app, including data entering and receiving of the medication alert, was the responsibility of the caregiver. Patients in other clinical trials or visually impaired or with acute kidney disease, rapidly progressive nephropathies or malignancies, or impaired cognitive abilities were excluded.

To test the platform in a real-life setting, GPs were asked to see the volunteers at the beginning of the testing phase, after 3 months, and after 6 months. During the first visit, patients were invited to download the Smit-CKD app from Google Play or the App Store. At all 3 visits, GPs recorded in the platform clinical data, laboratory data, and current medications (with date and time of administration) and set the frequency of measurement of blood pressure, body weight, and blood glucose (the latter for diabetics only). At each visit, patients were asked to complete the questionnaire about adherence to therapy. Throughout the testing phase, patients received alerts on their app according to GP settings reminding them to measure clinical data and take medications. Registered measurements and feedback on medications taken were sent in real time to the Smit-CKD server via internet connection. Compliance of the

volunteers in the use of the app was carefully monitored by their GPs, who checked the amount of clinical data transmitted to the platform and feedback to the received alerts on a weekly basis. In case of missing feedback, the participant was contacted to rule out app or mobile phone malfunction.

Study Outcomes

The outcomes of this study were user satisfaction, willingness to use the app, and potential usefulness of the integrated system.

Patient satisfaction was determined after the testing phase using a 7-question questionnaire (Textbox 2). We considered the tools available in the literature too complex and time consuming for our participants, so we designed a simpler version for this study. A similar questionnaire was administered to GPs for the web component (Textbox 2). In both questionnaires, satisfaction was expressed on a 5-point scale (1=poor satisfaction and 5= high satisfaction). Willingness to use the app was indirectly measured based on feedback received via the platform (entered measurements, answers to therapy alerts). Potential usefulness of the app was measured in terms of adherence to the therapy. For this purpose, the questionnaire for the adherence to therapy was administered 3 times, at baseline and after 3 and 6 months, and results from each visit were compared.

Textbox 2. Questionnaires administered to patients and general practitioners to measure the satisfaction level in the use of the integrated server-app system.

Patient satisfaction questionnaire

- How easy is it to use the app?
- How easy do you find connecting the supplied blood pressure monitor to the app?
- How easy do you find it to enter data in the app?
- How useful do you find the alerts that remind you to take the measurements?
- How useful do you find the alerts that remind you to take the medications?
- How often does the app crash or have problems?
- What is your overall opinion of the app?

General practitioner satisfaction

- How easy is it to use the web platform?
- How easy do you find entering data on the web platform?
- How easy do you find searching for information you need from the platform (eg, data entered for a specific patient)?
- How easy do you find the navigation between the various tabs on the platform?
- Was the support provided for use of the platform sufficient?
- How often does the web platform crash or have problems?
- What is your overall opinion regarding the web platform?

Ethics Approval

The study protocol was approved by the ethical committee of Commissione per l'Etica e l'Integrità nella Ricerca, National Research Council. All participants gave their informed consent. The integrated system was designed in accordance with the current European legislation regarding data processing (EU Regulation 679/2016). Data encryption was applied to protect data in case of accidental release of information; pseudonymization guarantees that personal data cannot be attributed to an identified or identifiable natural person. In addition, to protect data from lawful access to the platform, differentiated log-in credentials were created. To avoid illegal access to the computer system, the adoption of 2-factor authentication was implemented. Finally, the information technology system server location meets the necessary requirements in terms of security, redundancy, and operational recovery in the event of a disaster (disaster recovery).

Statistical Analysis

Data were expressed as mean and standard deviation for normally distributed variables, median and IQR for nonnormally distributed variables, and frequency and percentage for categorical variables. Clinical data were anonymously extracted from the web platform. Data on adherence to therapy were obtained by entering in the final data set the score of the respective questionnaires. Differences in adherences to therapy across visits (expressed as proportion of adherent patients) were analyzed using the N-1 chi-square test as recommended by Campbell [39] and Richardson [40]. Statistical analysis was performed with open source MedCalc (version 18, MedCalc Software Ltd).

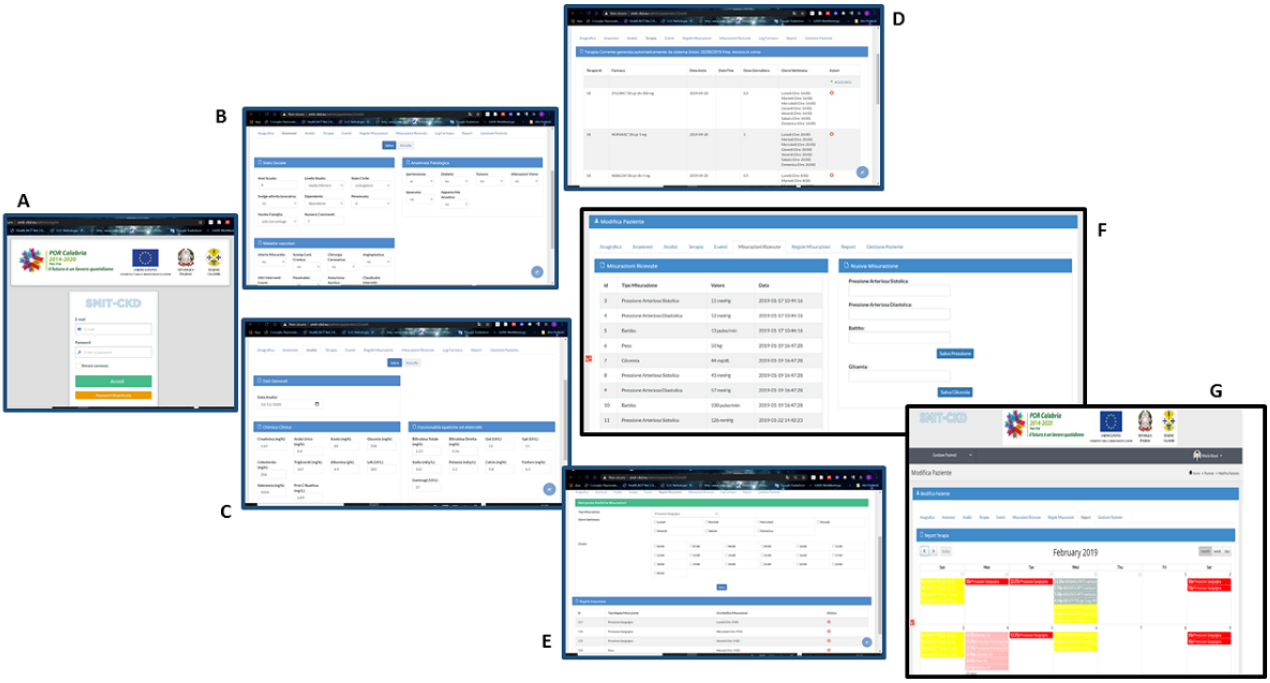
Results

Architectural Features and Functionality of the Web Platform

The Smit-CKD server (Figure 1) is accessible via 2-step authentication [41]. Access to the system is granted with 3

profiles (system administrator, medical supervisor, and doctor) with access to all or some tabs of the portal, according to their role. Furthermore, sensitive data can be accessed by the referring GP only.

Figure 1. Smit-CKD server: (A) home page/log-in, (B) anamnesis section, (C) laboratory examinations section, (D) therapeutic prescriptions, (E) measurements rules, (F) measurements received by app, (G) feedback received by app: shift from red to pale red indicates patient recorded blood pressure, glucose, or body weight; yellow indicates they took the prescribed medications.



The server component consists of a series of tabs, each designed to collect medical information. The first 3 tabs are dedicated to patient anamnesis, laboratory measurements, and ongoing therapies. Each medication is selected from a list of pre-entered medications; once the drug is selected, dosage and time of administration can be chosen.

Another tab allows the GP to enter the schedule of measurement to be performed by the patient (blood pressure, weight, blood glucose). Both settings are transmitted to the Smit-CKD app installed on the patient's smartphone, which sets the appropriate alarms and reminders. Finally, 2 tabs allow GPs to monitor all measurements performed by the patients and transmitted via app and feedback from patients about the medication alarm.

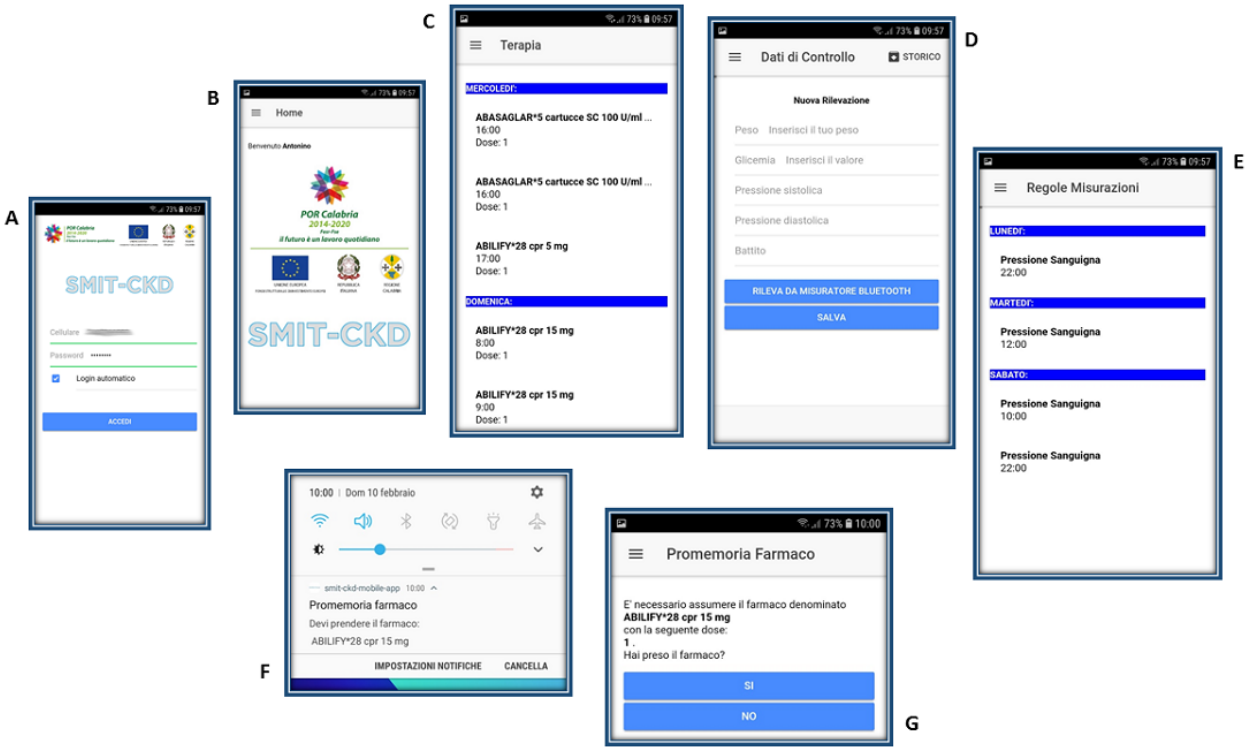
Architectural Features and Functionality of the Mobile App

The Smit-CKD app (Figure 2) has been developed for Android and iOS. An internet connection allows app and server to communicate. Access is granted by username and password after accepting use conditions. From the menu on the home screen of the Smit-CKD app, users can perform the following tasks:

- Pair the smartphone with a blood pressure monitor equipped with Bluetooth interface (configure blood pressure monitor)
- Display medications prescribed by the doctor (therapy)
- Enter control data, such as weight, blood glucose, and blood pressure, in case no Bluetooth monitor has been paired (control data). Data entered via app are transmitted in real time to the referring GP
- Visualize the history of all measurements taken or entered and consult the report of measurements sent to the remote server (history)

In addition to these activities, the app is designed to send personalized timed notifications to users reminding them to take a prescribed drug or measure weight, blood glucose, or blood pressure. After receiving the notification, the user is required to select YES or NO from the input box to indicate if the required action (for example "take the pill xxx") has been performed. This allows GPs to monitor the progress of the treatment.

Figure 2. Smit-CKD app: (A) log-in, (B) home page, (C) overview of therapy prescribed by doctor, (D) section for entering data, with overview of old measurements (history section), (E) overview of measurement rules, (F) example of reminder for therapy, (G) patient selects yes or no if they take or do not take the medicine; information is transmitted to doctor via Smit-CKD server, as seen in Figure 1.



Test Phase

The integrated system Smit-CKD server and mobile app was tested with the help of 22 participants enrolled by 4 GPs. The main clinical and demographic characteristics are reported in Table 1. Mean age was 70 (SD 11) years, with a male proportion of 59% (13/22). Participants were treated with antihypertensive drugs for a median time of 10 (IQR 6.50-21) years. According to the questionnaires administered during the study, a large majority (19/22, 86%) of the participants were satisfied with the treatment and considered their disease well controlled. Only 9% (2/22) of the participants manifested side effects more than once per month, maintaining a high level of satisfaction. Overall, patients considered the management of medications easy and were satisfied with their knowledge of their disease. Before using the app, 64% (14/22) regularly took all drugs, with forgetfulness the most common cause of missed doses. The rate of regular intake rose to 82% (18/22) after 6 months of using the app (Table 2). However, the difference was not significant (P=.18), probably due to the small sample.

The level of compliance with the use of the app was satisfactory, with 16 participants entering clinical data (blood pressure, body weight, and glucose) regularly and sending feedback of drug intake. Among the 6 low compliance users, 2 discontinued app use without giving a reason and 4 were minimally compliant due to technical problems related to an obsolete version of Android installed on their smartphone, which caused frequent malfunctioning of the app.

Among users, the percentage of satisfaction was high overall, as shown in Figure 3. The questionnaire administered to GPs to investigate the level of satisfaction with the use of the Smit-CKD server returned even better results, as all the GPs gave a score of 5 to all items in the questionnaire.

Table 1. Main demographic and clinical characteristics in the study population (n=22).

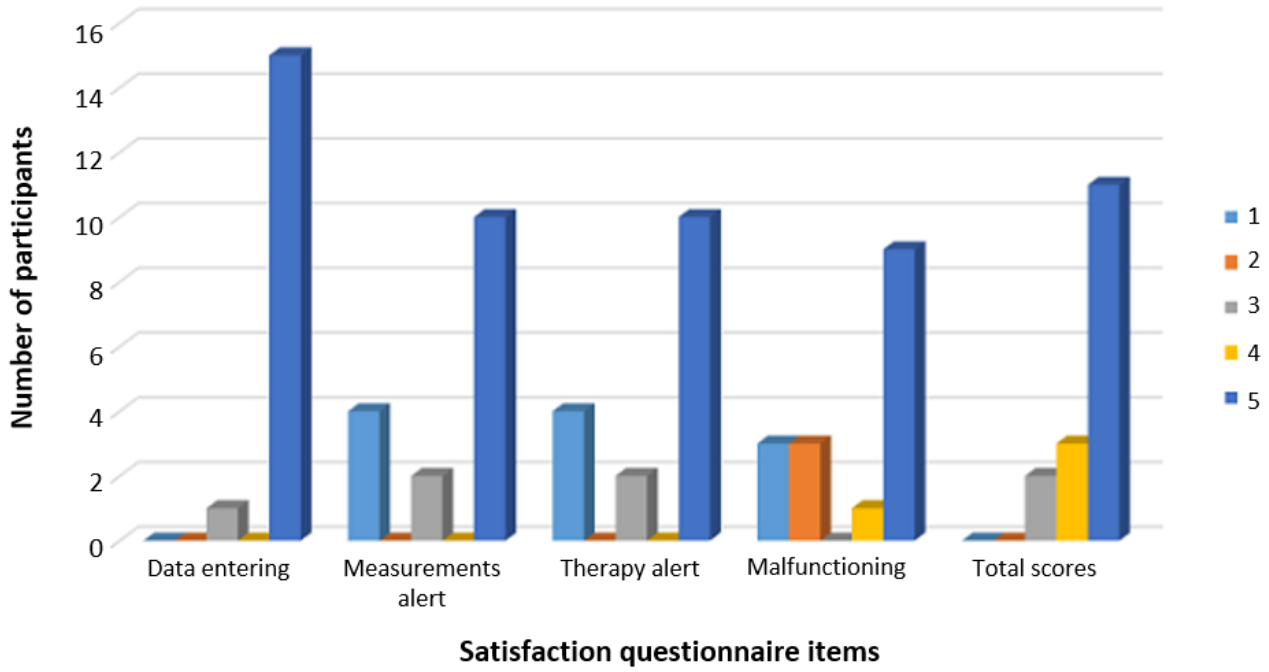
	Value
Age (years), mean (SD)	70 (11)
Male, n (%)	13 (59)
Diabetes (yes), n (%)	10 (46)
Cardiovascular comorbidities, n (%)	7 (32)
Cholesterol (mg/dL), mean (SD)	165 (42)
Hemoglobin (g/dL), mean (SD)	13.1 (2.0)
Albumin (g/dL), mean (SD)	4.2 (1.0)
Calcium (mg/dL), mean (SD)	9.4 (0.7)
Phosphate (mg/dL), mean (SD)	3.9 (1.3)
Creatinine (mg/dL), median (IQR)	
Male	2.0 (1.9-2.3)
Female	1.9 (1.7-3.2)
Marital status, n (%)	
Married	14 (64)
Widow	6 (27)
Separated/divorced	1 (4.5)
Never married	1 (4.5)
Education level, n (%)	
Low/medium education	13 (59)
High school diploma	5 (23)
University degree	2 (9)
No education	2 (9)
Currently working or retired, n (%)	18 (82)

Table 2. Questionnaire responses on adherence to therapy at baseline and after 6 months.

Question	Baseline, n (%)	After 6 months, n (%)
Satisfaction with the treatment		
Satisfied	19 (86)	19 (86)
Neutral	2 (9)	2 (9)
Unsatisfied	1 (5)	1 (5)
Control of kidney disease		
Well controlled	17 (77)	17 (77)
Neutral	2 (9)	2 (9)
Not controlled	3 (14)	3 (14)
Side effects		
≤1 per month	20 (91)	20 (91)
>1 per month	2 (9)	2 (9)
Satisfaction with side effects		
Highly satisfied	21 (95)	21 (95)
Not satisfied	2 (5)	1 (5)
Ease of medication regimen		
Very difficult	2 (9)	3 (14)
Neutral	2 (9)	2 (9)
Very easy	18 (82)	17 (77)
Knowledge about the disease		
Satisfied	17 (77)	17 (77)
Neutral	2 (9)	2 (9)
Unsatisfied	3 (14)	3 (14)
Medication intake		
All medication	14 (64)	18 (82)
Almost all	4 (18)	2 (9)
Part of	4 (18)	2 (9)
Reason for missed doses		
Forgetfulness	7[a]	3[a]
Too many pills	1[a]	1[a]

[a]Numbers refer only to those who took almost all or some of the medications (baseline n=8 and after 6 months n=4).

Figure 3. Level of satisfaction of the APP users, expressed on a scale with possible values from 1 to 5 (with 1 equivalent to poor satisfaction - 5 to high approval). Ease of use, data entering, alerts and bugs were considered.



Discussion

Principal Findings

In this paper we present a new integrated system designed for both GPs and patients consisting of a web-based platform (Smit-CKD server) and an app (Smit-CKD app) with the aim of improving medication compliance and educating patients in self-monitoring of the most common risk factors for CKD and cardiovascular disease. The beta version of the integrated system was tested on a small number of GPs and patients, well representative of the target population, confirming willingness and ease of use; in addition, even with a small sample, our data suggest the potential usefulness for improving adherence to therapy.

Growing coverage of mobile cellular networks has recently made mHealth capable of changing the approach to health systems worldwide. In this scenario, mobile apps are now regarded as potentially useful for changing health behavior and promoting self-management [42,43]. Among apps specifically designed for CKD patients, the *My Kidneys*, *My Health handbook* provides the user with educational information about detection of kidney disease and advice for a healthier life, including a calculator to compute the individual risk of kidney disease. *H2O Overload: Fluid Control for Heart-Kidney Health* is designed to help fluid intake control. *CKD Go!* allows creation of personalized action plans based on the estimated glomerular filtration rate and urine albumin-creatinine ratio, well known risk factors of CKD progression. CKD patients are often subject to dietary restrictions in order to control the progression of the disease. *My Food Coach*, designed by the National Kidney Foundation, offers nutritional advice from health care professionals. The American Association of Kidney Patients *myHealth Nutrition Guide* is an interactive app that provides the user with the nutrient values of more than 300 commonly consumed foods and many fast food restaurant options. Similarly, *Kidney APPetite* helps monitor daily nutrients and fluid intake. *Wholesome* collects healthy recipes from the web and contains personal recommendations to optimize nutrition. Focused on more specific dietary restrictions, *Oxalator* helps the user following a low oxalate diet, while *Phosphorus Foods Diet Guide* and *MyKidneyDiet – Phosphate Tracker* allows the user to monitor the content of phosphate in their diet.

Markossian et al [44] recently proposed a mobile app for self-management in stage 3-4 CKD patients. Features include the monitoring of clinical parameters, such as weight, blood pressure, and glucose, some aspects related to COVID-19 infection, and medication tracking. In addition, the automated system recognizes values out of range and suggests the patient contact the referring clinician; virtual visits can also be implemented.

The Smit-CKD app adds the self-monitoring of clinical parameters (blood pressure, blood glucose, and body weight), reminders to improve adherence to therapy, and transmission in real time of clinical information to the referring GP, thus allowing continuous monitoring of the health status of the patient and adherence to the medication regimen. This exchange allows the GPs to promptly intervene if data are out of the normal range or adherence to therapy is low; consequently, the patient is responsible for the self-monitoring of the main risk factors of CKD while feeling constantly monitored by their doctor.

Strengths and Limitations

A strength of the proposed integrated system is good tolerability reported by GPs and users, who appreciated the ease of use.

Furthermore, the alerts were helpful in reminding users to take their medications; this was supported by an increase in the adherence to therapy of 18% after 6 months of use.

However, our study has some important limitations. First of all, due to the low number of users and the study design, we cannot prove that this improvement in adherence to therapy is due to the use of the app. Second, even though blood pressure measurements and laboratory exams, useful for CKD monitoring, were collected during the testing phase, the limited number of users and short duration of follow-up prevent us from assessing the real impact of our system on CKD progression and clinical outcomes. However, this paper is not meant for describing the results of a clinical study but rather to introduce the new integrated system and its features developed for patients affected by CKD and their GPs. Finally, user satisfaction was not measured using validated tools but with a new, simpler questionnaire specifically designed for this study.

Conclusion

Our results suggest that the Smit-CKD app, easy to use and well accepted by patients, may improve adherence to therapy and empower patients affected by CKD. The use of the counterpart Smit-CKD server by GPs showed to be helpful in tracking the evolution of the disease in real time, preventing negative clinical outcomes, and increasing involvement of patients in the management of their clinical condition. The use of the integrated system at a national level could allow a more effective monitoring of patients in the early stages of CKD, contributing to slowing disease progression and delaying the first visit to the nephrology unit. Additional studies designed to test this app in a randomized controlled setting are required to clarify if the use of this integrated system will have long-term effects on medication adherence and if self-monitoring of risk factors will improve clinical outcomes in this population.

Acknowledgments

The Smit-CKD project was supported by Programma Operativo Regionale Calabria–Fondo Europeo di Sviluppo Regionale–Fondo Sociale Europeo 2014-2020. We thank Dr Alessi Maria Caterina, Dr Umberto Buccafurri, Dr Domenico Marra, and Dr Vittoria Pizzi for their contributions during the pilot phase.

Authors' Contributions

AV, VAP, GT, and CT designed the study and wrote the first draft of the manuscript. VAP and MCV collected the data. CT analyzed the data. SMez, AV, ML, SMer, and SMor contributed to the development of the integrated system. RI, GA, DC, RP, and SC critically reviewed the manuscript. The final version of the manuscript was approved by all the authors.

Conflicts of Interest

None declared.

References

1. Hill NR, Fatoba ST, Oke JL, Hirst JA, O'Callaghan CA, Lasserson DS, et al. Global prevalence of chronic kidney disease: a systematic review and meta-analysis. PLoS One 2016 Jul 6;11(7):e0158765 [FREE Full text] [doi: 10.1371/journal.pone.0158765] [Medline: 27383068]
2. Lameire N, Jager K, Van Biesen W, de Bacquer D, Vanholder R. Chronic kidney disease: a European perspective. Kidney Int Suppl 2005 Dec;68(99):S30-S38 [FREE Full text] [doi: 10.1111/j.1523-1755.2005.09907.x] [Medline: 16336574]
3. Zoccali C. Traditional and emerging cardiovascular and renal risk factors: an epidemiologic perspective. Kidney Int 2006 Jul;70(1):26-33 [FREE Full text] [doi: 10.1038/sj.ki.5000417] [Medline: 16723985]
4. Grundy SM, Benjamin IJ, Burke GL, Chait A, Eckel RH, Howard BV, et al. Diabetes and cardiovascular disease: a statement for healthcare professionals from the American Heart Association. Circulation 1999 Sep 07;100(10):1134-1146. [doi: 10.1161/01.cir.100.10.1134] [Medline: 10477542]
5. Kjeldsen SE. Hypertension and cardiovascular risk: general aspects. Pharmacol Res 2018 Mar;129:95-99. [doi: 10.1016/j.phrs.2017.11.003] [Medline: 29127059]
6. Poirier P, Giles TD, Bray GA, Hong Y, Stern JS, Pi-Sunyer FX, American Heart Association, Obesity Committee of the Council on Nutrition, Physical Activity, and Metabolism. Obesity and cardiovascular disease: pathophysiology, evaluation, and effect of weight loss: an update of the 1997 American Heart Association Scientific Statement on Obesity and Heart Disease from the Obesity Committee of the Council on Nutrition, Physical Activity, and Metabolism. Circulation 2006 Feb 14;113(6):898-918. [doi: 10.1161/CIRCULATIONAHA.106.171016] [Medline: 16380542]
7. Baroletti S, Dell'Orfano H. Medication adherence in cardiovascular disease. Circulation 2010 Mar 30;121(12):1455-1458. [doi: 10.1161/circulationaha.109.904003]
8. Leslie K, McCowan C, Pell J. Adherence to cardiovascular medication: a review of systematic reviews. J Public Health (Oxf) 2019 Mar 01;41(1):e84-e94 [FREE Full text] [doi: 10.1093/pubmed/fdy088] [Medline: 29850883]
9. Molnar AO, Bota S, Jeyakumar N, McArthur E, Battistella M, Garg AX, et al. Potentially inappropriate prescribing in older adults with advanced chronic kidney disease. PLoS One 2020 Aug 20;15(8):e0237868 [FREE Full text] [doi: 10.1371/journal.pone.0237868] [Medline: 32818951]

10. Kantor ED, Rehm CD, Haas JS, Chan AT, Giovannucci EL. Trends in prescription drug use among adults in the United States from 1999-2012. JAMA 2015 Nov 03;314(17):1818-1831 [FREE Full text] [doi: 10.1001/jama.2015.13766] [Medline: 26529160]
11. Payne RA, Avery AJ, Duerden M, Saunders CL, Simpson CR, Abel GA. Prevalence of polypharmacy in a Scottish primary care population. Eur J Clin Pharmacol 2014 May 1;70(5):575-581. [doi: 10.1007/s00228-013-1639-9] [Medline: 24487416]
12. Aston J, Wilson KA, Terry DRP. The treatment-related experiences of parents, children and young people with regular prescribed medication. Int J Clin Pharm 2019 Feb 26;41(1):113-121 [FREE Full text] [doi: 10.1007/s11096-018-0756-z] [Medline: 30478490]
13. Millar E, Gurney J, Stanley J, Stairmand J, Davies C, Semper K, et al. Pill for this and a pill for that: a cross-sectional survey of use and understanding of medication among adults with multimorbidity. Australas J Ageing 2019 Jun 16;38(2):91-97. [doi: 10.1111/ajag.12606] [Medline: 30556358]
14. Tabi K, Randhawa AS, Choi F, Mithani Z, Albers F, Schnieder M, et al. Mobile apps for medication management: review and analysis. JMIR Mhealth Uhealth 2019 Sep 11;7(9):e13608 [FREE Full text] [doi: 10.2196/13608] [Medline: 31512580]
15. Pouls BPH, Vriezekolk JE, Bekker CL, Linn AJ, van Onzenoort HAW, Vervloet M, et al. Effect of interactive ehealth interventions on improving medication adherence in adults with long-term medication: systematic review. J Med Internet Res 2021 Jan 08;23(1):e18901 [FREE Full text] [doi: 10.2196/18901] [Medline: 33416501]
16. Tabi K, Randhawa AS, Choi F, Mithani Z, Albers F, Schnieder M, et al. Mobile apps for medication management: review and analysis. JMIR Mhealth Uhealth 2019 Sep 11;7(9):e13608 [FREE Full text] [doi: 10.2196/13608] [Medline: 31512580]
17. Timmers T, Janssen L, Kool RB, Kremer JA. Educating patients by providing timely information using smartphone and tablet apps: systematic review. J Med Internet Res 2020 Apr 13;22(4):e17342 [FREE Full text] [doi: 10.2196/17342] [Medline: 32281936]
18. Siddique AB, Krebs M, Alvarez S, Greenspan I, Patel A, Kinsolving J, et al. Mobile apps for the care management of chronic kidney and end-stage renal diseases: systematic search in app stores and evaluation. JMIR Mhealth Uhealth 2019 Sep 04;7(9):e12604 [FREE Full text] [doi: 10.2196/12604] [Medline: 31486408]
19. Chen Y, Hung C, Huang C, Lee J, Yu J, Ho Y. The impact of synchronous telehealth services with a digital platform on day-by-day home blood pressure variability in patients with cardiovascular diseases: retrospective cohort study. J Med Internet Res 2022 Jan 10;24(1):e22957 [FREE Full text] [doi: 10.2196/22957] [Medline: 35006089]
20. Fundoiano-Hershcovitz Y, Hirsch A, Dar S, Feniger E, Goldstein P. Role of digital engagement in diabetes care beyond measurement: retrospective cohort study. JMIR Diabetes 2021 Feb 18;6(1):e24030 [FREE Full text] [doi: 10.2196/24030] [Medline: 33599618]
21. Agher D, Sedki K, Despres S, Albinet J, Jaulent M, Tsopra R. Encouraging behavior changes and preventing cardiovascular diseases using the prevent connect mobile health app: conception and evaluation of app quality. J Med Internet Res 2022 Jan 20;24(1):e25384 [FREE Full text] [doi: 10.2196/25384] [Medline: 35049508]
22. Claes J, Cornelissen V, McDermott C, Moyna N, Pattyn N, Cornelis N, et al. Feasibility, acceptability, and clinical effectiveness of a technology-enabled cardiac rehabilitation platform (Physical Activity Toward Health-I): randomized controlled trial. J Med Internet Res 2020 Feb 04;22(2):e14221 [FREE Full text] [doi: 10.2196/14221] [Medline: 32014842]
23. Ni Z, Wu B, Yang Q, Yan LL, Liu C, Shaw RJ. An mHealth intervention to improve medication adherence and health outcomes among patients with coronary heart disease: randomized controlled trial. J Med Internet Res 2022 Mar 09;24(3):e27202 [FREE Full text] [doi: 10.2196/27202] [Medline: 35262490]
24. Jupp JCY, Sultani H, Cooper CA, Peterson KA, Truong TH. Evaluation of mobile phone applications to support medication adherence and symptom management in oncology patients. Pediatr Blood Cancer 2018 Nov 26;65(11):e27278. [doi: 10.1002/pbc.27278] [Medline: 29943893]
25. Menditto E, Costa E, Midão L, Bosnic-Anticevich S, Novellino E, Bialek S, MASK group. Adherence to treatment in allergic rhinitis using mobile technology: the MASK study. Clin Exp Allergy 2019 Apr;49(4):442-460. [doi: 10.1111/cea.13333] [Medline: 30597673]
26. Lukkanalikitkul E, Kongpetch S, Chotmongkol W, Morley MG, Anutrakulchai S, Srichan C, et al. Optimization of the chronic kidney disease-peritoneal dialysis app to improve care for patients on peritoneal dialysis in northeast thailand: user-centered design study. JMIR Form Res 2022 Jul 06;6(7):e37291 [FREE Full text] [doi: 10.2196/37291] [Medline: 35793137]
27. Dawber T, Kannel W. An epidemiologic study of heart disease: the Framingham study. Nutr Rev 1958 Jan;16(1):1-4. [doi: 10.1111/j.1753-4887.1958.tb00605.x] [Medline: 13493903]
28. Mahmood SS, Levy D, Vasan RS, Wang TJ. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. Lancet 2014 Mar 15;383(9921):999-1008 [FREE Full text] [doi: 10.1016/S0140-6736(13)61752-3] [Medline: 24084292]
29. Tsao CW, Vasan RS. Cohort Profile: The Framingham Heart Study (FHS): overview of milestones in cardiovascular epidemiology. Int J Epidemiol 2015 Dec 23;44(6):1800-1813 [FREE Full text] [doi: 10.1093/ije/dyv337] [Medline: 26705418]

30. Schnabel RB, Yin X, Gona P, Larson MG, Beiser AS, McManus DD, et al. 50 year trends in atrial fibrillation prevalence, incidence, risk factors, and mortality in the Framingham Heart Study: a cohort study. The Lancet 2015 Jul;386(9989):154-162. [doi: 10.1016/s0140-6736(14)61774-8]
31. Mora S, Yanek LR, Moy TF, Fallin MD, Becker LC, Becker DM. Interaction of body mass index and framingham risk score in predicting incident coronary disease in families. Circulation 2005 Apr 19;111(15):1871-1876. [doi: 10.1161/01.CIR.0000161956.75255.7B] [Medline: 15837938]
32. Ghehi C, Gabillard D, Moh R, Badje A, Kouamé G, Oouttara E, et al. High correlation between Framingham equations with BMI and with lipids to estimate cardiovascular risks score at baseline in HIV-infected adults in the Temprano trial, ANRS 12136 in Côte d'Ivoire. PLoS One 2017 Jun 5;12(6):e0177440 [FREE Full text] [doi: 10.1371/journal.pone.0177440] [Medline: 28582393]
33. Kunst AE, Mackenbach JP. The size of mortality differences associated with educational level in nine industrialized countries. Am J Public Health 1994 Jun;84(6):932-937. [doi: 10.2105/ajph.84.6.932] [Medline: 8203689]
34. Robards J, Evandrou M, Falkingham J, Vlachantoni A. Marital status, health and mortality. Maturitas 2012 Dec;73(4):295-299 [FREE Full text] [doi: 10.1016/j.maturitas.2012.08.007] [Medline: 23007006]
35. Magacho E, Ribeiro L, Chaoubah A, Bastos M. Adherence to drug therapy in kidney disease. Braz J Med Biol Res 2011 Mar;44(3):258-262 [FREE Full text] [doi: 10.1590/s0100-879x2011007500013] [Medline: 21344138]
36. Morisky DE, Green LW, Levine DM. Concurrent and predictive validity of a self-reported measure of medication adherence. Med Care 1986 Jan;24(1):67-74. [doi: 10.1097/00005650-198601000-00007] [Medline: 3945130]
37. Barendse SM, Speight J, Bradley C. The Renal Treatment Satisfaction Questionnaire (RTSQ): a measure of satisfaction with treatment for chronic kidney failure. Am J Kidney Dis 2005 Mar;45(3):572-579. [doi: 10.1053/j.ajkd.2004.11.010] [Medline: 15754280]
38. Zeng X, Liu J, Tao S, Hong HG, Li Y, Fu P. Associations between socioeconomic status and chronic kidney disease: a meta-analysis. J Epidemiol Community Health 2018 Apr 02;72(4):270-279. [doi: 10.1136/jech-2017-209815] [Medline: 29437863]
39. Campbell I. Chi-squared and Fisher–Irwin tests of two-by-two tables with small sample recommendations. Statist Med 2007 Aug 30;26(19):3661-3675. [doi: 10.1002/sim.2832] [Medline: 17315184]
40. Richardson JTE. The analysis of 2 × 2 contingency tables, yet again. Stat Med 2011 Apr 15;30(8):890-892. [doi: 10.1002/sim.4116] [Medline: 21432882]
41. Smit-CKD. URL: http://www.smit-ckd.eu/ [accessed 2022-11-07]
42. Zhao J, Freeman B, Li M. Can mobile phone apps influence people's health behavior change? an evidence review. J Med Internet Res 2016 Oct 31;18(11):e287 [FREE Full text] [doi: 10.2196/jmir.5692] [Medline: 27806926]
43. Rudolf I, Pieper K, Nolte H, Junge S, Dopfer C, Sauer-Heilborn A, et al. Assessment of a mobile app by adolescents and young adults with cystic fibrosis: pilot evaluation. JMIR Mhealth Uhealth 2019 Nov 21;7(11):e12442 [FREE Full text] [doi: 10.2196/12442] [Medline: 31750841]
44. Markossian TW, Boyda J, Taylor J, Etingen B, Modave F, Price R, et al. A mobile app to support self-management of chronic kidney disease: development study. JMIR Hum Factors 2021 Dec 15;8(4):e29197 [FREE Full text] [doi: 10.2196/29197] [Medline: 34914614]

Abbreviations

CKD: chronic kidney disease
GP: general practitioner
mHealth: mobile health

Edited by A Mavragani; submitted 05.07.22; peer-reviewed by M Burnier, V Buss; comments to author 26.08.22; revised version received 26.10.22; accepted 05.11.22; published 15.11.22.

Please cite as:
Vilasi A, Panuccio VA, Morante S, Villa A, Versace MC, Mezzatesta S, Mercuri S, Inguanta R, Aiello G, Cutrupi D, Puglisi R, Capria S, Li Vigni M, Tripepi G, Torino C
Monitoring Risk Factors and Improving Adherence to Therapy in Patients With Chronic Kidney Disease (Smit-CKD Project): Pilot Observational Study
JMIR Bioinform Biotech 2022;3(1):e36766
URL: https://bioinform.jmir.org/2022/1/e36766
doi:10.2196/36766
PMID:

Original Paper

Novel Molecular Networks and Regulatory MicroRNAs in Type 2 Diabetes Mellitus: Multiomics Integration and Interactomics Study

Manoj Khokhar[1], MSc; Dipayan Roy[1], MD; Sojit Tomo[1], MD; Ashita Gadwal[1], MSc; Praveen Sharma[1], PhD; Purvi Purohit[1], PhD

Department of Biochemistry, All India Institute of Medical Sciences, Jodhpur, India

Corresponding Author:
Purvi Purohit, PhD
Department of Biochemistry
All India Institute of Medical Sciences
CRL-2, first floor, Basni Industrial Area
MIA 2nd Phase, Basni
Jodhpur, 342005
India
Phone: 91 9928388223
Email: dr.purvipurohit@gmail.com

Abstract

Background: Type 2 diabetes mellitus (T2DM) is a metabolic disorder with severe comorbidities. A multiomics approach can facilitate the identification of novel therapeutic targets and biomarkers with proper validation of potential microRNA (miRNA) interactions.

Objective: The aim of this study was to identify significant differentially expressed common target genes in various tissues and their regulating miRNAs from publicly available Gene Expression Omnibus (GEO) data sets of patients with T2DM using in silico analysis.

Methods: Using differentially expressed genes (DEGs) identified from 5 publicly available T2DM data sets, we performed functional enrichment, coexpression, and network analyses to identify pathways, protein-protein interactions, and miRNA-mRNA interactions involved in T2DM.

Results: We extracted 2852, 8631, 5501, 3662, and 3753 DEGs from the expression profiles of GEO data sets GSE38642, GSE25724, GSE20966, GSE26887, and GSE23343, respectively. DEG analysis showed that 16 common genes were enriched in insulin secretion, endocrine resistance, and other T2DM-related pathways. Four DEGs, *MAML3*, *EEF1D*, *NRG1*, and *CDK5RAP2*, were important in the cluster network regulated by commonly targeted miRNAs (hsa-let-7b-5p, hsa-mir-155-5p, hsa-mir-124-3p, hsa-mir-1-3p), which are involved in the advanced glycation end products (AGE)-receptor for advanced glycation end products (RAGE) signaling pathway, culminating in diabetic complications and endocrine resistance.

Conclusions: This study identified tissue-specific DEGs in T2DM, especially pertaining to the heart, liver, and pancreas. We identified a total of 16 common DEGs and the top four common targeting miRNAs (hsa-let-7b-5p, hsa-miR-124-3p, hsa-miR-1-3p, and has-miR-155-5p). The miRNAs identified are involved in regulating various pathways, including the phosphatidylinositol-3-kinase-protein kinase B, endocrine resistance, and AGE-RAGE signaling pathways.

(JMIR Bioinform Biotech 2022;3(1):e32437) doi:10.2196/32437

KEYWORDS

type 2 diabetes mellitus; interactomics; integrative genomics; protein-protein interaction; microRNAs; miRNA; bioinformatics; multiomics; genomics; gene expression

Introduction

Interactions among DNA, RNA, and proteins regulate their functions and have an immense effect on the underlying mechanistic processes in the pathophysiology of many diseases. Owing to the advent of newer technologies such as microarray and genome sequencing, it is now possible to investigate and analyze an enormous amount of genomic and proteomic data to predict disease pathology, outcome, and possible therapeutic targets [1]. Diabetes is a metabolic disorder characterized by

hyperglycemia and glycosuria, which, if left untreated, leads to an array of complications and associated comorbidities [2]. These can include obesity, cardiomyopathy, nephropathy, retinopathy, neuropathy, and peripheral vascular disease, which have a lasting adverse effect on the quality of the patient's life. To date, diabetes has affected almost half a billion individuals worldwide [3]. The absence of effective treatment strategies for this disease makes it a challenge to manage. The obligatory lifestyle changes and multiple treatment modalities, along with lifelong disease monitoring, depict an urgent and unmet need to develop newer and specific preventive and treatment strategies. Mortality rates in patients with type 2 diabetes mellitus (T2DM) are higher than those of individuals without diabetes and are linked to increased cardiovascular, renovascular, and neuropathic risks [4,5]. Thus, to reduce the morbidity and mortality associated with T2DM, it is important to gain a better understanding of its pathogenic pathways and regulation mechanisms. There is accumulating evidence that microRNAs (miRNAs) play an essential role in diabetes by reducing the expression of their various target genes [6,7]. It is also crucial to select the right target for disease treatment strategies in the early discovery phases, thus maximizing the drug's success rates in the latter phases [8].

Currently, there is a vast amount of genomic data on diabetes and its complications. However, from its detection to the management of its late-stage complications, many areas still need to be explored and lacunae need to be filled. The role of molecular integration networks regulating the pathogenesis of T2DM in specific tissues is unknown. In this study, we have undertaken an in silico approach with existing tissue-specific microarray data of patients with diabetes to address this particular area by detecting novel diabetes-associated genes, their regulatory miRNAs, and their interactions to predict the pivotal pathways in tissues that are associated with disease onset and progression.

We selected five data sets from the Gene Expression Omnibus (GEO) database comprising the expression profiles of patients with diabetes and corresponding controls, and identified 16 differentially expressed genes (DEGs) overlapping the three preassigned groups. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses were further used to classify the DEGs into cellular component (CC), biological process (BP), and molecular function (MF) classes. We selected four clusters from the protein-protein interaction (PPI) network and identified the seed genes. We further investigated the miRNA and hub gene network. Finally, we explored the 16 hub genes for biological pathway enrichment and their targeting miRNAs.

Methods

Data Collection

We searched several keywords, including "type 2 diabetes mellitus," "tissue," "pancreas," "liver," "heart," "expression profiling by array," and "*Homo sapiens*," in the GEO data sets, among which five were selected for this study: GSE38642 [9-11], GSE25724 [12], GSE20966 [13], GSE26887 [14], and GSE23343 [15] (Table 1).

Table 1. Description of Gene Expression Omnibus data sets for three groups of organs.

Sample organ	T2DM[a]			Control			Platform	Country	Year
	Samples, n	Sex (M/F)	Mean age (years)	Samples, n	Sex (M/F)	Mean age (years)			
Pancreas									
GSE25724 [12]	6	3/3	58.1	7	4/3	70.5	Affymetrix Human Genome U133A Array	Italy	2010
GSE20966 [13]	10	7/3	60.3	10	6/4	67.3	Affymetrix Human X3P Array	United States	2010
GSE38642 [9-11]	9	5/4	57.0	54	31/23	56.6	Affymetrix Human Gene 1.0 ST Array	Sweden	2012
Heart (GSE26887) [14]	7	6/1	65.1	5	2/3	48.4	Affymetrix Human Gene 1.0 ST Array	Italy	2012
Liver (GSE23343) [15]	10	4/6	—[b]	7	4/3	—	Affymetrix Human Genome U133 Plus 2.0 Array	Japan	2010

[a]T2DM: Type 2 diabetes mellitus.

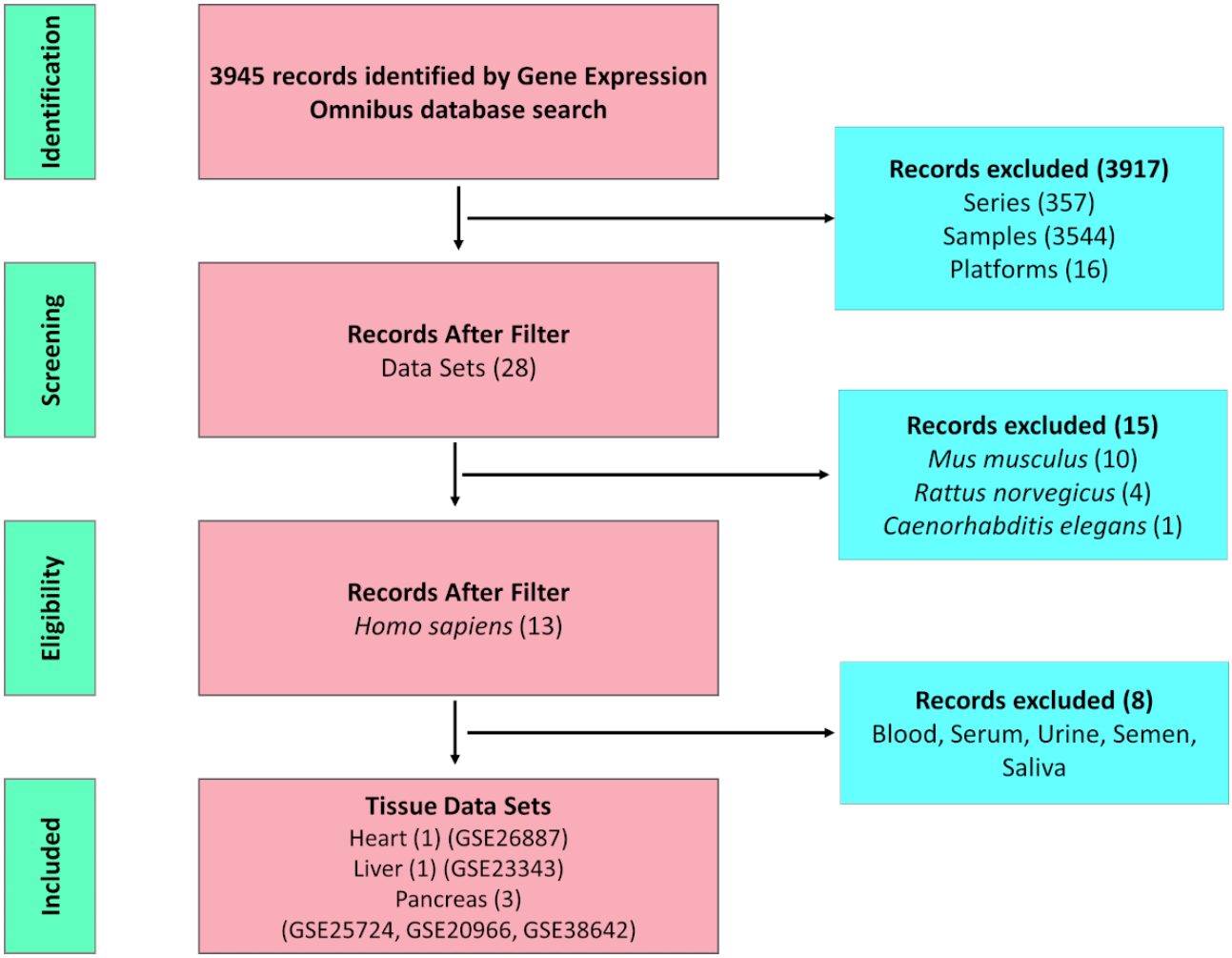
[b]Information not available.

Inclusion and Exclusion Criteria

Data were restricted to human (*Homo sapiens*) samples, data set as the data type, expression profiling by array, tissue samples, and T2DM compared to controls (without diabetes). Thus, data from other organisms (*Mus musculus*, *Rattus norvegicus*, *Xenopus laevis*); series data; expression profiling by other methods (eg, massively parallel signature sequencing, reverse transcription-polymerase chain reaction, serial analysis of gene expression, genome variation or occupancy profiling by single-nucleotide polymorphism array, genome tiling array); nontissue samples (eg, blood, serum, semen, saliva, urine, body fluid); and data from patients with type 1 diabetes, gestational diabetes, or prediabetes were excluded.

The data collection process is summarized in Figure 1.

Figure 1. Flow diagram illustrating the process of data collection and number of data sets considered for inclusion.

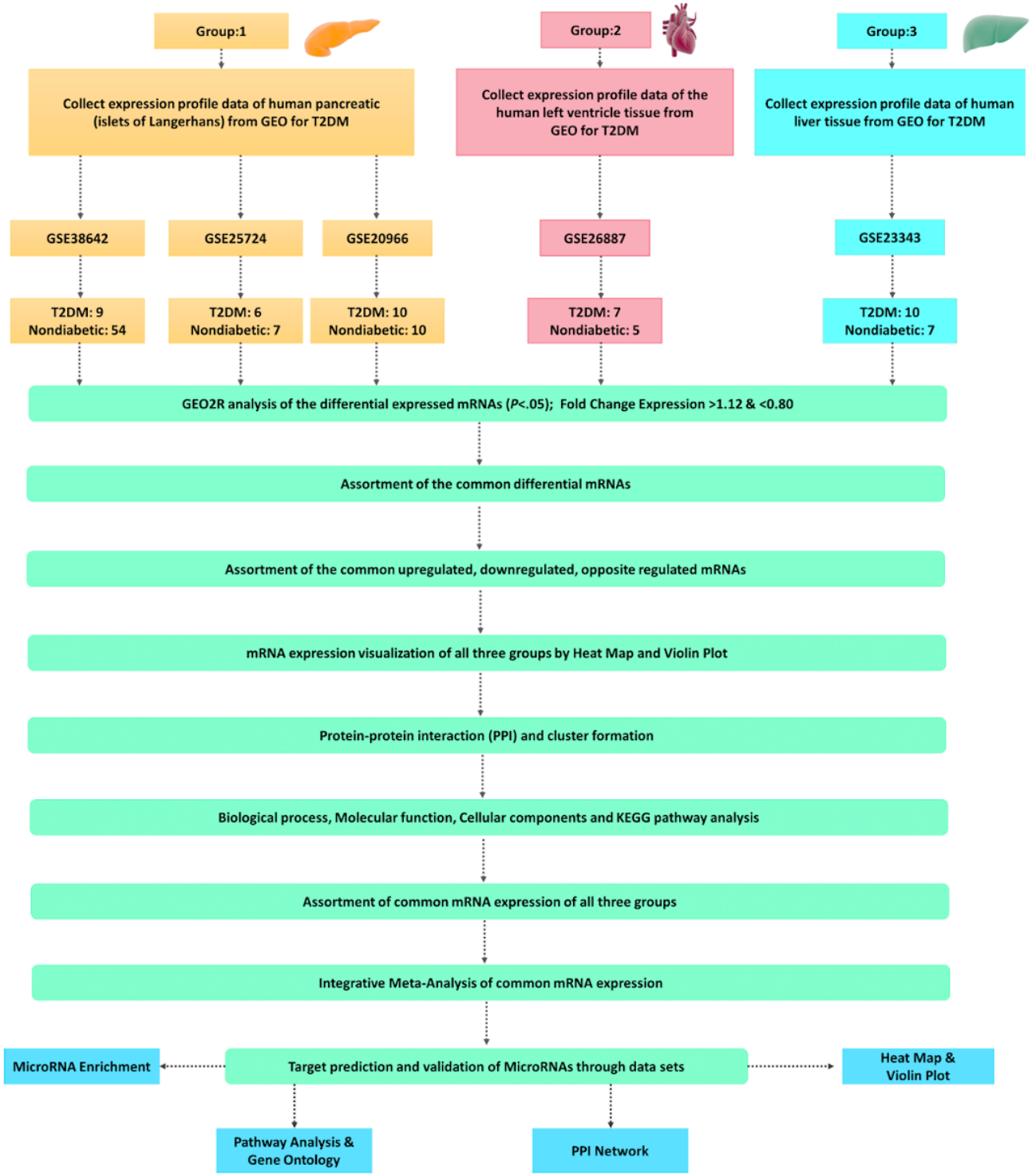


Identification and Assortment of Differentially Expressed mRNAs

The DEGs were obtained from the five data sets using the online interactive tool GEO2R [16]. The cutoff for the selection was kept at the default of $P<.05$. The relaxed P-value cutoff was fixed for the initial selection because (1) we were subjecting the selected genes for a repeated analysis using ImaGEO software with a cutoff adjusted $P<.05$, and (2) the application of a stringent P-value cutoff in the initial selection did not enable obtaining an adequate number of genes from each data set for undertaking a meta-analysis. The overlapping DEGs among the three data sets of pancreatic tissues from patients with T2DM and controls (GSE38642, GSE25724, and GSE20966) were identified using the Venn diagram tool [17,18]. Subsequently, the common DEGs of these three data sets (GSE38642, GSE25724, and GSE20966) with those of the other two data sets for heart (GSE26887) and liver (GSE23343) samples were identified separately. The fold change expression distribution was visualized by heat maps and violin plots using the R *limma* (linear models for microarray data) package and Orange Data Mining software [19,20].

To check the quality of the data, quality control plots were assessed in the form of volcano plots, mean difference plots, and mean-variance trends. A volcano plot visualizes the DEGs by plotting the statistical significance against the magnitude of change, whereas the mean difference plot displays the log2 fold change against the average log2 expression level. The mean-variance trend, generated using the R packages plotSA and vooma, assesses the variance of the data. The workflow for the data processing and analysis is portrayed in Figure 2.

Figure 2. Flowchart of data processing and analysis. GEO: Gene Expression Omnibus; KEGG: Kyoto Encyclopedia of Genes and Genomes; T2DM: type 2 diabetes mellitus.



Functional Enrichment and KEGG Pathway Analysis

The DEGs were divided into three groups according to the tissue (Figure 2). The functional enrichment of each group related to T2DM was analyzed with the Database for Annotation, Visualization and Integrated Discovery (DAVID) tool for significant MF, CC, and BP GO terms. KEGG pathway analysis was performed with piNET, a versatile tool that integrates protein signatures with transcriptomic and proteomic signatures [21-23]. DAVID includes an analysis of KEGG pathways and enrichment significance of GO terms from the three categories (MF, CC, BP). We defined $P<.05$ as significantly enriched. The nonsignificant findings were manually removed.

PPI Network Construction and Identification of Hub Genes

The DEGs in the three groups were used to construct the PPI network using Search Tool for the Retrieval of Interacting Genes/Protein (STRING) [24]. We established the PPI network using only the overlapping DEGs with greater than 0.4 confidence score cutoffs. The "combined scores" were computed by integrating the probabilities from the various different types

of evidence (by evidence channels), while correcting for the probability of randomly observing an interaction [25]. The number of interactions (by confidence level) were divided into four groups: (1) highest confidence (score≥0.90), (2) high confidence or better (score≥0.70), (3) medium confidence or better (score≥0.40), (4) low confidence links (score≥0.15). We chose medium confidence as the default setting given in STRING.

The interaction networks for each group were constructed by Cytoscape [26,27]. The Molecular Complex Detection (MCODE) [28] plugin of Cytoscape was employed to visualize significant genes in all three groups with a degree cutoff=2, node score cutoff=0.2, k-score=2, and maximum depth=100. The criteria for selecting the top 3 clusters were set as MCODE score≥3 and number of nodes≥3.

Integrative Gene Expression Meta-analysis

ImaGEO is a web tool for gene expression meta-analysis that was used to perform a comprehensive meta-analysis from all five data sets. For the retrieval and preprocessing of the data, the GEOquery package in R was used, followed by quality control, gene filtering expression, meta-analysis, and functional analysis. The meta-analysis was based on the functional modules with the MetaDE R package. For this study, we used the "effect size" parameter estimation with a fixed-effects model and an adjusted *P* value threshold of .05. The allowable missing values was kept at the default of 10%.

Target Prediction, Validation, and miRNA–Hub Gene Interaction

The top 10 targeting miRNAs of the hub genes were predicted by the well-established miRNA target prediction database miRNet 2.0 [29] with *H. sapiens* (human) as the selected organism. Default values for the degree of interaction and betweenness were retained. Common miRNAs and targeted mRNAs of all groups were sorted by the Venn diagram tool [30]. The network of all targeting miRNAs and the coexpressed mRNAs was created with FunRich and Cytoscape software. To validate the targeting miRNAs, we further sorted miRNA data sets in T2DM for comparison of differentially expressed miRNAs.

Functional Enrichment and KEGG Pathway Analysis for MiRNAs

All common miRNAs were enriched by MicroRNA Enrichment Turned Network (MIENTURNET) and KEGG pathway analysis [31]. MIENTURNET is a web tool based on the shiny package in R studio for both statistical and network-based analyses of miRNA-target enrichment. Functional enrichment was retrieved for the input list of genes, with the minimum interaction threshold set at 2 and an adjusted *P* value of .05. The input list infers possible experimental or computational evidence of miRNA-based regulation.

Results

Identification of DEGs in all Combined Groups

The five mRNA expression profiles of the GSE38642, GSE25724, GSE20966, GSE26887, and GSE23343 data sets, including 125 samples of the pancreas, heart, and liver tissues of patients with T2DM and controls without diabetes, were included in this study. We extracted 2852, 8631, 5501, 4210, and 3754 DEGs, respectively. The following sections describe the analysis of the DEGs derived from the datasets, and shown in Figures 3-14. In the pancreas data sets (GSE38642, GSE25724, GSE20966), a total of 321 common mRNAs were identified, 69 of which were upregulated and 95 were downregulated (Figure 3A-H and Supplementary Tables S1-S3 in Multimedia Appendix 1). The quality control plots for the DEGs are shown in Figure S1 and Figure S2 of Multimedia Appendix 1.

These Group 1 (pancreas) DEGs were then overlapped with the heart expression profile data set GSE26887, revealing a total of 70 common differentially expressed mRNAs, 5 of which were downregulated and 5 were upregulated. A total of 28 mRNAs with regulation in the opposite direction were identified (Group 2) (Figure 5A-K, Tables S4-S7 in Multimedia Appendix 1). Further, the Group 1 DEGs were overlapped with the liver data set GSE23343, and a total of 82 common differentially expressed mRNAs were obtained, out of which 8 were upregulated, 1 was downregulated, and 27 were regulated in opposite directions (Figure 7A-I, Tables S8-S11 in Multimedia Appendix 1).

DEGs for all three groups were used to establish the PPI networks (Figure 6E, Figure 8E, Figure 9A).

Figure 3. Differential mRNA expression of all three data sets (GSE38642, GSE25724, GSE20966) for Group 1 (pancreas tissues) in type 2 diabetes mellitus. (A-C) Heat maps of all, downregulated and upregulated differentially expressed genes (DEGs). Fold change expression (FCE) levels are displayed in ascending order from blue to yellow. (D-F) Venn diagrams of the total downregulated and upregulated DEGs of the three data sets. (G, H) Violin plots showing the entire FCE distribution of all three data sets for upregulated and downregulated common DEGs.

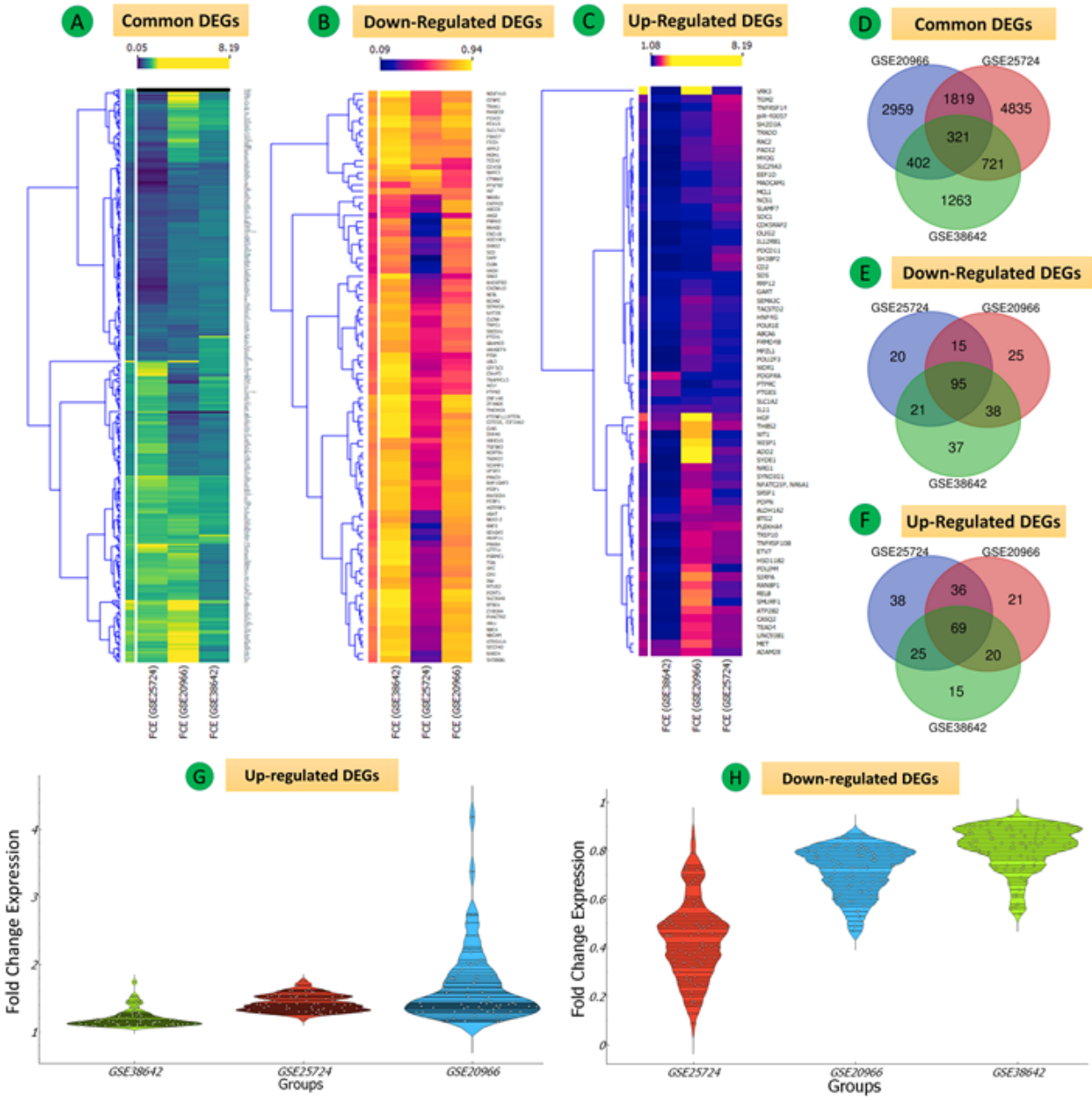


Figure 4. Differential mRNA expression of all three data sets (GSE38642, GSE25724, GSE20966) for Group 1 (pancreas tissues) in type 2 diabetes mellitus. (A-D) Enrichment analysis of common DEGs. (A) Kyoto Enclyclopedia of Genes and Genomes (KEGG) pathway enrichment: the connections are shown using red nodes (pathways) or brown nodes (DEGs) through the brown edges in a circle. The larger the size of the grey node, the more connected it is within the network. The density of red color indicates the number of connecting DEGs. (B) Gene Ontology cellular component terms. (C) Gene Ontology biological process terms. (D) Gene Ontology molecular function terms. Significant pathways represent adjusted $P<.05$ (false discovery rate).

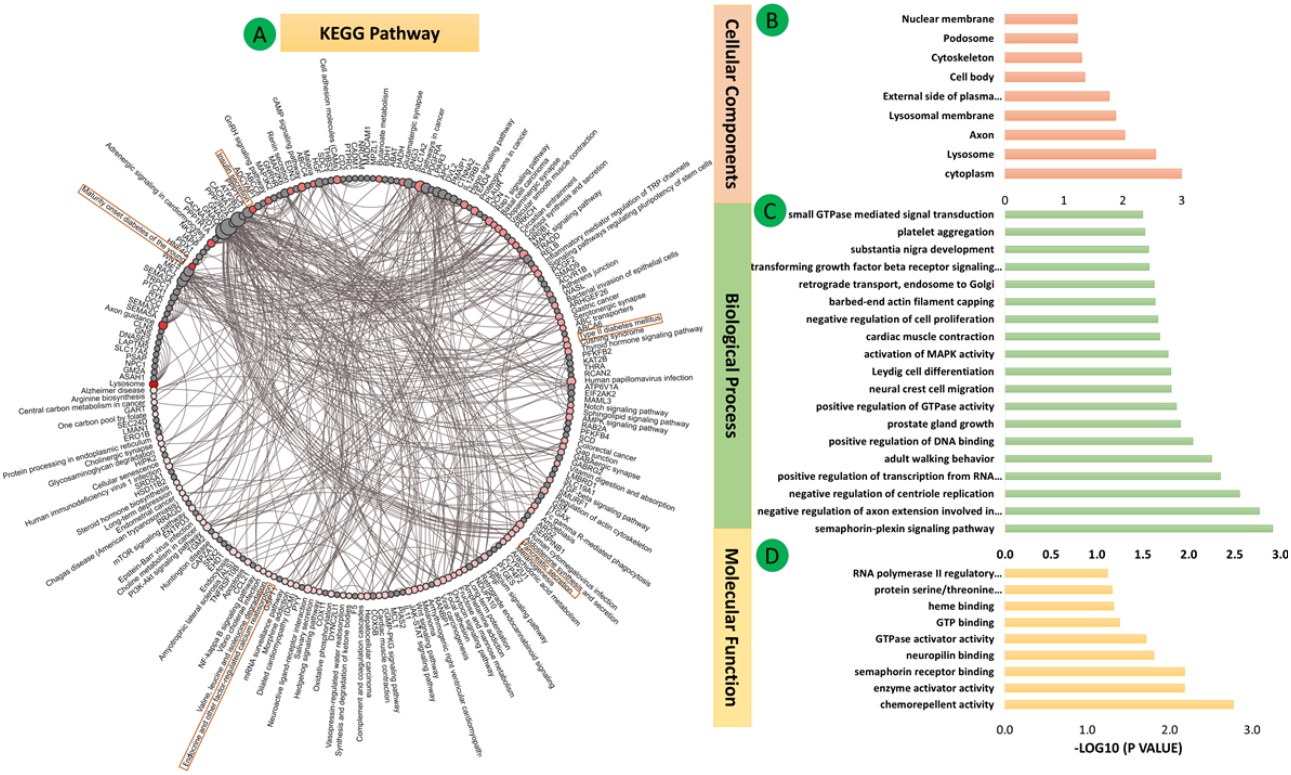


Figure 5. (A-D) Heat maps of mRNA expression for the three data sets (GSE38642, GSE25724, GSE20966) of Group 1 (pancreas) and the GSE26887 (heart) data set showing all differentially expressed genes (DEGs), DEGs regulated in opposite directions, upregulated DEGs, and downregulated DEGs. (E-G) Venn diagrams of complete, upregulated, and downregulated common DEGs. The upper part of the heat map shows fold change expression (FCE) values represented by varying color densities. (H-K) Violin plots showing the entire FCE distribution of all four data sets of Group 1 (pancreas) and the heart data set.

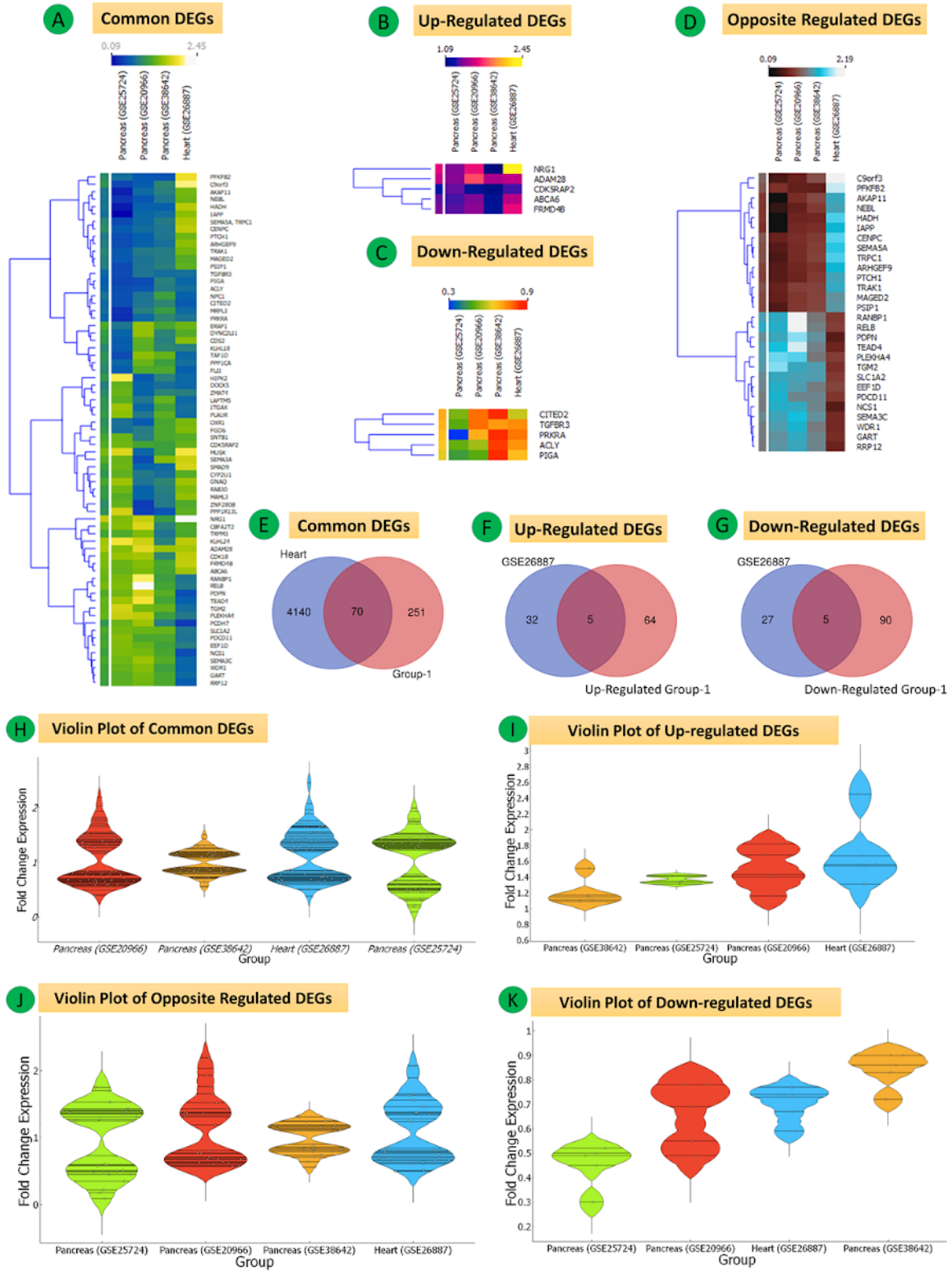


Figure 6. (A) Enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. The connections are shown using red nodes (pathways) or brown nodes (DEGs) through the brown edges in a circle. The larger the size of the grey node, the more connected it is within the network. The density of red color indicates the number of connecting DEGs. (B-D) Gene Ontology enrichment for (B) cellular component, (C) biological process, and (D) molecular function terms. Significant pathways represent adjusted $P<.05$ (false discovery rate). (E) Top hub genes in the network. (F, G) Clusters using MCODE.

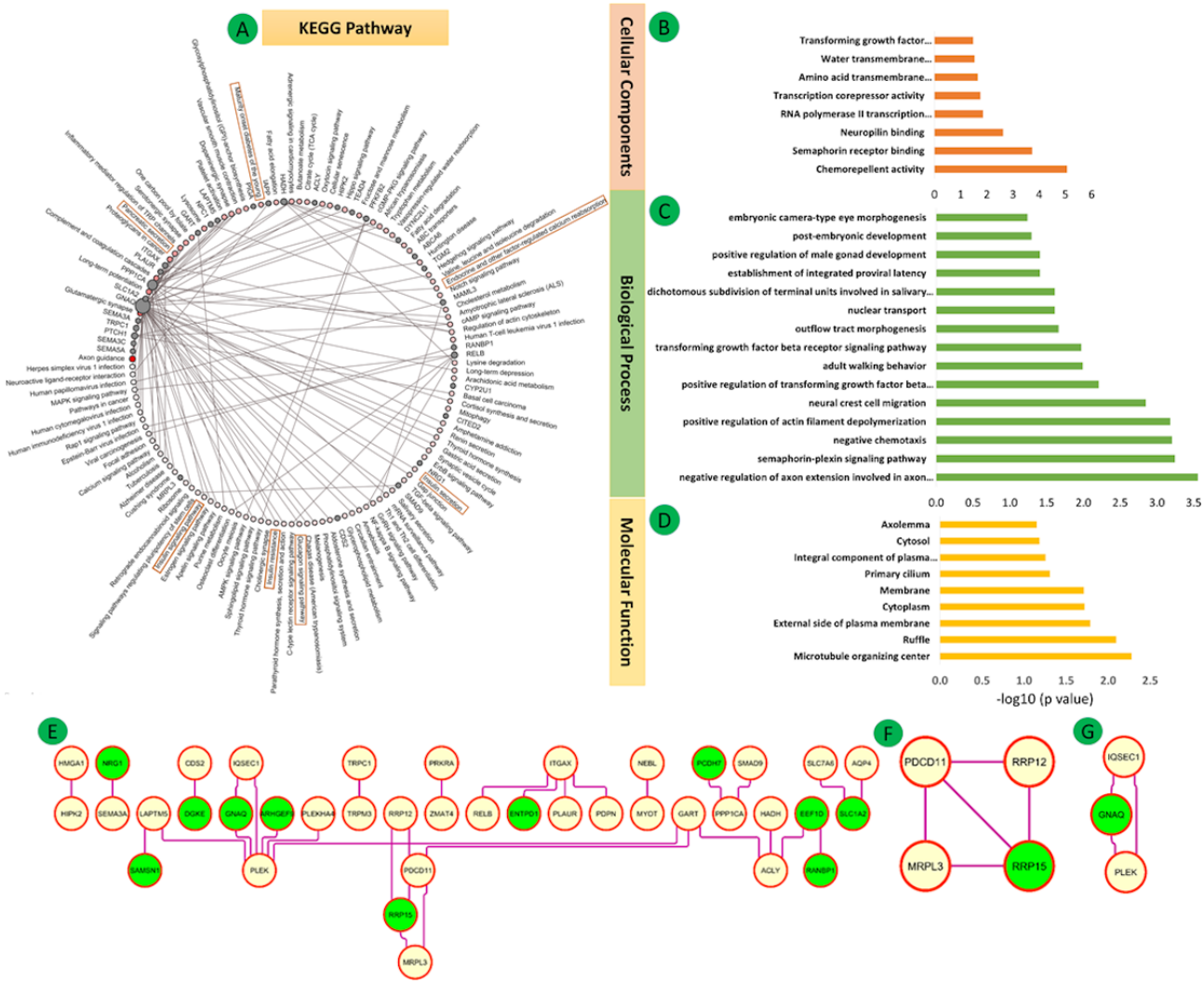


Figure 7. mRNA expression of three data sets (GSE38642, GSE25724, and GSE20966) of Group 1 (pancreas) and the GSE23343 data set (liver). (A-C) Venn diagrams of complete, upregulated, and downregulated common differentially expressed genes (DEGs). (D-F) Heat maps of common, oppositely regulated, and common upregulated DEGs. The upper part of the heat map shows the fold change in expression values reflected by respective color densities. (G-I) Violin plots showing the entire fold change expression (FCE) distribution of all four data sets for complete common DEGs, oppositely regulated DEGs, and common upregulated DEGs.

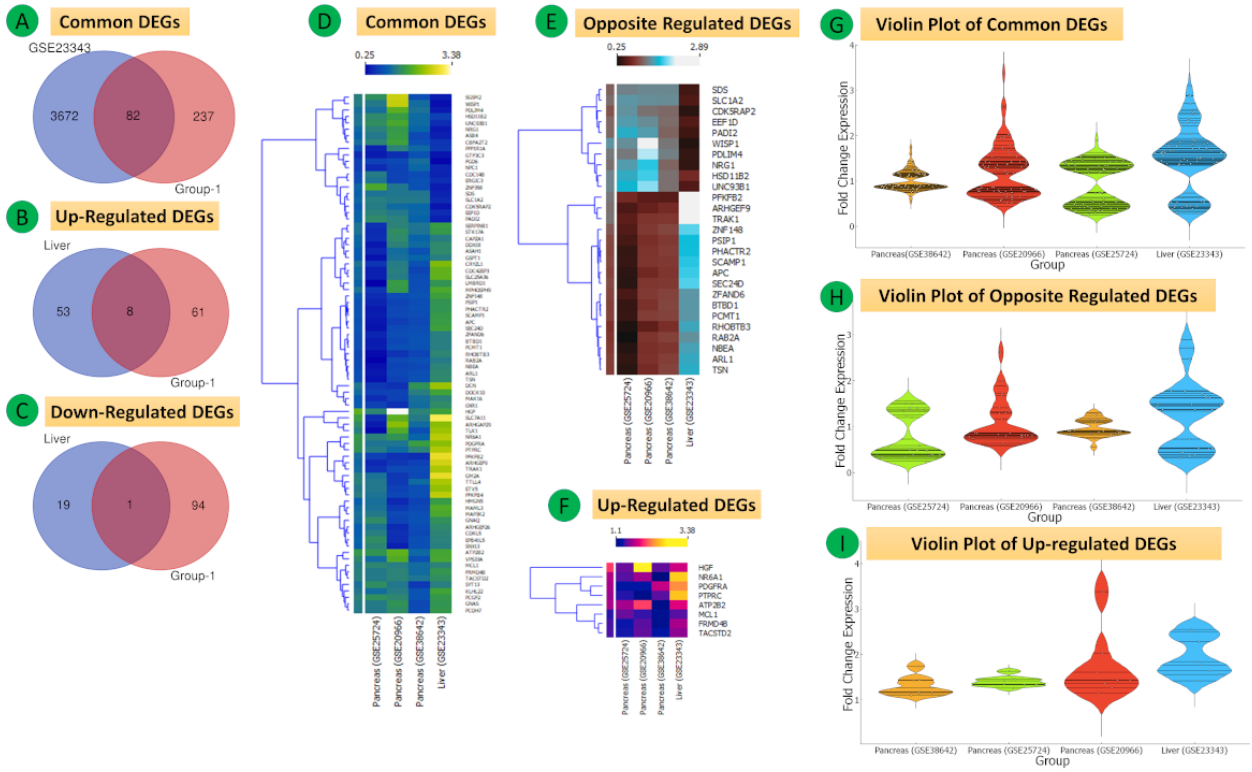


Figure 8. (A-D) Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway and Gene Ontology functional enrichment analysis of common DEGs. The connections are shown using red nodes (pathways) or brown nodes (DEGs) through the brown edges in a circle. The larger the size of the grey node, the more connected it is within the network. The density of red color indicates the number of connecting DEGs. (E) Protein-protein interaction networks of 82 overlapping DEGs of GSE23343 and co-expressed genes of Group 1 (321 DEGs) composed of 82 nodes and 56 edges. (F, G) Clusters from the network. Significant pathways represent adjusted $P<.05$ (false discovery rate).

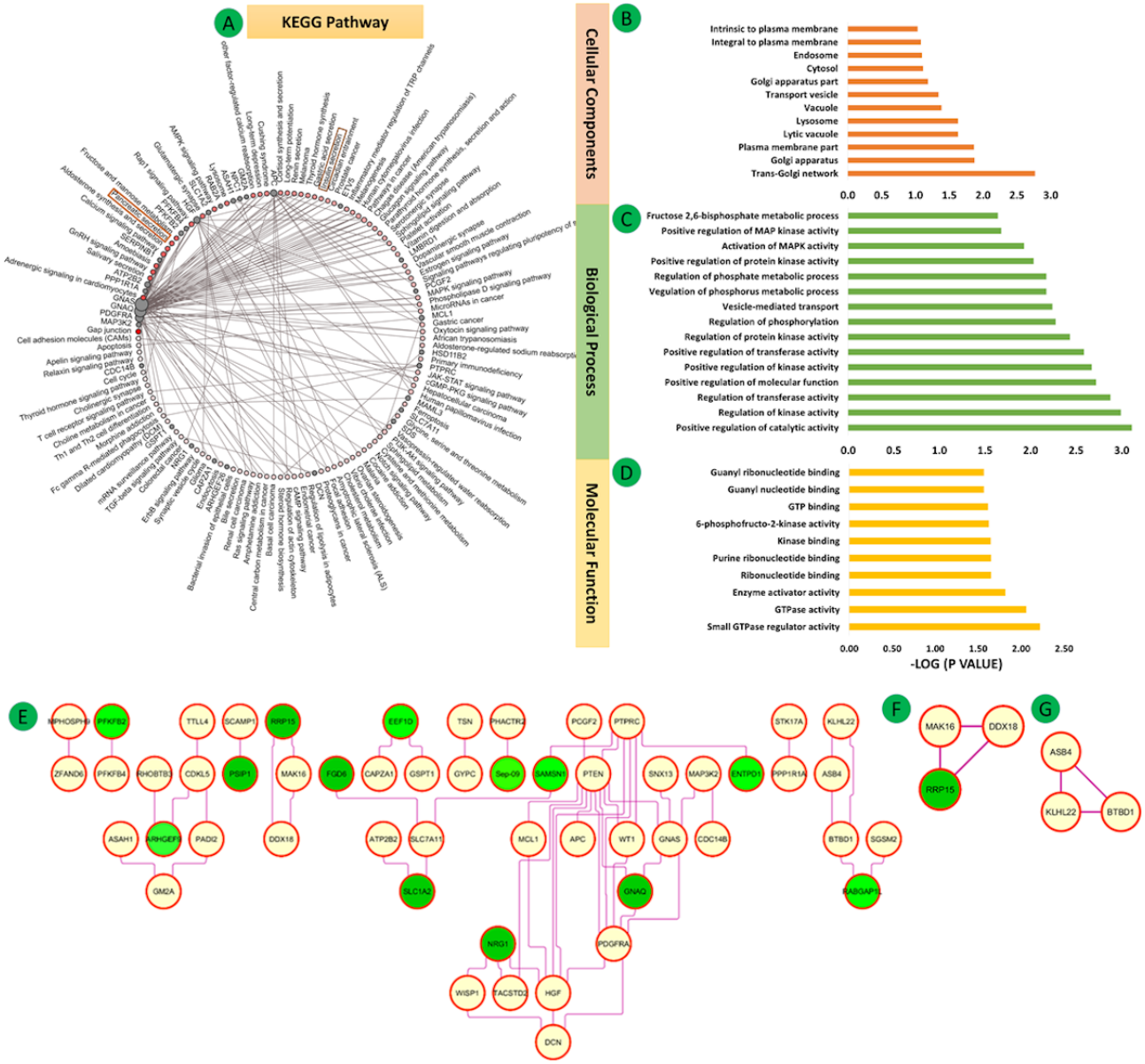


Figure 9. (A) Common and the top hub genes (green) in the protein-protein interaction network. (B-D) Clusters of the network.

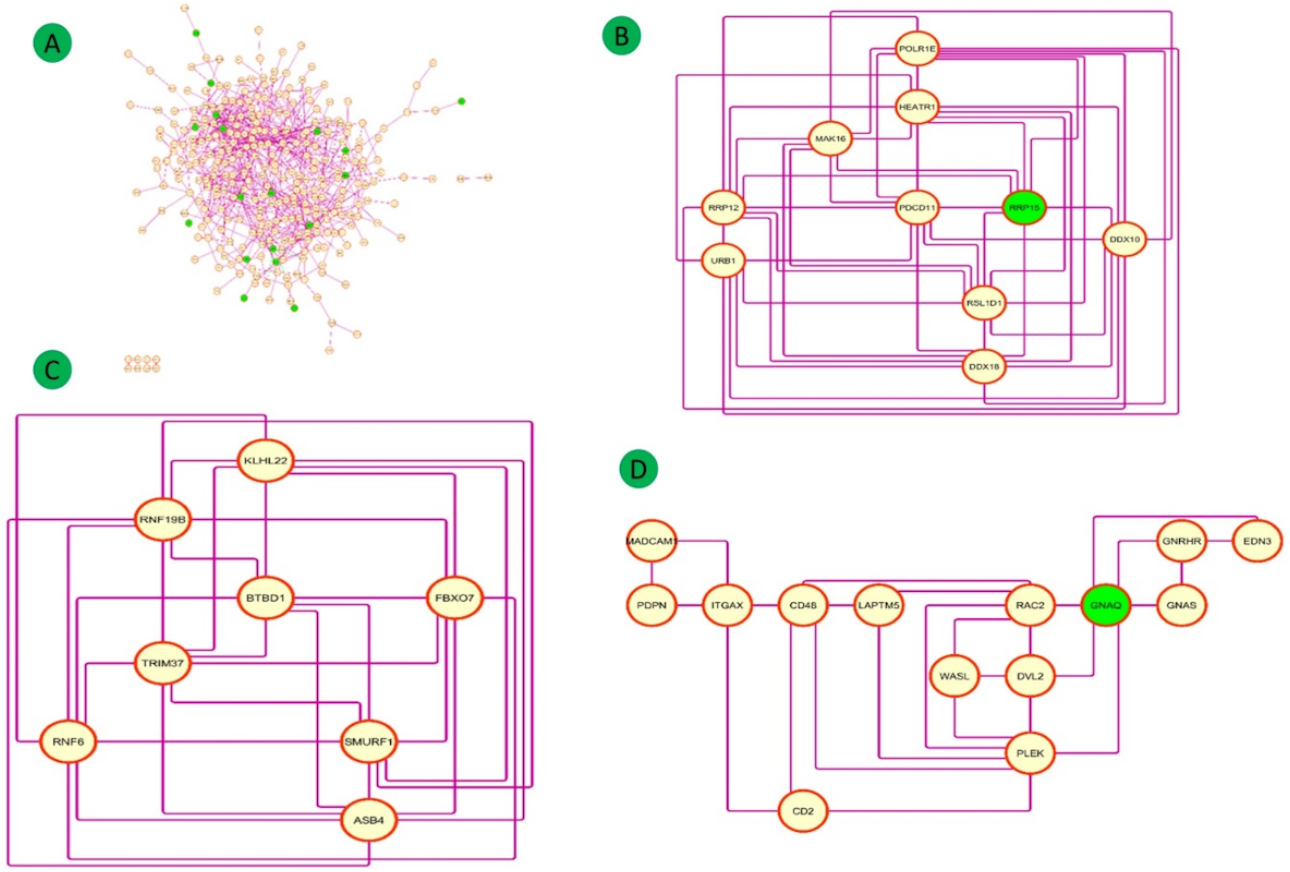


Figure 10. (A) Heat map of 16 common seed genes from the five data sets (pancreas, heart, and liver). The fold change expression levels are displayed in ascending order from blue to yellow. (B) Violin plot showing the entire fold change expression (FCE) distribution of all 16 common seed genes. (C) Venn diagram of common differentially expressed genes (DEGs). (D) Disease-gene interaction network. (E) Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways.

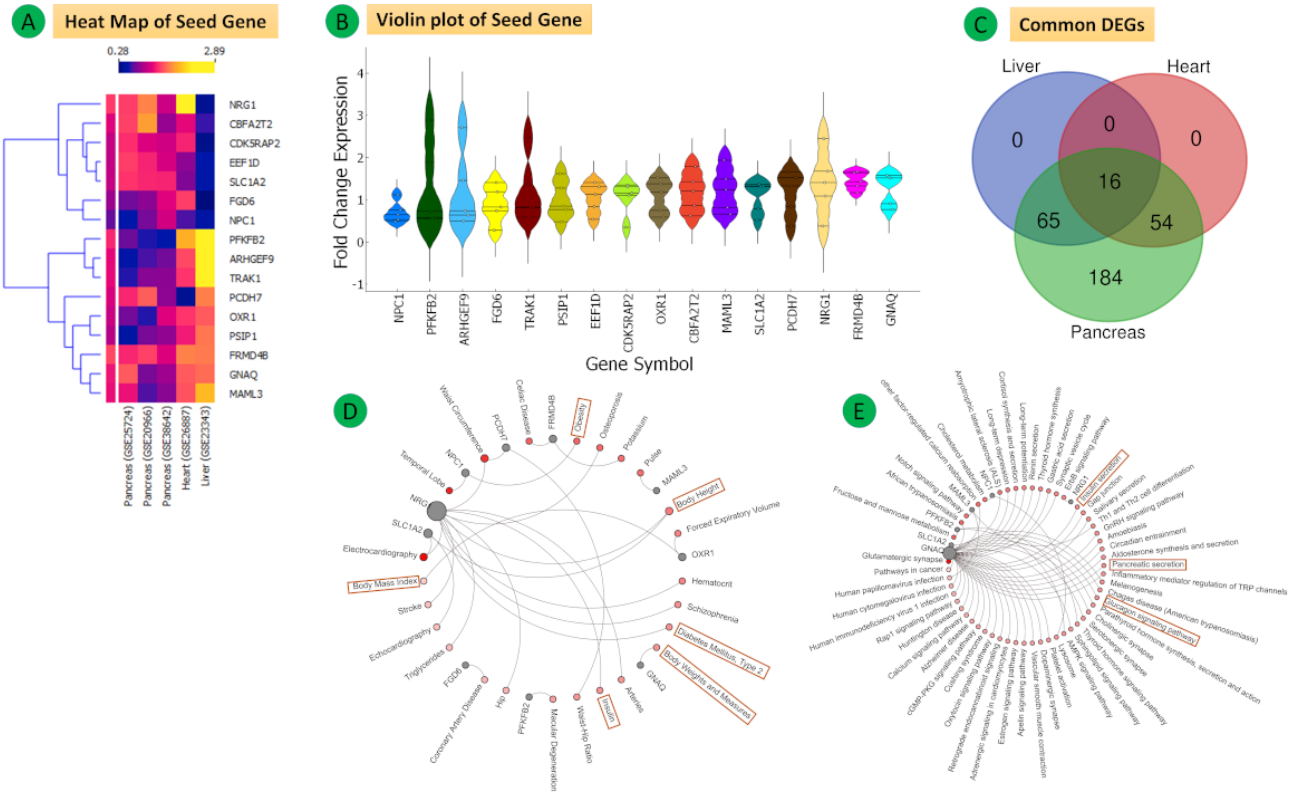


Figure 11. (A) Fold change expression levels of 16 common DEGs. (B) Top hub genes in the network (green) according to the criterion. (C, D, E, F) Clusters determined using MCODE.

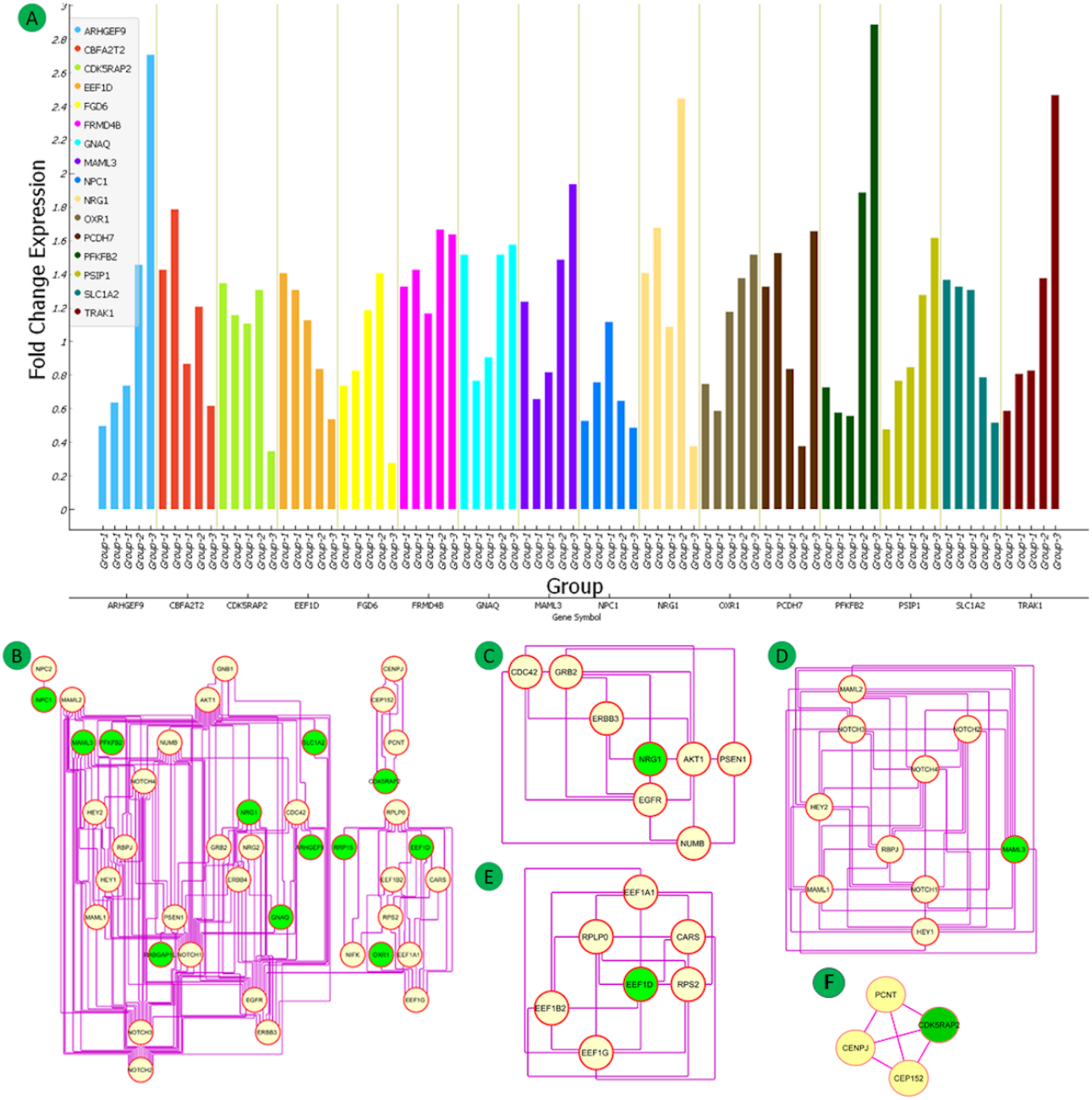


Figure 12. Protein-microRNAs interactions (top 10 ranked) for (A) pancreas data sets, (B) heart data set, (C) liver data set, and (D) 16 common differentially expressed genes (DEGs) of all five data sets.

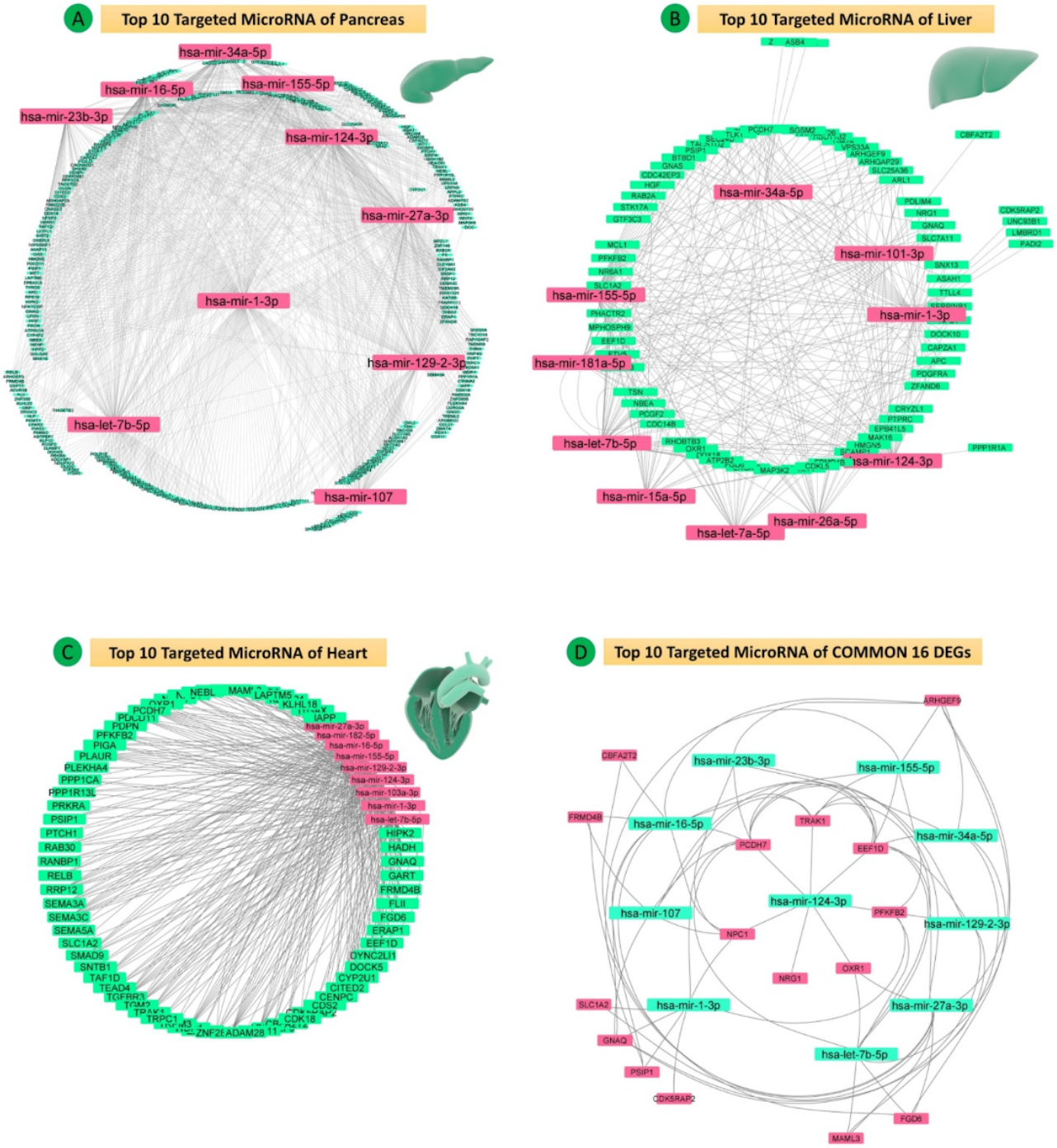


Figure 13. (A-E) Venn diagrams for (A) all microRNAs, (B) hsa-miR-1-3p, (C) hsa-miR-1-5p, (D) hsa-miR-155-5p, and (E) hsa-let-7b-5p with 16 common differentially expressed genes of the five data sets. (F) MicroRNA enrichment analysis of top four common targeted microRNAs.

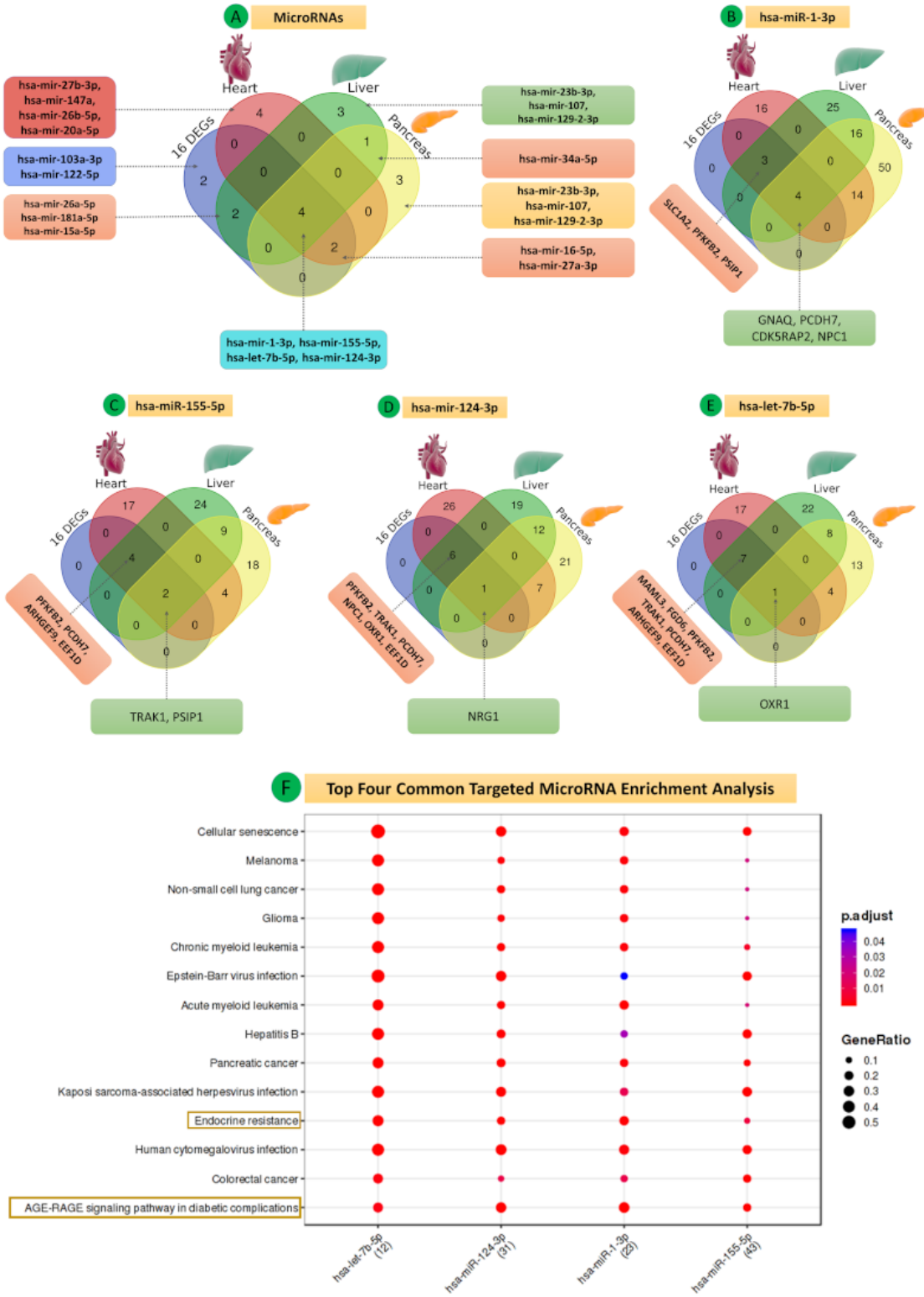
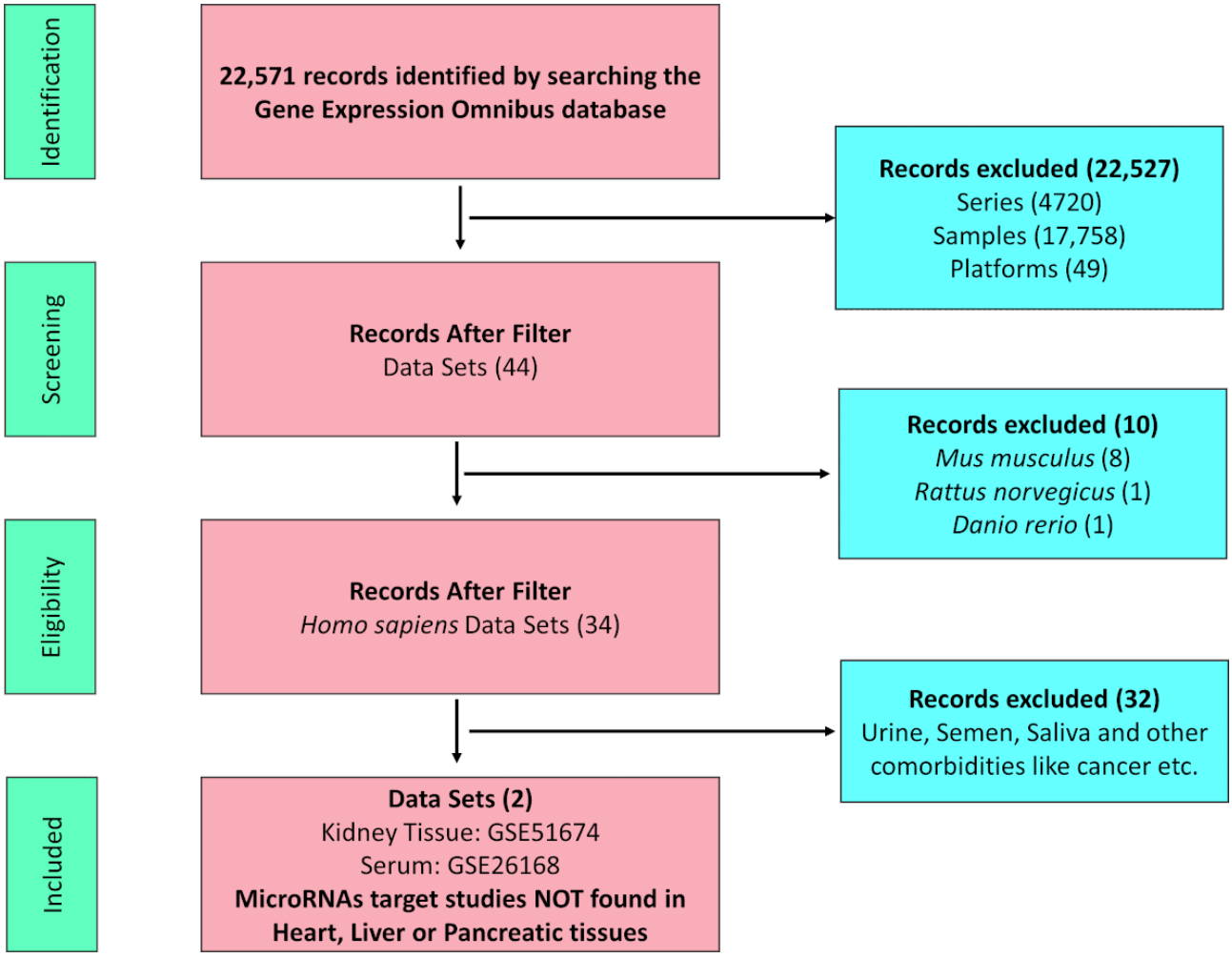


Figure 14. Flow diagram for the process of microRNA data collection with the number of data sets considered for inclusion.



Functional Enrichment and KEGG Pathway Analysis

The enrichments for the three GO classes (BP, CC, and MF) of the 321 DEGs of Group 1 are shown in Figure 4B-D (also see Tables S12-S14 of Multimedia Appendix 1). KEGG pathway analysis showed that these genes were enriched in maturity-onset diabetes of the young, malaria, lysosome, insulin secretion, adrenergic signaling in cardiomyocytes, cell adhesion molecules, and T2DM pathways (Figure 4A, Table S15 of Multimedia Appendix 1).

The enrichments for the three GO classes of the 70 DEGs of Group 2 are shown in Figure 6B-D (also see Tables S16-S18 in Multimedia Appendix 1). The genes were mainly enriched in gap junction, melanoma, calcium signaling pathway, and GnRH signaling pathway (Figure 6A and Table S19 of Multimedia Appendix 1).

The enrichment terms for the three GO classes for the 82 DEGs in Group 3 are shown in Figure 8B-D (also see Tables S20-S22 in Multimedia Appendix 1). These genes were enriched in axon guidance (Figure 8A and Table S23 in Multimedia Appendix 1).

PPI Network and Hub Gene Identification

Group 1

The 321 overlapping DEGs of the GSE38642, GSE25724, GSE20966 pancreas data sets were used to establish the PPI network, which constituted 321 nodes, 737 edges, and a PPI enrichment *P* value <.001 at medium confidence (0.4) (Figure 9A). The top three significant clusters within the PPI were selected.

Cluster 1 (MCODE Score=9.556, 10 nodes, 43 edges) included the genes *POLR1E*, *DDX10*, *URB1*, *HEATR1*, *DDX18*, *PDCD11*, *RSL1D1*, *RRP12*, *MAK16*, and *RRP15*, which are mainly associated with insulin pathway, transforming growth factor (TGF)-β receptor signaling, and the mammalian target of rapamycin (mTOR) signaling pathway (Figure 9B).

Cluster 2 (MCODE score=8.000, 8 nodes, 28 edges) included the genes *TRIM37*, *BTBD1*, *RNF19B*, *ASB4*, *KLHL22*, *SMURF1*, *FBXO7*, and *RNF6*, which are mainly associated with insulin pathway, insulin-like growth factor 1 (IGF1) pathway, class I phosphatidylinositol-3-kinase (PI3K) signaling events mediated by protein kinase B (AKT), TGF-β receptor signaling, mTOR signaling pathway, platelet-derived growth factor receptor-beta signaling pathway, and epidermal growth factor (EGF) receptor (ERBB1) signaling pathway (Figure 9C).

Cluster 3 (MCODE score=4.000, 14 nodes, 26 edges) included the genes *CD2*, *CD48*, *EDN3*, *GNAS*, *ITGAX*, *PDPN*, *GNRHR*, *RAC2*, *MADCAM1*, *WASL*, *GNAQ*, *PLEK*, *LAPTM5*, and *DVL2*, which are associated with platelet activation, signaling, and aggregation; hemostasis, cell surface interactions at the vascular wall; integrin family cell surface; and IGF1 pathway (Figure 9D).

Group 2

The 70 overlapping DEGs of GSE26887 and coexpressed genes with Group 1 (321 genes) were used to establish the PPI network composed of 70 nodes, 32 edges, and a PPI enrichment *P* value of .05 at medium confidence (0.4). The top two significant clusters within the PPI were selected using the MCODE plugin of Cytoscape software (Figure 6E).

Cluster 1 (MCODE score=3.333, 4 nodes, 5 edges) included the genes *RRP12*, *PDCD11*, *RRP15*, and MRPL3 (Figure 6F). Cluster 2 (MCODE score=3.000, 3 nodes, 3 edges) included the genes *PLEK*, *GNAQ*, and *IQSEC1*, which are mainly associated with platelet activation, signaling, and aggregation; hemostasis; class I PI3K signaling events mediated by AKT; insulin pathway; mTOR signaling pathway; IGF1 pathway; and EGF receptor (ERBB1) signaling pathway (Figure 6G).

Group 3

The 82 overlapping DEGs of GSE23343 and the coexpressed genes of Group 1 (321 DEGs) were used to establish the PPI network composed of 82 nodes, 56 edges, and a PPI enrichment *P* value of .02 at medium confidence (0.4). The top two significant clusters are shown in Figure 8E.

Cluster 1 (MCODE score=03, 3 nodes, 3 edges) included the genes *BTBD1*, *ASB4*, and *KLHL22*, which were mainly associated with PI3K/AKT signaling in cancer (Figure 8F).

Cluster 2 (MCODE score=03, 3 nodes, 3 edges) included the genes *DDX18*, *MAK16*, and *RRP15*, which were mainly associated with insulin pathway, mTOR signaling pathway, IGF1 pathway, and EGF receptor (ERBB1) signaling pathway (Figure 8G).

Common Genes Among All Groups

A total of 16 overlapping DEGs were identified in all three groups. The hub genes of all data sets were *ARHGEF9*, *CBFA2T2*, *CDK5RAP2*, *EEF1D*, *FGD6*, *FRMD4B*, *GNAQ*, *MAML3*, *NPC1*, *NRG1*, *OXR1*, *PCDH7*, *PFKFB2*, *PSIP1*, *SLC1A2*, and *TRAK1* (Table S24 in Multimedia Appendix 1). All 16 hub genes belonging to the five data sets were analyzed with the help of an expression heat map, violin plot, and Venn diagram, and their fold change expression levels were compared by bar plots and analyzed by the disease-gene interaction network and KEGG pathway (Figure 10A-E and Figure 10F; Table S25 in Multimedia Appendix 1).

The PPI network of the 16 hub genes and their related genes was established by protein STRING analysis. We selected 4 clusters from the PPI network using MCODE (Figure 11B). Cluster 1 (MCODE score=10, 10 nodes, 45 edges) included the genes *MAML1*, *HEY2*, *NOTCH3*, *MAML3*, *NOTCH2*, *MAML2*, *NOTCH1*, *HEY1*, *RBPJ*, and *NOTCH4*. The analysis also showed that cluster 1 contains *MAML3* as a seed gene (Figure 11D). Cluster 2 (MCODE score=6.667, 7 nodes, 20 edges) included the genes *EEF1A1*, *EEF1B2*, *EEF1G*, *RPLP0*, *RPS2*, *CARS*, and *EEF1D*, with *EEF1D* as a seed gene (Figure 11E). Cluster 3 (MCODE score=5.714, 8 nodes, 20 edges) included the genes *AKT1*, *NUMB*, *EGFR*, *ERBB3*, *GRB2*, *CDC42*, *PSEN1*, and *NRG1*, with *NRG1* as a seed gene (Figure 11C). Cluster 4 (MCODE score=4, 4 nodes, 6 edges) included the genes *CDK5RAP2*, *CEP152*, *CENPJ*, and *PCNT*, with *CDK5RAP2* as the seed gene (Figure 11F).

Integrative Gene Expression and Meta-analysis

The number of genes with an adjusted *P* value <.05 for each data set revealed 4, 0, 3533, 171, and 1 significant genes from the meta-analysis, including *ARHGEF9*, *SAMSN1*, *SLC1A2*, *RABGAP1L*, *OXR1*, *GNAQ*, *CBFA2T2*, and *RRP15*. The 16 hub genes obtained from the gene expression meta-analysis are shown in Table S26 of Multimedia Appendix 1.

MicroRNA and Hub Gene Network

To investigate the regulatory relationship of the identified hub genes, their targeting miRNAs, and coexpressed network, the top 10 ranked DEG-targeting miRNAs were selected based on degree and betweenness values. The top 10 targeting miRNAs for the three groups were hsa-let-7b-5p, hsa-mir-107, hsa-mir-124-3p, hsa-mir-129-2-3p, hsa-mir-1-3p, hsa-mir-155-5p, hsa-mir-16-5p, hsa-mir-23b-3p, hsa-mir-27a-3p, and hsa-mir-34a-5p in Group 1 (pancreas); hsa-mir-16-5p, hsa-mir-124-3p, hsa-mir-1-3p, hsa-mir-27a-3p, hsa-let-7b-5p, hsa-mir-155-5p, hsa-mir-20a-5p, hsa-mir-26b-5p, hsa-mir-27b-3p, and hsa-mir-147a in Group 2 (heart); and hsa-mir-1-3p, hsa-mir-155-5p, hsa-mir-124-3p, hsa-let-7b-5p, hsa-mir-34a-5p, hsa-mir-101-3p, hsa-mir-15a-5p, hsa-mir-26a-5p, hsa-mir-181a-5p, and hsa-let-7a-5p in Group 3 (liver). The common hub genes were targeted by hsa-mir-16-5p, hsa-mir-27a-3p, hsa-let-7a-5p, hsa-let-7b-5p, hsa-mir-101-3p, hsa-mir-1-3p, hsa-mir-124-3p, hsa-mir-103a-3p, hsa-mir-122-5p, and hsa-mir-155-5p.

Four common miRNAs (hsa-let-7b-5p, hsa-mir-155-5p, hsa-mir-124-3p, hsa-mir-1-3p) were found in all three groups, targeting the 16 hub DEGs. The miRNAs and PPI networks representing multiple targeted nodes (DEGs) of particular miRNAs for all groups are shown in Figure 12A-D.

The common DEGs found in all three groups are targeted by hsa-miR-1-3p (*GNAQ*, *PCDH7*, *CDK5RAP2*, *NPC1*), hsa-let-7b-5p (*OXR1*), hsa-mir-155-5p (*TRAK1*, *PSIP1*), and hsa-mir-124-3p (*NRG1*). The common targeting important miRNAs (hsa-let-7b-5p, hsa-mir-155-5p, hsa-mir-124-3p, hsa-mir-1-3p) were mainly involved in the advanced glycation end products (AGE)-receptor for advanced glycation end products (RAGE) signaling pathway in diabetic complication and endocrine resistance (Figure 13A-F, Tables S27-S40 in Multimedia Appendix 1).

Target MiRNA Validation from Available Data Sets

To validate our miRNA prediction, we searched the database again and performed a thorough review of available miRNA data sets for T2DM. Our search yielded two miRNA data sets

from renal tissue (GSE51674) and serum (GSE26168) samples. The flow diagram for the miRNA data set search is shown in Figure 14. However, we were not able to find any miRNA data set pertaining to the heart, pancreas, or liver tissue. Interestingly, on analysis of the data sets obtained from the renal tissue and serum, we observed a significant alteration for our predicted miRNAs in the renal tissue, which was conspicuously absent in the serum (Table 2). We assessed the expression of our predicted miRNAs in the renal tissue and serum by comparing the adjusted *P* values for both sample types. This analysis revealed that although the expression of miRNAs was significantly altered in renal tissues from patients with T2DM, the same was not observed in serum when compared with healthy controls. Our analysis highlights a paradoxical difference in the alteration of miRNAs in tissue and serum in T2DM.

Table 2. Validation of the fold change in expression levels of common microRNAs in the GSE51674 (kidney) and GSE26168 (serum) data sets.

MicroRNA	Adjusted *P* value	*P* value	*t*	B	FC[a]	logFC
GSE51674 (kidney)						
hsa-miR-124*[b]	<.001	<.001	–4.83	–0.85	0.50	–1.00
hsa-miR-1	<.001	<.001	6.06	0.92	19.42	4.28
hsa-miR-155	<.001	<.001	20.96	12.12	74.56	6.22
hsa-let-7b	<.001	<.001	4.60	–1.20	2.55	1.35
GSE26168 (serum)						
hsa-miR-124*	.46	.23	1.23	–6.22	1.01	0.02
hsa-miR-1	.86	.82	0.23	–6.95	1.00	0.00
hsa-miR-155	.46	.08	1.86	–5.33	1.04	0.05
hsa-let-7b	.46	.16	1.45	–5.94	237.61	7.89

[a]FC: fold change.

[b]*indicates the star strand for miR-124.

Functional Enrichment of MiRNAs

The functional enrichment and pathway analysis by MIENTURNET revealed the top significant pathways for hsa-let-7b-5p, hsa-miR-124-3p, hsa-miR-1-3p, and hsa-miR-155-5p, including the PI3K-AKT signaling pathway (hsa-let-7b-5p, hsa-miR-124-3p, hsa-miR-1-3p, hsa-miR-155-5p), endocrine resistance (hsa-let-7b-5p, hsa-miR-124-3p, hsa-miR-1-3p, hsa-miR-155-5p), AGE-RAGE signaling pathway in diabetic complications (hsa-let-7b-5p, hsa-miR-124-3p, hsa-miR-1-3p, hsa-miR-155-5p), lipid and atherosclerosis (hsa-let-7b-5p, hsa-miR-1-3p, hsa-miR-155-5p), insulin signaling pathway (hsa-let-7b-5p, hsa-miR-124-3p, hsa-miR-1-3p), mitogen-activated protein kinase (MAPK) signaling pathway (hsa-let-7b-5p, hsa-miR-124-3p, hsa-miR-155-5p), fluid sheer stress and atherosclerosis (hsa-miR-124-3p, hsa-miR-1-3p), adipocytokine signaling pathway (hsa-miR-124-3p), diabetic cardiomyopathy (hsa-miR-124-3p, hsa-miR-1-3p), insulin resistance (hsa-miR-124-3p), carbohydrate digestion and absorption (hsa-miR-124-3p), regulation of lipolysis in adipocytes (hsa-miR-124-3p), glucagon signaling pathway (hsa-miR-124-3p), and TGF-β signaling pathway (hsa-miR-155-5p) (see Figure 13 and Table S41 of Multimedia Appendix 1).

Discussion

Principal Findings

Diabetes develops because of dysregulated β-cell and adipose-tissue responses to chronic fuel excess, which result in so-called nutrient spillover, insulin resistance, and metabolic stress. The latter causes multiple organ damage. However, insulin resistance, while forcing β-cells to work harder, may also have an important defensive role against nutrient-related toxic effects in tissues such as the heart [32]. The liver, which primarily regulates glucose homeostasis in the body, has a strong association with diabetes. Liver disease in diabetes can further be classified into liver disease related to diabetes, hepatogenous diabetes, and liver disease occurring coincidentally with diabetes mellitus [33]. Recently, knowledge on the pathogenesis and management of diabetes mellitus has been expanding; however, the disease is far from being effectively managed in a large proportion of patients. In silico analysis of disease pathways and exploration of various disease-related genes and their regulatory molecules have revealed unforeseen vistas. In this study, we analyzed tissue-specific microarray gene expression data sets from publicly available repositories employing a network-based bioinformatics pipeline. We identified DEGs common to different tissues of patients with T2DM and constructed diseasome networks to provide insights into the interactions of the genes. These DEGs enabled the identification of associated dysregulated molecular pathways in tissues and related GO terms. A large number of pathways and GO categories were reduced by manual curation after filtering using a *P* value threshold of .05.

Our analysis supports that diabetes is a multifactorial disease caused by multiple complex systems, with an abundant crossover between signaling pathways. For each data set included in the study, comprehensive analysis focusing on biological function and interaction of T2DM-related genes provided valuable

information to understand the pathogenic effect of DEGs in various organs, including the heart, liver, and pancreas, of patients with diabetes. In this study, five mRNA expression profile data sets (GSE38642, GSE25724, GSE20966, GSE26887, and GSE23343), including 125 samples of the pancreas, heart, and liver tissues of patients with T2DM and controls without diabetes, were analyzed. A total of 16 seed genes were obtained after the final analysis. Some of these genes have been reported to play significant roles in T2DM and its related comorbidities. In a similar study that included DEG screening from a genome-wide association study (GWAS) catalog, Gupta and Vadde [34] identified four hub gene candidates, related signaling pathways, target miRNAs, and transcription factors. However, their selection criteria of the data sets chosen for analysis were different than those adopted in this study, which possibly accounts for the difference in results.

Neuregulin 1 (NRG1) and ERBB receptors are involved in glucose homeostasis. NRG1-ERBB pathway activation affects glucose metabolism in the liver. Mice with chronic NRG1 treatment showed increased p38 phosphorylation in the liver and improved glucose tolerance [35]. Myocardial NRG1/ERBB is altered during postmyocardial infarction heart failure associated with diabetes. NRG1 can improve the antioxidative function of the mitochondria, and thereby increase the proliferation and decrease the apoptosis of cardiomyocytes via ERBB/AKT signaling. This can explain the upregulated expression of *NRG1* found in the cardiac tissue of patients with T2DM in our study. Moreover, the dysregulated insulin signaling pathway modifies titin-based cardiomyocyte tension, modulates diastolic function, impairs cyclic guanosine monophosphate (cGMP)–cGMP-dependent protein kinase signaling, and elevates protein kinase C-α activity, thereby causing titin-based cardiomyocyte stiffening in diabetic hearts. Chronic NRG1 application has shown promising results in the modulation of titin properties in T2DM-associated heart failure with a preserved ejection fraction [36]. Further, there are reports showing that hyperglycemia impairs NRG1/ERBB2 signaling by disrupting the balance between NRG1 isoforms, decreasing the expression of erbin, and correspondingly activating the MAPK pathway, ultimately aiding in the development of diabetic peripheral neuropathy [37]. Again, the downregulation of *NRG1* expression in the liver found in this study points toward dysregulated glucose homeostasis.

PFKFB2 encodes 6-phosphofructo-2-kinase/fructose 2,6-bisphosphatase (PFK2/FBPase-2) isoform 2, a bifunctional enzyme involved in the synthesis and degradation of fructose 2,6-bisphosphate. Enhanced hepatic glycolysis in mice achieved by overexpressing PFK2/FBPase-2 in the liver resulted in reduced body weight and visceral fat content. PFK2/FBPase-2 is also a binding partner for glucokinase, which plays a pivotal role in the rate-limiting step of glucose-stimulated insulin secretion in pancreatic β-cells, and regulates obesity, insulin secretory dysfunction, and T2DM [38,39]. The loss of PFK2 content as a result of reduced insulin signaling impairs its regulatory function of glycolysis and elevates the levels of early glycolytic intermediates. Although this may be beneficial in the fasting state to conserve systemic glucose, it represents a pathological impairment in diabetes mellitus [40]. Interestingly, *PFKFB2*, among a few other genes, showed opposing expression changes in the pancreas (downregulation) and heart (upregulation). This is likely due to the impaired insulin secretion pathway in pancreatic β-cells, in which PFKFB2 plays an important role [39]. Moreover, PFKFB2 is known to alleviate myocardial injury; hence, the increased expression level in the heart is possibly a protective mechanism [41].

CDK5 regulatory subunit associated protein (*CDK5RAP*) 1, 2, and 3 were all found to be differentially upregulated in four data sets, except GSE23343 in which these genes were downregulated. These genes have been associated with neuronal development and spindle checkpoint function [42]. FRMD4B plays a vital role in cardiac activity regulation. However, the effect varies in different populations due to polymorphisms. FRMD4B has shown to be associated with ischemic heart failure in a European population but not in other populations [43]. The G-protein Gq, encoded by *GNAQ*, is a crucial key regulator of the insulin secretion pathway that is involved in glucose metabolism, and a functional *GNAQ* promoter haplotype was associated with altered Gq expression and with insulin resistance and obesity in women with polycystic ovary syndrome [44]. The Niemann-Pick type C1 (NPC1) protein regulates the transport of cholesterol and fatty acids from late endosomes/lysosomes and has a central role in maintaining lipid homeostasis. In humans, GWAS and post-GWAS highlighted the implication of common variants in *NPC1* in adult-onset obesity, body fat mass, and T2DM. Heterozygous human carriers of rare loss-of-function coding variants in *NPC1* display an increased risk of morbid adult obesity [45]. Another significant DEG pair was orexin A and B, which regulate a variety of physiological functions. The biological effects of these neuropeptides occur through OXR1, a G-protein coupled receptor. There is growing evidence that orexins regulate body weight, glucose homeostasis, and insulin sensitivity, and promote energy expenditure, thus protecting against obesity by interacting with brown adipocytes. Further, orexins control brown and white adipocytes as well as pancreatic α- and β-cell functions [46,47]. Single-cell RNA sequencing from samples of patients with gestational diabetes mellitus revealed *SLC1A2* as a novel marker for syncytiotrophoblasts [48]. Such cell-type-specific marker genes in particular disease states can open new avenues of tissue-targeted therapeutic intervention. Among the other DEGs, *EEF1D* regulates lipid synthesis via the PI3K/AKT, PPAR, and AMPK pathways [49]. CBFA2T2 is a key regulator of adipogenic differentiation through CEBPA [50]. Further, these seed genes were analyzed as possible miRNA targets in silico, which revealed the top 10 miRNAs for each of the pancreas, liver, and heart tissues, as well as for the 16 seed genes. The role of miRNAs in the regulation of the underlying pathogenic mechanisms of diabetes and diabetic complications is well established [7,51] Several of the target miRNAs for the seed genes have already been explored in T2DM, and our in silico analysis further confirms their candidature as potential biomarkers as well as therapeutic targets. In fact, miR-124-3p was interconnected to 7 of the 16 seed genes. Pan et al [52] studied mouse primary hepatocytes and observed that regulation of miR-124-3p plays an important role in turning the hepatocytes into insulin-producing cells. A

recent analysis of weighted genes in diabetic retinopathy concluded miR-124-3p to be a pivotal regulatory molecule in the underlying pathogenesis [6]. Furthermore, in isolated myocardial cells, *NRG1* expression was observed to be downregulated while miR-124-3p expression was upregulated in ischemia/reperfusion injury [53], which also supports our finding of this miRNA-mRNA target interaction. The miRNA hsa-miR-124-3p affects the immune status of patients with T2DM through its interaction with the obesity-related immune cytokines [54].

Three other miRNAs, namely miR-155-5p, miR-1-3p, and let-7b-5p, were also commonly identified in all three groups. Likewise, the role of miR-155-5p in diabetes has been widely studied, especially as a marker in diabetic kidney disease (DKD) [55-57]. The expression of miR-155-5p is positively associated with urinary microalbumin and has good diagnostic and prognostic value in patients with DKD [56]. Further, dihydromyricetin attenuates renal interstitial fibrosis by regulating PTEN signaling, a critical element in the pathogenesis of DKD, through miR-155-5p [58,59]. Recently, Zhou and colleagues [60] showed that metformin can relieve inflammation and fibrosis in patients with DKD by acting through an inflammation axis involving miR-155-5p. Some recent studies have also shown that miR-155-5p interferes with immune dysregulation in COVID-19 patients with diabetes or other comorbidities [61,62]. Further, all four miRNAs were found to be involved in regulating the endocrine resistance and AGE-RAGE pathways, which is in line with recent findings [63].

The differing trend in miRNA expression observed in our comparison of miRNA data sets from serum and renal tissue in T2DM highlights the necessity to further explore the tissue-specific alterations in T2DM to better comprehend its role in various tissues.

Limitations

The main limitation of this study is that it was based on an in silico analysis; therefore, further validation of the identified novel hub genes and miRNAs is still required based on laboratory experiments with human T2DM samples. The data sets were compiled using different arrays on the Affymetrix platform, and the patient populations belong to multiple ethnic groups, which may account for some of the variability in the results. Furthermore, the predicted miRNAs in this study could not be validated within the same tissue data sets. However, the functional enrichment for the miRNAs highlighted some significant pathways related to T2DM, its complications, and its pathogenic mechanisms.

Conclusion

The aim of this study was to identify the tissue-specific differential expression of genes, especially pertaining to the heart, liver, and pancreas, in T2DM. From Group 1 (pancreas: 374 DEGs), Group 2 (heart: 86 DEGs), and Group 3 (liver: 97 DEGs), we identified a total of 16 common DEGS (*ARHGEF9*, *CBFA2T2*, *CDK5RAP2*, *EEF1D*, *FGD6*, *FRMD4B*, *GNAQ*, *MAML3*, *NPC1*, *NRG1*, *OXR1*, *PCDH7*, *PFKFB2*, *PSIP1*, *SLC1A2*, and *TRAK1*) in the selected data sets. Further, we identified the top four common miRNAs (hsa-let-7b-5p, hsa-miR-124-3p, hsa-miR-1-3p, has-miR-155-5p) targeting the 16 common DEGs. Although we were not able to find any miRNA data set pertaining to the heart, pancreas, or liver tissue, we observed significant alterations of our predicted miRNAs in renal tissue. Interestingly, this significant alteration was conspicuously absent in the serum. The miRNAs identified in this study are involved in regulating various pathways, including the PI3K-AKT signaling pathway, endocrine resistance, and the AGE-RAGE signaling pathway. Moreover, the differing trend in miRNA expression observed in our comparison of miRNA data sets from the serum and renal tissue in T2DM highlights the necessity to further explore the tissue-specific alteration in T2DM to better comprehend its role in various tissues.

Acknowledgments

The authors are grateful to All India Institute of Medical Sciences Jodhpur for providing the research facility to perform this in silico experiment. MK is supported by a senior research fellowship of The University Grants Commission of India (NOV2017-361200).

Authors' Contributions

Concept and design: MK, DR, PP; data acquisition, analysis, and interpretation: MK, ST, AG, PP; manuscript drafting: MK, DR, ST, AG, PS, PP; manuscript revision: MK, DR, ST, PS, PP; project supervision: PP.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary data: Figures S1-S2, Tables S1-S41.
[PDF File (Adobe PDF File), 1836 KB - bioinform_v3i1e32437_app1.pdf]

References

1. Behera PM, Behera DK, Panda A, Dixit A, Padhi P. In silico expressed sequence tag analysis in identification of probable diabetic genes as virtual therapeutic targets. Biomed Res Int 2013;2013:704818. [doi: 10.1155/2013/704818] [Medline: 23509765]
2. Hruby A, Hu FB. The epidemiology of obesity: a big picture. Pharmacoeconomics 2015 Jul 4;33(7):673-689 [FREE Full text] [doi: 10.1007/s40273-014-0243-x] [Medline: 25471927]
3. DF Diabetes Atlas Ninth edition 2019. International Diabetes Federation. URL: https://diabetesatlas.org/upload/resources/material/20200302_133351_IDFATLAS9e-final-web.pdf [accessed 2021-07-09]
4. Vaidya V, Gangan N, Sheehan J. Impact of cardiovascular complications among patients with Type 2 diabetes mellitus: a systematic review. Expert Rev Pharmacoecon Outcomes Res 2015 Jun 31;15(3):487-497. [doi: 10.1586/14737167.2015.1024661] [Medline: 25824591]
5. Pordzik J, Jakubik D, Jarosz-Popek J, Wicik Z, Eyileten C, De Rosa S, et al. Significance of circulating microRNAs in diabetes mellitus type 2 and platelet reactivity: bioinformatic analysis and review. Cardiovasc Diabetol 2019 Aug 30;18(1):113 [FREE Full text] [doi: 10.1186/s12933-019-0918-x] [Medline: 31470851]
6. You Z, Zhang Y, Li B, Zhu X, Shi K. Bioinformatics analysis of weighted genes in diabetic retinopathy. Invest Ophthalmol Vis Sci 2018 Nov 01;59(13):5558-5563. [doi: 10.1167/iovs.18-25515] [Medline: 30480744]
7. Roy D, Modi A, Khokhar M, Sankanagoudar S, Yadav D, Sharma S, et al. MicroRNA 21 emerging role in diabetic complications: a critical update. Curr Diabetes Rev 2021 Feb;17(2):122-135. [doi: 10.2174/1573399816666200503035035] [Medline: 32359340]
8. Plenge RM. Disciplined approach to drug discovery and early development. Sci Transl Med 2016 Jul 27;8(349):349ps15. [doi: 10.1126/scitranslmed.aaf2608] [Medline: 27464747]
9. Taneera J, Lang S, Sharma A, Fadista J, Zhou Y, Ahlqvist E, et al. A systems genetics approach identifies genes and pathways for type 2 diabetes in human islets. Cell Metab 2012 Jul 03;16(1):122-134 [FREE Full text] [doi: 10.1016/j.cmet.2012.06.006] [Medline: 22768844]
10. Taneera J, Fadista J, Ahlqvist E, Zhang M, Wierup N, Renström E, et al. Expression profiling of cell cycle genes in human pancreatic islets with and without type 2 diabetes. Mol Cell Endocrinol 2013 Aug 15;375(1-2):35-42. [doi: 10.1016/j.mce.2013.05.003] [Medline: 23707792]
11. Kanatsuna N, Taneera J, Vaziri-Sani F, Wierup N, Larsson HE, Delli A, Lernmark. Autoimmunity against INS-IGF2 protein expressed in human pancreatic islets. J Biol Chem 2013 Oct 04;288(40):29013-29023 [FREE Full text] [doi: 10.1074/jbc.M113.478222] [Medline: 23935095]
12. Dominguez V, Raimondi C, Somanath S, Bugliani M, Loder MK, Edling CE, et al. Class II phosphoinositide 3-kinase regulates exocytosis of insulin granules in pancreatic beta cells. J Biol Chem 2011 Feb 11;286(6):4216-4225 [FREE Full text] [doi: 10.1074/jbc.M110.200295] [Medline: 21127054]
13. Marselli L, Thorne J, Dahiya S, Sgroi DC, Sharma A, Bonner-Weir S, et al. Gene expression profiles of beta-cell enriched tissue obtained by laser capture microdissection from subjects with type 2 diabetes. PLoS One 2010 Jul 13;5(7):e11499 [FREE Full text] [doi: 10.1371/journal.pone.0011499] [Medline: 20644627]
14. Greco S, Fasanaro P, Castelvecchio S, D'Alessandra Y, Arcelli D, Di Donato M, et al. MicroRNA dysregulation in diabetic ischemic heart failure patients. Diabetes 2012 Jun;61(6):1633-1641 [FREE Full text] [doi: 10.2337/db11-0952] [Medline: 22427379]
15. Misu H, Takamura T, Takayama H, Hayashi H, Matsuzawa-Nagata N, Kurita S, et al. A liver-derived secretory protein, selenoprotein P, causes insulin resistance. Cell Metab 2010 Nov 03;12(5):483-495 [FREE Full text] [doi: 10.1016/j.cmet.2010.09.015] [Medline: 21035759]
16. GEO2R. National Center for Biotechnology Information (NCBI). URL: https://www.ncbi.nlm.nih.gov/geo/geo2r/ [accessed 2022-02-04]
17. Pathan M, Keerthikumar S, Chisanga D, Alessandro R, Ang C, Askenase P, et al. A novel community driven software for functional enrichment analysis of extracellular vesicles data. J Extracell Vesicles 2017 Dec;6(1):1321455 [FREE Full text] [doi: 10.1080/20013078.2017.1321455] [Medline: 28717418]
18. Pathan M, Keerthikumar S, Ang C, Gangoda L, Quek CY, Williamson NA, et al. FunRich: An open access standalone functional enrichment and interaction network analysis tool. Proteomics 2015 Aug 17;15(15):2597-2601. [doi: 10.1002/pmic.201400515] [Medline: 25921073]
19. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015 Apr 20;43(7):e47 [FREE Full text] [doi: 10.1093/nar/gkv007] [Medline: 25605792]
20. Demsar J, Curk T, Gorup C, Hocevar T, Milutinovic M, Mozina M, et al. Orange: data mining toolbox in Python. J Machine Learn Res 2013 Aug;14:2349-2353 [FREE Full text]
21. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 2009 Dec 18;4(1):44-57. [doi: 10.1038/nprot.2008.211] [Medline: 19131956]
22. Huang D, Sherman B, Lempicki R. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 2009 Jan;37(1):1-13 [FREE Full text] [doi: 10.1093/nar/gkn923] [Medline: 19033363]

23. Shamsaei B, Chojnacki S, Pilarczyk M, Najafabadi M, Niu W, Chen C, et al. piNET: a versatile web platform for downstream analysis and visualization of proteomics data. Nucleic Acids Res 2020 Jul 02;48(W1):W85-W93 [FREE Full text] [doi: 10.1093/nar/gkaa436] [Medline: 32469073]
24. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 2019 Jan 08;47(D1):D607-D613 [FREE Full text] [doi: 10.1093/nar/gky1131] [Medline: 30476243]
25. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, et al. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res 2005 Jan 01;33(Database issue):D433-D437 [FREE Full text] [doi: 10.1093/nar/gki005] [Medline: 15608232]
26. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003 Nov;13(11):2498-2504 [FREE Full text] [doi: 10.1101/gr.1239303] [Medline: 14597658]
27. Otasek D, Morris JH, Bouças J, Pico AR, Demchak B. Cytoscape Automation: empowering workflow-based network analysis. Genome Biol 2019 Sep 02;20(1):185 [FREE Full text] [doi: 10.1186/s13059-019-1758-4] [Medline: 31477170]
28. Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 2003 Jan 13;4(1):2 [FREE Full text] [doi: 10.1186/1471-2105-4-2] [Medline: 12525261]
29. Chang L, Zhou G, Soufan O, Xia J. miRNet 2.0: network-based visual analytics for miRNA functional analysis and systems biology. Nucleic Acids Res 2020 Jul 02;48(W1):W244-W251 [FREE Full text] [doi: 10.1093/nar/gkaa467] [Medline: 32484539]
30. Calculate and draw custom Venn diagrams. Bioinformatics & Evolutionary Genomics. URL: http://bioinformatics.psb.ugent.be/webtools/Venn/ [accessed 2022-02-04]
31. Licursi V, Conte F, Fiscon G, Paci P. MIENTURNET: an interactive web tool for microRNA-target enrichment and network-based analysis. BMC Bioinformatics 2019 Nov 04;20(1):545 [FREE Full text] [doi: 10.1186/s12859-019-3105-x] [Medline: 31684860]
32. Nolan CJ, Damm P, Prentki M. Type 2 diabetes across generations: from pathophysiology to prevention and management. Lancet 2011 Jul 09;378(9786):169-181. [doi: 10.1016/S0140-6736(11)60614-4] [Medline: 21705072]
33. Hamed AE, Elsahar M, Elwan NM, El-Nakeep S, Naguib M, Soliman HH, et al. Managing diabetes and liver disease association. Arab J Gastroenterol 2018 Dec;19(4):166-179. [doi: 10.1016/j.ajg.2018.08.003] [Medline: 30420265]
34. Gupta MK, Vadde R. Identification and characterization of differentially expressed genes in type 2 diabetes using in silico approach. Comput Biol Chem 2019 Apr;79:24-35. [doi: 10.1016/j.compbiolchem.2019.01.010] [Medline: 30708140]
35. Ennequin G, Caillaud K, Chavanelle V, Teixeira A, Etienne M, Li X, et al. Neuregulin 1 treatment improves glucose tolerance in diabetic db/db mice, but not in healthy mice. Arch Physiol Biochem 2020 Oct 17;126(4):320-325. [doi: 10.1080/13813455.2018.1534243] [Medline: 30449185]
36. Hopf A, Andresen C, Kötter S, Isić M, Ulrich K, Sahin S, et al. Diabetes-induced cardiomyocyte passive stiffening is caused by impaired insulin-dependent titin modification and can be modulated by neuregulin-1. Circ Res 2018 Jul 20;123(3):342-355. [doi: 10.1161/circresaha.117.312166]
37. Pan P, Dobrowsky RT. Differential expression of neuregulin-1 isoforms and downregulation of erbin are associated with Erb B2 receptor activation in diabetic peripheral neuropathy. Acta Neuropathol Commun 2013 Jul 17;1(1):39 [FREE Full text] [doi: 10.1186/2051-5960-1-39] [Medline: 24252174]
38. Muller Y, Piaggi P, Hanson R, Kobes S, Bhutta S, Abdussamad M, et al. A cis-eQTL in PFKFB2 is associated with diabetic nephropathy, adiposity and insulin secretion in American Indians. Hum Mol Genet 2015 May 15;24(10):2985-2996 [FREE Full text] [doi: 10.1093/hmg/ddv040] [Medline: 25662186]
39. Arden C, Hampson L, Huang G, Shaw J, Aldibbiat A, Holliman G, et al. A role for PFK-2/FBPase-2, as distinct from fructose 2,6-bisphosphate, in regulation of insulin secretion in pancreatic beta-cells. Biochem J 2008 Apr 01;411(1):41-51. [doi: 10.1042/BJ20070962] [Medline: 18039179]
40. Bockus LB, Matsuzaki S, Vadvalkar SS, Young ZT, Giorgione JR, Newhardt MF, et al. Cardiac insulin signaling regulates glycolysis through phosphofructokinase 2 content and activity. J Am Heart Assoc 2017 Dec 04;6(12):3846 [FREE Full text] [doi: 10.1161/JAHA.117.007159] [Medline: 29203581]
41. Gao J, Feng W, Lv W, Liu W, Fu C. HIF-1/AKT signaling-activated PFKFB2 alleviates cardiac dysfunction and cardiomyocyte apoptosis in response to hypoxia. Int Heart J 2021 Mar 30;62(2):350-358. [doi: 10.1536/ihj.20-315] [Medline: 33678793]
42. Marselli L, Thorne J, Dahiya S, Sgroi DC, Sharma A, Bonner-Weir S, et al. Gene expression profiles of Beta-cell enriched tissue obtained by laser capture microdissection from subjects with type 2 diabetes. PLoS One 2010 Jul 13;5(7):e11499 [FREE Full text] [doi: 10.1371/journal.pone.0011499] [Medline: 20644627]
43. Lee HJ, Yun JH, Kim H, Jang HB, Park SI, Lee H. 244-LB: Glutamate is associated with type 2 diabetes through PLG regulation. Diabetes 2019 Jun 04;68(Supplement 1):244-LB. [doi: 10.2337/db19-244-lb]
44. Saini C, Petrenko V, Pulimeno P, Giovannoni L, Berney T, Hebrok M, et al. A functional circadian clock is required for proper insulin secretion by human pancreatic islet cells. Diabetes Obes Metab 2016 Apr 22;18(4):355-365. [doi: 10.1111/dom.12616] [Medline: 26662378]

45. Lamri A, Pigeyre M, Garver W, Meyre D. The extending spectrum of NPC1-related human disorders: from Niemann-Pick C1 disease to obesity. Endocr Rev 2018 Apr 01;39(2):192-220 [FREE Full text] [doi: 10.1210/er.2017-00176] [Medline: 29325023]
46. Skrzypski M, Billert M, Nowak KW, Strowski MZ. The role of orexin in controlling the activity of the adipo-pancreatic axis. J Endocrinol 2018 Aug;238(2):R95-R108. [doi: 10.1530/JOE-18-0122] [Medline: 29848609]
47. Sellayah D, Sikder D. Orexin receptor-1 mediates brown fat developmental differentiation. Adipocyte 2012 Jan 01;1(1):58-63 [FREE Full text] [doi: 10.4161/adip.18965] [Medline: 23700511]
48. Yang Y, Guo F, Peng Y, Chen R, Zhou W, Wang H, et al. Transcriptomic profiling of human placenta in gestational diabetes mellitus at the single-cell level. Front Endocrinol 2021 May 7;12:679582. [doi: 10.3389/fendo.2021.679582] [Medline: 34025588]
49. Hou Y, Xie Y, Yang S, Han B, Shi L, Bai X, et al. EEF1D facilitates milk lipid synthesis by regulation of PI3K-Akt signaling in mammals. FASEB J 2021 May;35(5):e21455. [doi: 10.1096/fj.202000682RR] [Medline: 33913197]
50. Luo J, Dou L, Yang Z, Zhou Z, Huang H. CBFA2T2 promotes adipogenic differentiation of mesenchymal stem cells by regulating CEBPA. Biochem Biophys Res Commun 2020 Aug 20;529(2):133-139. [doi: 10.1016/j.bbrc.2020.05.120] [Medline: 32703401]
51. Khokhar M, Roy D, Bajpai NK, Bohra GK, Yadav D, Sharma P, et al. Metformin mediates MicroRNA-21 regulated circulating matrix metalloproteinase-9 in diabetic nephropathy: an in-silico and clinical study. Arch Physiol Biochem 2021 Jun 04:1-11. [doi: 10.1080/13813455.2021.1922457] [Medline: 34087084]
52. Pan G, Liu Q, Xin H, Liu J. The key regulation of miR-124-3p during reprogramming of primary mouse hepatocytes into insulin-producing cells. Biochem Biophys Res Commun 2020 Feb 05;522(2):315-321. [doi: 10.1016/j.bbrc.2019.11.058] [Medline: 31761319]
53. Liu Y, Ke X, Guo W, Wang X, Peng C, Liao Z, et al. Circ-RHOJ.1 regulated myocardial cell proliferationand apoptosis via targeting the miR-124-3p/NRG-1 axisin myocardial ischemia/reperfusion injury. Arch Med Sci 2019 Aug 7:1-14. [doi: 10.5114/aoms.2019.87205]
54. Duan J, Liu H, Chen J, Li X, Li P, Zhang R. Changes in gene expression of adipose tissue CD14 cells in patients with type 2 diabetes mellitus and their relationship with environmental factors. Zhong Nan Da Xue Xue Bao Yi Xue Ban 2021 Jan 28;46(1):1-10 [FREE Full text] [doi: 10.11817/j.issn.1672-7347.2021.190558] [Medline: 33678630]
55. Wang Y, Zheng Z, Jia Y, Yang Y, Xue Y. Role of p53/miR-155-5p/sirt1 loop in renal tubular injury of diabetic kidney disease. J Transl Med 2018 May 30;16(1):146 [FREE Full text] [doi: 10.1186/s12967-018-1486-7] [Medline: 29848325]
56. Bai X, Luo Q, Tan K, Guo L. Diagnostic value of VDBP and miR-155-5p in diabetic nephropathy and the correlation with urinary microalbumin. Exp Ther Med 2020 Nov 11;20(5):86 [FREE Full text] [doi: 10.3892/etm.2020.9214] [Medline: 32968443]
57. Wang G, Wu B, Zhang B, Wang K, Wang H. LncRNA CTBP1-AS2 alleviates high glucose-induced oxidative stress, ECM accumulation, and inflammation in diabetic nephropathy via miR-155-5p/FOXO1 axis. Biochem Biophys Res Commun 2020 Nov 05;532(2):308-314. [doi: 10.1016/j.bbrc.2020.08.073] [Medline: 32868076]
58. Guo L, Tan K, Luo Q, Bai X. Dihydromyricetin promotes autophagy and attenuates renal interstitial fibrosis by regulating miR-155-5p/PTEN signaling in diabetic nephropathy. Bosn J Basic Med Sci 2020 Aug 03;20(3):372-380. [doi: 10.17305/bjbms.2019.4410] [Medline: 31668144]
59. Khokhar M, Roy D, Modi A, Agarwal R, Yadav D, Purohit P, et al. Perspectives on the role of PTEN in diabetic nephropathy: an update. Crit Rev Clin Lab Sci 2020 Nov 20;57(7):470-483. [doi: 10.1080/10408363.2020.1746735] [Medline: 32306805]
60. Zhou Y, Ma X, Han J, Yang M, Lv C, Shao Y, et al. Metformin regulates inflammation and fibrosis in diabetic kidney disease through TNC/TLR4/NF-κB/miR-155-5p inflammatory loop. World J Diabetes 2021 Jan 15;12(1):19-46 [FREE Full text] [doi: 10.4239/wjd.v12.i1.19] [Medline: 33520106]
61. Khokhar M, Purohit P, Roy D, Tomo S, Gadwal A, Modi A, et al. Acute kidney injury in COVID 19 - an update on pathophysiology and management modalities. Arch Physiol Biochem 2020 Dec 15:1-14. [doi: 10.1080/13813455.2020.1856141] [Medline: 33320717]
62. Khokhar M, Tomo S, Purohit P. MicroRNAs based regulation of cytokine regulating immune expressed genes and their transcription factors in COVID-19. Meta Gene 2022 Feb;31:100990 [FREE Full text] [doi: 10.1016/j.mgene.2021.100990] [Medline: 34722158]
63. Agarwal RG, Khokhar M, Purohit P, Modi A, Bajpai NK, Bohra GK, et al. A clinical and in-silico study of MicroRNA-21 and growth differentiation factor-15 expression in pre-diabetes, type 2 diabetes and diabetic nephropathy. Minerva Endocrinol 2022 Feb 01:online ahead of print. [doi: 10.23736/S2724-6507.22.03646-6] [Medline: 35103454]

Abbreviations

AGE: advanced glycation end products
AKT: protein kinase B
BP: biological process
CC: cellular component

CDK5RAP: CDK5 regulatory subunit associated protein
cGMP: cyclic guanosine monophosphate
DAVID: Database for Annotation, Visualization and Integrated Discovery
DEG: differentially expressed gene
DKD: diabetic kidney disease
EGF: epidermal growth factor
ERBB1: epidermal growth factor receptor
GEO: Gene Expression Omnibus
GO: Gene Ontology
GWAS: genome-wide association study
IGF1: insulin-like growth factor 1
KEGG: Kyoto Encyclopedia of Genes and Genomes
limma: linear models for microarray data
MAPK: mitogen-activated protein kinase
MCODE: Molecular Complex Detection
MF: molecular function
MIENTURNET: MicroRNA Enrichment Turned Network
miRNA: microRNA
mTOR: mammalian target of rapamycin
NPC1: Niemann-Pick type C1
NRG1: neuregulin 1
PFK2: 6-phosphofructo-2-kinase/fructose 2,6-bisphosphatase isoform 2
PI3K: phosphoinositide 3-kinase
PPI: protein-protein interaction
RAGE: receptor of advanced glycation end products
STRING: Search Tool for the Retrieval of Interacting Genes/Proteins
T2DM: type 2 diabetes mellitus
TGF: transforming growth factor

Edited by A Mavragani; submitted 28.07.21; peer-reviewed by BS Chrisman, M Giri, M Hetti Arachchilage; comments to author 23.09.21; revised version received 18.11.21; accepted 27.12.21; published 23.02.22.

Please cite as:
Khokhar M, Roy D, Tomo S, Gadwal A, Sharma P, Purohit P
Novel Molecular Networks and Regulatory MicroRNAs in Type 2 Diabetes Mellitus: Multiomics Integration and Interactomics Study
JMIR Bioinform Biotech 2022;3(1):e32437
URL: https://bioinform.jmir.org/2022/1/e32437
doi:10.2196/32437
PMID:

Original Paper

An Analysis of Different Distance-Linkage Methods for Clustering Gene Expression Data and Observing Pleiotropy: Empirical Study

Joydhriti Choudhury[1*]; Faisal Bin Ashraf[1*]

Brac University, Dhaka, Bangladesh

[*]all authors contributed equally

Corresponding Author:
Faisal Bin Ashraf
Brac University
Mohakhali
Dhaka
Bangladesh
Phone: 880 9617445125
Email: faisal.ashraf@bracu.ac.bd

Abstract

Background: Large amounts of biological data have been generated over the last few decades, encouraging scientists to look for connections between genes that cause various diseases. Clustering illustrates such a relationship between numerous species and genes. Finding an appropriate distance-linkage metric to construct clusters from diverse biological data sets has thus become critical. Pleiotropy is also important for a gene's expression to vary and create varied consequences in living things. Finding the pleiotropy of genes responsible for various diseases has become a major research challenge.

Objective: Our goal was to establish the optimal distance-linkage strategy for creating reliable clusters from diverse data sets and identifying the common genes that cause various tumors to observe genes with pleiotropic effect.

Methods: We considered 4 linking methods—single, complete, average, and ward—and 3 distance metrics—Euclidean, maximum, and Manhattan distance. For assessing the quality of different sets of clusters, we used a fitness function that combines silhouette width and within-cluster distance.

Results: According to our findings, the maximum distance measure produces the highest-quality clusters. Moreover, for medium data set, the average linkage method, and for large data set, the ward linkage method works best. The outcome is not improved by using ensemble clustering. We also discovered genes that cause 3 different cancers and used gene enrichment to confirm our findings.

Conclusions: Accuracy is crucial in clustering, and we investigated the accuracy of numerous clustering techniques in our research. Other studies may aid related works if the data set is similar to ours.

(JMIR Bioinform Biotech 2022;3(1):e30890) doi:10.2196/30890

KEYWORDS

gene clustering; gene expression; distance metric; linkage method; hierarchical clustering; pleiotropy

Introduction

A substantial amount of genetic data began to accumulate in the hands of bioinformatics experts at the turn of the 21st century. The process was sped by advances in technology hardware and improved computer algorithms. Scientists began storing all of this genomic information in sequential data [1] and intensity matrix [2] formats. Different types of sequences, such as protein, DNA, and RNA sequences, are kept in sequential data format, and the intensity matrix preserves gene behavior under various conditions. To record and analyze gene behavior on sample individuals, these conditions can vary under varied light intensities.

Microarray [3] is a type of intensity matrix in which each row represents a single gene, and each column indicates that gene's behavior in a given situation. A microarray data set's sample structure is shown in Table 1. Four genes express themselves at 3 different times or circumstances. Depending on the normalization approach used, the values stored in a microarray data set can be both positive and negative.

Table 1. Sample microarray data.

Genes	Time 1	Time 2	Time 3
Gene 1	0.25	0.22	0.65
Gene 2	–0.75	1.25	–0.63
Gene 3	0.05	0.66	0.75
Gene 4	1.25	–0.52	0.15

Researchers have been extracting valuable biological information from microarray data. The construction of a phylogenetic tree is one of the most extensively used methodologies [4]. The evolutionary relationships between numerous species are shown by the phylogenetic tree. In the case of genes, it calculates gene similarity to create a gene tree that depicts how particular genes have evolved [5]. Although phylogenetic trees are based on sequence data because mutations occur in any species' genome sequence, genome sequences are comparatively large and need a lot of computing power and memory. Gene expression represents phenotypes of a gene, and different genes exhibit variable levels of expression under the same conditions [6]. As a result, we can employ phenotype, which is a measurement of the genes' reflection due to genotype differences. The expression level of genes calculates how near they are to one another using the microarray data set as an input, because the transcriptional activity of similar genes should be similar [7]. A tree is built by connecting all closely related genes one by one, with each leaf representing a single gene and branches separating one group of genes from another [8,9]. This hierarchical tree can aid in the creation of more precise groupings. It assists biologists in determining and comprehending the function of an unknown gene. As a result, developing appropriate metrics for clustering microarray data is a significant scientific challenge.

Different clustering approaches have been presented to extract information from the microarray data set [10]. Clustering algorithms divide unclassified data into distinct classified groups [11], with the most comparable data points grouped together. As a result, if an unknown element belongs to a recognized cluster, it becomes easier for the researcher to forecast its properties. Clustering is a technique used in bioinformatics to organize microarray data and predict properties of unknown genes based on which cluster they belong to [11]. Furthermore, bioinformatics workflow [12] and immune repertoire profiling [13] are classified using hierarchical clustering, a sort of clustering technique. It also has applications in the prediction of nonsmall cell lung cancer metastasis [14], the high-confidence identification of B cell clones [15], and the identification of cell type from a single cell transcriptome [16]. It is also used to create a phylogenetic tree using microarray data [15]. The hierarchical clustering methodology uses a distance algorithm to calculate the distance between distinct genes after inputting microarray data. The distance is then used to connect closely related genes in clusters using a linkage approach.

Various distance methods are employed depending on the data set's characteristics. The way the 2 distance methods determine the difference between 2 distant data points is the fundamental distinction between them. Euclidean [17], Chebyshev [18], and other distance approaches are common. After applying the distance approach, the hierarchical clustering technique connects related genes using several types of linking methods to form a cluster. single linkage method [19], complete linkage method [20], average linkage method [20], and others are some of the most used linkage methods. Linkage methods connect genes in a bottom-up manner, eventually resulting in a hierarchical tree, often known as a phylogenetic tree. As computational ability and technology progress, it has become increasingly important to establish reliable clusters of related genes to understand unknown genes in sensitive domains such as health care and disease prediction.

Pleiotropy is another key phenomenon identified in the investigation of gene functions behind many diseases. Pleiotropy occurs when a single gene influences many phenotypic features [21]. There are numerous examples of multiple genes working together to cause a single disease [22-24]. Furthermore, it appears that a single gene is responsible for several disorders [25]. Even though we can identify diseases caused by the same gene, the gene's impact on each disease is different. It may appear to be more active in some disorders than in others. As a result, we can visualize the impact of a gene on other diseases if we can detect commonalities in their expressions for different diseases and quantify the distance.

In this work, we used a variety of data sets to investigate different distance-linkage combinations for hierarchical clustering. These clusters have revealed which gene groupings are closely connected to one another. We also assessed the fitness of those groupings and attempted to determine which distance-linkage combination produced the greatest results. We validated our findings using 8 different data sets. Furthermore, we used the best measure to identify common genes responsible for various tumors. Gene enrichment scores about their influence on various diseases were used to corroborate our findings.

Methods

This section goes over our proposed methodology. First, we provided the proposed workflow for determining the optimum clustering distance-linkage approach. Then we went over several distance metrics, linkage methods, and our selection procedure for comparing the performance of various combinations. Finally, the pleiotropic gene observation methodology is discussed.

Identifying the Best Distance-Linkage Method

Our investigation begins with the import of a microarray data set into our procedure. This microarray data set is typically a 2D array, with rows representing different genes and columns representing their intensity at various time stamps. To minimize the dimensionality of the data set, we will use Principal

Component Analysis. It is a sophisticated approach used by academics to remove irrelevant data from a data set while keeping its integrity.

Then, in our data set, we run a distance metric. A distance measure, in general, calculates the similarity of 2 genes and determines how far apart they are. We employed the following 3 different distance metrics: Euclidean, Manhattan, and maximum. We chose a linkage method to connect related genes and generate a hierarchical tree after picking the distance metric. We used the following 4 linkage methods: single, complete, average, and ward linkage methods. We constructed a hierarchical tree using the distance-linkage method, where each leaf represents a gene, and the branches reflect the dissimilarity among them. The tree was then cut to various heights, resulting in several sets of genes for each cut point. Subsequently, we identified the appropriate cut point for that hierarchical tree by calculating how well those genes are clustered on different cut points. We used "Average Silhouette Width" and "Distance within Cluster" to calculate the fitness of the groups formed by different cut locations. The optimal fitness value is calculated using these fitness values. We determined the best combination of distance and linkage methods for a single data set by repeating this process with different combinations of distance and linkage methods. Figure 1 depicts the algorithm.

Figure 1. Proposed algorithm for finding the best distance-linkage combination. Input: Microarray data set. Output: Distance-linkage combination.

```
D <- List of distance metrics
L <- List of linkage metrics
Best <- { }
score <- 0

for each d in D:
  for each l in L:
    create hierarchical tree using d and l from the data set
    fitness <- 0
    repeat
      f <- cut the tree at different heights and calculate the fitness of cluster
      if f > fitness then
        fitness <- f
      if fitness > score then
        score <- fitness
        Best <- (d,l)
    until fitness – 0 or fitness < f

return Best
```

For a particular data set, D, optimal fitness value can be expressed by the following equations:

Where d  distance methods and, I  linkage methods.

Used Distance Methods

Euclidean distance uses Pythagorean formula to calculate the distance between 2 genes. For n dimensional space, we can write that formula as follows:

Unlike Euclidean distance, Manhattan distance takes the modulus value of the subtraction. For n-dimensional space, the equation of Manhattan Distance will be as follows:

Maximum distance, on the other hand, calculates the subtraction value for each column before selecting the highest number. The formula for n-dimensional space is as follows:

Used Linkage Methods

The single linkage approach connects 2 clusters by taking the shortest distance between them. The equation for the single linkage method to calculate the distance between any element and another element in another group is as follows:

Where p is an element in cluster P and q is an element of cluster Q.

To compute the distance, the complete technique uses the farthest points in 2 clusters and connects the clusters with the shortest distance. The equation for the entire linking approach is as follows:

The average method determines the average value for each gene inside the cluster, then connects them one by one on each layer to form a hierarchical tree. Equation 7 is the average linkage method update formula.

Where m is all the instances of cluster a, and n is all the instances of cluster b.

A centroid point is determined using Ward linkage (much like the centroid method). The squared distance value of each point in each cluster is then calculated using that centroid. It then sums all the squared distance values obtained by the 2 clusters together. It takes the smallest total value produced by a cluster pair and merges them on that level after repeating the same technique for every cluster on the same level. Equation 8 is the Ward linkage method update formula.

Metrics Used to Calculate Fitness

The fitness of the clusters we acquired after cutting the hierarchical tree at a specific height was calculated using the following 2 metrics: average silhouette width (ASW) and distance inside cluster. The following formula is used to compute silhouette width:

Where a(i) is the average distance from object i and all the other points of the cluster in which i belongs; b(i) is the distance of the closest point in other cluster; and s(i) is the silhouette value between 2 clusters.

ASW is the average of all the silhouette values. Generally, it varies from –1 to 1, and the value closer to 1 is considered better.

The distance within a cluster is used to determine how close the elements are. Each cluster's centroid is chosen during this process. The distance between each object in the cluster and the centroid is then determined as an average. This calculation's formula is as follows:

Where dist(c,i) is the distance between centroid c and element i in a cluster; E is the set of elements in the cluster; and |E| is the number of elements in the cluster.

From the characteristics, we can understand that ASW measures the quality of clusters. A greater ASW indicates good quality of clusters, that is, for a data set D, distance metric d and linkage method l,

Where S_i is the ASW for cut point i.

However, distance within clusters measures how compact the data points are in the clusters. Therefore, better-quality clusters will have lower distance within clusters, that is,

Where W_i is the distance within clusters for cut point i.

Thus, to compare the quality of clusters we acquired at different cut points i in the hierarchical tree, our fitness function combines these 2 criteria. When these 2 relationships are combined, our fitness function becomes as follows:

From this function, we can find out the optimal fitness for a specific combination of metrics in a certain data set.

Cluster Ensemble

We will try ensemble clustering [26] to see if it works better once we have tried different clustering combinations. Three ensemble clustering techniques were employed, which are as follows: (1) similarity partitioning based on clusters; (2) hypergraph partitioning algorithm [27-29]; (3) meta-clustering algorithm.

Cluster-Based Similarity Partitioning

It starts by creating an n×n binary matrix in which the input is 1 if two objects belong to the same cluster and 0 otherwise. Every clustering approach is put through it. The final ensemble cluster is then generated using an entry-wise average of all clustering approaches.

Hypergraph Partitioning Algorithm

The data set is represented as a hypergraph by this algorithm. The hypergraph is then partitioned to determine the smallest number of edges. It produces the ensemble cluster based on the smallest number of edges.

Metaclustering Algorithm

The metaclustering algorithm starts by creating numerous clusters from a data set. The dissimilarity between those clusters is then calculated, and a metacluster is generated as a result of that measurement. In this approach, the ensemble is represented by the final metacluster.

One of the most important characteristics of these algorithms is that the number of clusters that the algorithm will build must be declared at the start. For the specified data set, we used the cluster number created by the best distance-linkage combination.

Observing Pleiotropy for Different Cancers

We identified the genes responsible for various cancer tumors from the data sets and then evaluated their expression in different patients with cancer to report their various phenotypes in order to discover the pleiotropic behavior of distinct genes. We built a secondary data set by extracting the expression data for each gene from each data set after identifying the common genes across these disorders. Every primary data set must contain an equal number of time stamp values in order to build a 2D microarray data set. The data sets, however, have different numbers of columns. Central nervous system, for example, includes 60 time stamps for a single gene, but the ALL-AML (acute lymphoblastic leukemia-acute myeloid leukemia) data set has 72 time stamps. We cannot modify or remove any columns from the data set because doing so could compromise the data's integrity or result in the loss of valuable information. To address this issue, we estimated the mean, median, standard

deviation, and variance, which may be used to summarize numerical data [30], and we used these numbers to construct our secondary data set. We will design a hierarchical tree using the perfect distance-linkage method found in the previously presented method because we have a data set for each gene with pleiotropic behavior. For that particular gene, the diseases that are closest to each other share similar summarized statistics. As a result, these trees will aid our understanding of how a single gene exhibits various phenotypes in patients with cancer. Furthermore, the gene enrichment scores of these common genes for the disorders that are frequent will be used to corroborate our findings.

Ethical Considerations

Since no human or animal trial was conducted during this research, the authors did not apply for an ethical approval for the study.

Results

We will discuss the experimental outcomes we discovered in our research in this part. We started by explaining the data sets we used. The findings for various distance-linkage method combinations were then shown. We later presented our findings in terms of pleiotropy for the shared genes.

Data Set

We obtained gene expression data for various cancers from a publicly accessible database [31]. Every data set includes the disease-causing genes as well as their expression in various patients with the same condition. We also examined a data set from a variety of disorders to confirm that our findings were disease-agnostic. We used 7 data sets for various cancers. Table 2 lists the specifics of each data set.

The number of genes and patients, or the number of conditions for each gene, differs among these data sets. We used a diverse data set to discover the ideal metrics, which can be used to any gene expression data set. Furthermore, these databases contain certain genes that are widely used. We have created a secondary data set to explore and analyze those genes further.

Table 2. Description of data sets.

Data set	Data domain	Number of patients	Number of genes
CNS[a]	Central nervous system	60	7129
ALL-AML[b]	Acute lymphocytic leukemia	72	7129
Lung cancer	Lung cancer	181	12,533
Ovarian cancer	Ovarian cancer	253	15,154
Lymphoma	Lymphoma	62	4022
SRBCT[c]	Small round blue cell tumor	83	2308

[a]CNS: central nervous system.

[b]ALL-AML: acute lymphoblastic leukemia-acute myeloid leukemia.

[c]SRBCT: small round blue cell tumor.

Result of Experiments for Identifying the Best Distance-Linkage Method

In our experiment, we employed several combinations of distance measurements and connection algorithms to generate a hierarchical tree. To validate our founding, we used 3 distance metrics and 4 linking methods. We combined these 3 distance metrics and 4 linkage methods to build 12 hierarchical trees for each data set. We cut each tree on numerous cut points after building hierarchical trees. As a result, the tree has been separated into several distinct groups. We assessed the fitness value for each cut point and selected the highest as the ideal value for that hierarchical tree given that particular distance metric-linkage method combination.

A portion of a hierarchical tree of genes from the lung cancer data set is shown in Figure 2. This tree was constructed using the maximum-Ward combination. The full tree has a large number of leaves due to the data set's 12,533 genes. All the values using Equation 13 are calculated, and the best values for each combination of distance method and linkage metric are shown in Table 3.

Figure 2. Hierarchical tree created using the maximum-Ward method on lung cancer data set.

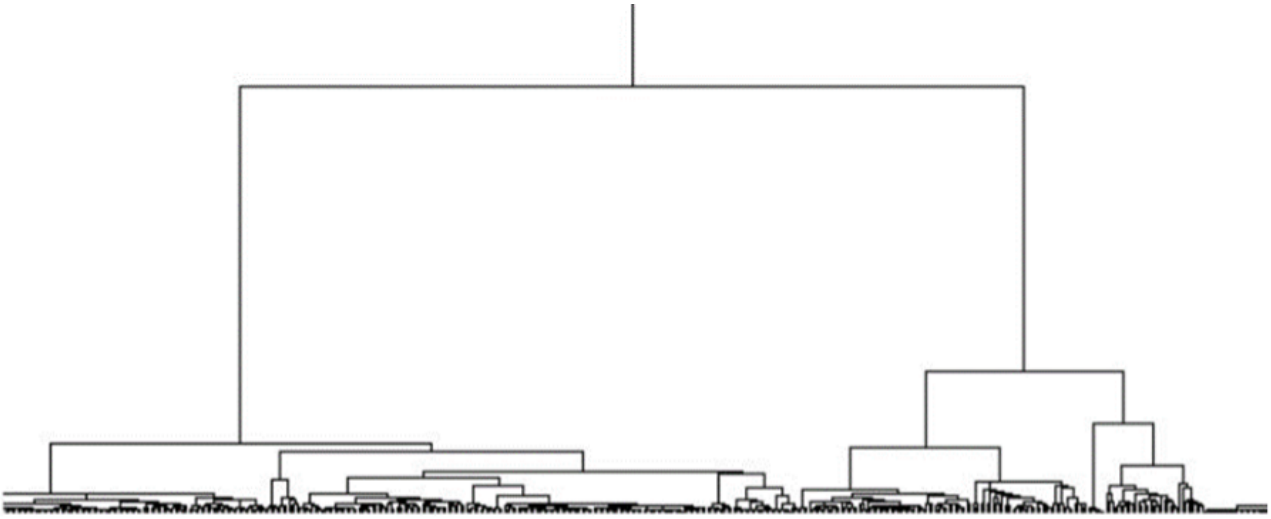


Table 3. Fitness value for different combinations of distance and linkage metrics.

Data set and linkage	Manhattan distance	Euclidean distance	Maximum distance
CNS[a]			
Average	3.80×10^{-13}	9.47×10^{-12}	3.50×10^{-11}
Complete	1.42×10^{-13}	6.78×10^{-12}	3.44×10^{-12}
Single	2.59×10^{-13}	5.72×10^{-12}	2.16×10^{-11}
Ward	4.49×10^{-14}	3.22×10^{-13}	3.09×10^{-12}
ALL-AML[b]			
Average	1.20×10^{-6}	1.45×10^{-5}	3.39×10^{-5}
Complete	8.89×10^{-7}	2.11×10^{-5}	1.51×10^{-5}
Single	1.11×10^{-6}	1.37×10^{-5}	1.24×10^{-5}
Ward	4.41×10^{-7}	2.64×10^{-6}	3.07×10^{-5}
Lung cancer			
Average	5.56×10^{-8}	1.48×10^{-6}	3.36×10^{-6}
Complete	5.35×10^{-8}	1.23×10^{-6}	1.52×10^{-9}
Single	5.33×10^{-8}	6.47×10^{-7}	5.86×10^{-7}
Ward	3.03×10^{-8}	1.19×10^{-6}	6.71×10^{-6}
Ovarian			
Average	1.25×10^{-5}	1.59×10^{-4}	2.87×10^{-4}
Complete	1.71×10^{-5}	7.49×10^{-5}	6.28×10^{-4}
Single	2.88×10^{-6}	3.12×10^{-4}	1.28×10^{-4}
Ward	2.49×10^{-4}	3.44×10^{-5}	9.31×10^{-4}
Lymphoma			
Average	1.29×10^{-7}	2.81×10^{-6}	9.66×10^{-6}
Complete	2.21×10^{-8}	2.34×10^{-6}	6.00×10^{-6}
Single	1.01×10^{-7}	2.81×10^{-6}	8.10×10^{-6}
Ward	1.23×10^{-8}	6.05×10^{-7}	2.82×10^{-6}
SRBCT[c]			
Average	1.52×10^{-7}	6.73×10^{-6}	4.41×10^{-5}
Complete	1.03×10^{-7}	4.72×10^{-6}	3.73×10^{-5}
Single	8.24×10^{-8}	4.34×10^{-6}	3.00×10^{-5}
Ward	3.88×10^{-9}	8.55×10^{-8}	2.67×10^{-6}

[a]CNS: Central Nervous System.

[b]ALL-AML: acute lymphoblastic leukemia-acute myeloid leukemia.

[c]SRBCT: small round blue cell tumor.

Ensemble Result

We chose the data set (ALL-AML) for testing and ran these 4 ensemble clustering techniques. For this data set, the maximum-average combination produced the best result, with a cluster number of 135. Table 4 displays the fitness values. We discovered that no ensemble clustering approach improves fitness value in any way.

Table 4. Fitness value for different ensemble techniques.

Ensemble techniques	Fitness value
CSPA[a]	4.32×10^{-6}
HGPA[b]	3.29×10^{-6}
MCLA[c]	1.53×10^{-5}
Maximum-average	3.39×10^{-5}

[a]CSPA: cluster-based similarity partitioning.

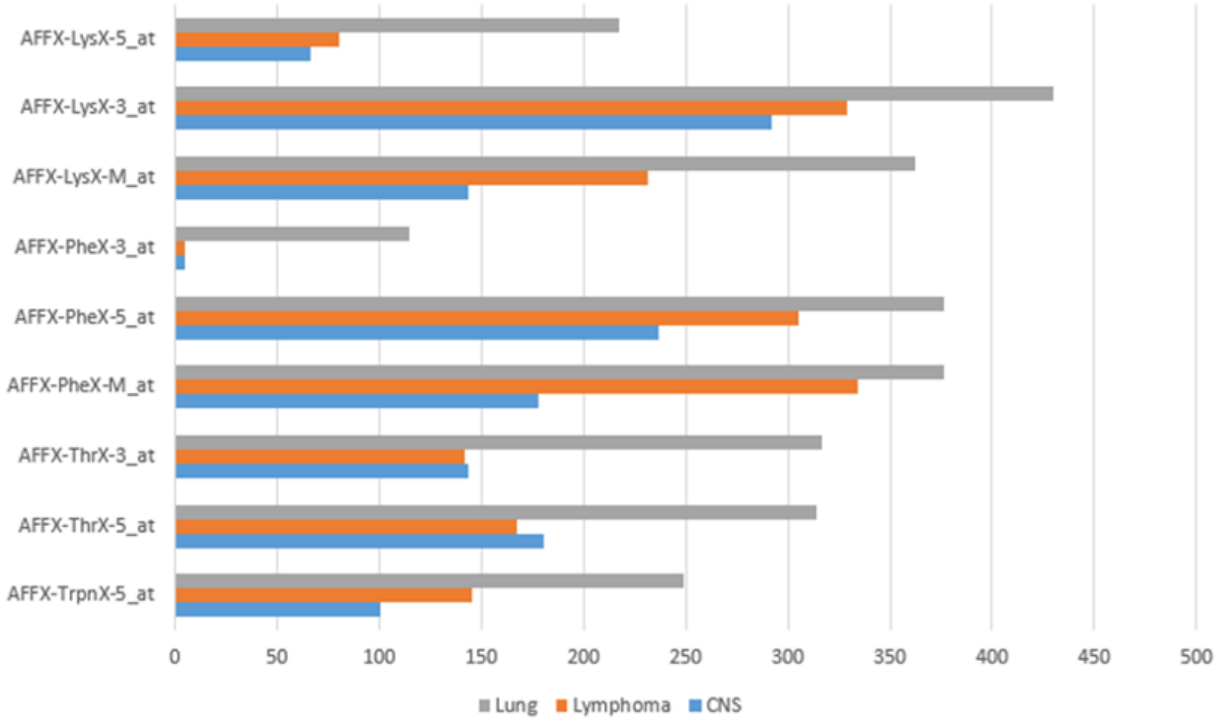
[b]HGPA: hyper graph partitioning algorithm.

[c]MCLA: metaclustering algorithm.

Result Analysis for Common Genes

Multiple tumors can be caused by a small number of genes. We discovered 9 genes linked to the following 3 types of cancer: central nervous system, lymphoma, and lung cancer. AFFX-TrpnX-5 at, AFFX-ThrX-5 at, AFFX-ThrX-3 at, AFFX-PheX-M at, AFFX-PheX-5 at, AFFX-PheX-3 at, AFFX-LysX-M at, AFFX-LysX-3 at, and AFFX-LysX-5 at were discovered to be common genes. We found the gene enrichment score publicly available at [32] to confirm our findings. Gene enrichment scores in various malignancies are given in Figure 3 for the discovered common genes.

Figure 3. Gene enrichment score vs cancer type.



Discussion

Principal Findings

The maximum distance method combined with the average linkage method produces better hierarchical trees in 4 data sets (central nervous system, leukemia, lymphoma, and SRBCT), according to the fitness values provided in Table 3. These data sets are medium in size, with 60-80 rows and 2000-7000 columns, as shown in Table 2. In the Spellmen data set, however, the maximum-average combination also excels. The other 4 data sets reflect human genes that are responsible for specific tumors, whereas Spellmen is a microarray data set of bacteria. However, the maximum distance approach with ward linkage method constructs a superior hierarchical tree compared with the other methods in 2 of the largest data sets, lung and ovarian. These 2 data sets are larger than the others, and they share no genes with the others.

The maximum distance metric outperforms the other 2 distance methods among the 3 most commonly used distance metrics. Maximum distance considers only 1 column where those 2 genes have the most variance when calculating distance between them. The Euclidean and Manhattan distance methods, on the other hand, would have taken distances across all columns. As a result, the dissimilarity values for the Euclidean and Manhattan distances are approaching the maximum distance. As a result, in clustering, the Euclidean and Manhattan distances place

points slightly farther apart than the Maximum distance. Furthermore, because all the columns indicate the same features of a gene evaluated at different time stamps, we can analyze the worst scenario (ie, the greatest differential in the expression of 2 genes at a certain moment). This is the most significant difference between these 2 genes. To put it another way, maximum distance calculates only the difference that matters. The Euclidean and Manhattan distances, on the other hand, are becoming buried in the massive amount of data. The maximum distance, on the other hand, may create undesirable clusters in a different data set with uniform variation across all columns.

When the data set is small, the average linkage approach performs well, and when the data set is huge, the ward method performs well. The single linkage approach may be faster than the average method for joining clusters, but it is not necessarily better. When determining the proximity of 2 clusters, it always considers only 2 points and ignores all others. The average linkage approach, on the other hand, considers all the points in the cluster when determining relatedness. When using the ward technique, the sum square error is used to determine similarity. When working with small or medium-sized data sets, the average linkage approach outperforms the ward linkage method, but as the data sets grow larger, the sum square error values take over and produce superior results compared with the average linkage method.

We tried to identify the optimal combination in our research and found that the maximum distance method performs better on hierarchical clustering when column variance is not uniform across the data set. However, if the data set is medium in size, with around 2000-7000 rows and 60-80 columns, the average linkage technique will outperform other linkage methods, and if the data set is very large, with 12,000-15,000 rows and 100-200 columns, the ward linkage approach will outperform other linkage methods. Furthermore, it has been discovered that ensemble clustering can improve performance by a very little amount at the cost of extra work.

We discovered 9 common genes that cause the following 3 diseases: lymphoma, central nervous system cancer, and lung cancer. We tried to figure out how these genes play a role in these 3 diseases using the data provided in the data sets. The maximum-average hierarchical clustering technique was chosen since it performed the best in the first experiment. We used gene enrichment score to confirm our findings on whether the 9 genes discovered have an impact on these 3 conditions. Figure 3 shows the gene enrichment scores for these genes. We can see that 8 of the 9 genes are important for all 3 cancers. Only 1 gene (AFFX-PheX-3 at) is more important than the other 2 in lung cancer. However, it is clear that our discovered genes have a significant impact on these 3 cancer forms.

Bioinformatics is becoming more and more involved in health sectors, such as disease detection and individualized medicine recommendation, as computational technology advances. Clustering techniques are becoming increasingly important in these industries. We investigated several distance-linkage combinations and attempted to find a solution. We hope that other researchers who use hierarchical clustering will profit from our findings and apply what they have learned to their own study. We also discovered common genes with multiple symptoms, which we confirmed using gene enrichment profiling. Knowing the pleiotropic nature of these genes will help scientists work on them to combat cancer.

Conclusion

In this study, we discovered a set of measures that will yield higher-quality clusters for gene expression data. Pleiotropic behavior of common genes for many disorders was also discovered. To validate our findings, we used a variety of data sets that varied in size and richness. We used a fitness function to compare cluster quality between sets of clusters while assessing cluster quality. For medium-sized data sets, we discovered that the maximum distance metric combined with average linkage works best. Ward linkage also works better with huge data sets. Furthermore, due to data dimension differences, we had to preprocess data while identifying common genes for various disorders. It is critical to identify genes with similar symptoms more precisely and to separate those genes more effectively. Furthermore, detecting a gene by applying the clustering technique to find comparable genes is a critical work for researchers, and if done correctly, might save countless lives. For all these reasons, correct clustering is becoming increasingly important in bioinformatics. Therefore, if their data set resembles our microarray data, researchers from other fields can employ this technology.

Conflicts of Interest

None declared.

References

1. Richter DC, Ott F, Auch AF, Schmid R, Huson DH. MetaSim: a sequencing simulator for genomics and metagenomics. In: Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches. Hoboken, NJ: Wiley; May 03, 2011.
2. White C, Chan DW, Zhang Z. Bioinformatics strategies for proteomic profiling. Clin Biochem 2004 Jul;37(7):636-641. [doi: 10.1016/j.clinbiochem.2004.05.004] [Medline: 15234244]
3. Smyth GK. Limma: linear models for microarray data. In: Bioinformatics and Computational Biology Solutions Using R and Bioconductor. New York, US: Springer; 2005:397-420.
4. Letunic I, Bork P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res 2011 Jul 05;39(Web Server issue):W475-W478 [FREE Full text] [doi: 10.1093/nar/gkr201] [Medline: 21470960]

5. Godini R, Fallahi H. A brief overview of the concepts, methods and computational tools used in phylogenetic tree construction and gene prediction. Meta Gene 2019 Sep;21:100586. [doi: 10.1016/j.mgene.2019.100586]
6. Carter GW, Prinz S, Neou C, Shelby JP, Marzolf B, Thorsson V, et al. Prediction of phenotype and gene expression for combinations of mutations. Mol Syst Biol 2007 Mar 27;3(1):96 [FREE Full text] [doi: 10.1038/msb4100137] [Medline: 17389876]
7. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Studying gene expression and function. In: Molecular Biology of the Cell, 4th edition. New York, US: Garland Science; 2002.
8. Letunic I, Bork P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 2007 Jan 01;23(1):127-128. [doi: 10.1093/bioinformatics/btl529] [Medline: 17050570]
9. Ashraf FB, Ajwad R, Mottalib MA. A novel gene-tree based approach to infer relations among disease-genes across different cancer types. 2019 Presented at: International Conference on Electrical, Computer and Communication Engineering (ECCE); February 07-09, 2019; Cox'sBazar, Bangladesh. [doi: 10.1109/ecace.2019.8678921]
10. Fu L, Medico E. FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data. BMC Bioinformatics 2007 Jan 04;8(1):3 [FREE Full text] [doi: 10.1186/1471-2105-8-3] [Medline: 17204155]
11. Jain AK, Murty MN, Flynn PJ. Data clustering. ACM Comput. Surv 1999 Sep;31(3):264-323. [doi: 10.1145/331499.331504]
12. Lord E, Diallo AB, Makarenkov V. Classification of bioinformatics workflows using weighted versions of partitioning and hierarchical clustering algorithms. BMC Bioinformatics 2015 Mar 03;16(1):68 [FREE Full text] [doi: 10.1186/s12859-015-0508-1] [Medline: 25887434]
13. Greiff V, Bhat P, Cook SC, Menzel U, Kang W, Reddy ST. A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. Genome Med 2015 May 28;7(1):49 [FREE Full text] [doi: 10.1186/s13073-015-0169-8] [Medline: 26140055]
14. Wang, Chen XF, Shu YQ. Prediction of non-small cell lung cancer metastasis-associated microRNAs using bioinformatics. Am J Cancer Res 2015;5(1):32-51 [FREE Full text] [Medline: 25628919]
15. Gupta NT, Adams KD, Briggs AW, Timberlake SC, Vigneault F, Kleinstein SH. Hierarchical Clustering Can Identify B Cell Clones with High Confidence in Ig Repertoire Sequencing Data. J Immunol 2017 Mar 15;198(6):2489-2499 [FREE Full text] [doi: 10.4049/jimmunol.1601850] [Medline: 28179494]
16. Xu C, Su Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. Bioinformatics 2015 Jun 15;31(12):1974-1980 [FREE Full text] [doi: 10.1093/bioinformatics/btv088] [Medline: 25805722]
17. Danielsson P. Euclidean distance mapping. Computer Graphics and Image Processing 1980 Nov;14(3):227-248. [doi: 10.1016/0146-664x(80)90054-4]
18. Klove T, Lin T, Tsai S, Tzeng W. Permutation Arrays Under the Chebyshev Distance. IEEE Trans. Inform. Theory 2010 Jun;56(6):2611-2617. [doi: 10.1109/tit.2010.2046212]
19. Everitt BS, Landau S, Leese M, Stahl D. Cluster Analysis, 5th Edition. New York, US: John Wiley & Son; 2011.
20. Pop M, Salzberg SL. Bioinformatics challenges of new sequencing technology. Trends Genet 2008 Mar;24(3):142-149 [FREE Full text] [doi: 10.1016/j.tig.2007.12.006] [Medline: 18262676]
21. Paaby AB, Rockman MV. The many faces of pleiotropy. Trends Genet 2013 Feb;29(2):66-73 [FREE Full text] [doi: 10.1016/j.tig.2012.10.010] [Medline: 23140989]
22. Cook JR, Carta L, Galatioto J, Ramirez F. Cardiovascular manifestations in Marfan syndrome and related diseases; multiple genes causing similar phenotypes. Clin Genet 2015;87(1):11-20. [doi: 10.1111/cge.12436] [Medline: 24867163]
23. McClellan JM, Susser E, King M. Schizophrenia: a common disease caused by multiple rare alleles. Br J Psychiatry 2007 Mar 02;190(3):194-199. [doi: 10.1192/bjp.bp.106.025585] [Medline: 17329737]
24. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. Science 2008 Nov 07;322(5903):881-888 [FREE Full text] [doi: 10.1126/science.1156409] [Medline: 18988837]
25. Davidsohn N, Pezone M, Vernet A, Graveline A, Oliver D, Slomovic S, et al. A single combination gene therapy treats multiple age-related diseases. Proc Natl Acad Sci U S A 2019 Nov 19;116(47):23505-23511 [FREE Full text] [doi: 10.1073/pnas.1910073116] [Medline: 31685628]
26. Vega-pons S, Ruiz-Shulcloper J. A survey of clustering ensemble algorithms. Int. J. Patt. Recogn. Artif. Intell 2011 Nov 21;25(03):337-372. [doi: 10.1142/s0218001411008683]
27. Strehl A, Ghosh J. Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions. Journal of machine learning research 2002 Feb 12;3:583-617. [doi: 10.1002/widm.32]
28. Fern XZ, Brodley CE. Solving cluster ensemble problems by bipartite graph partitioning. 2004 Jul 04 Presented at: Proceedings of the twenty-first international conference on Machine learning; July 4-8, 2004; Banff, AB, Canada. [doi: 10.1145/1015330.1015414]
29. Caruana R, Elhawary M, Nguyen N, Smith C. Meta clustering. 2006 Presented at: Sixth International Conference on Data Mining (ICDM'06); December 18-22, 2006; Hong Kong, China. [doi: 10.1109/icdm.2006.103]
30. Rees D. Summarizing data by numerical measures. In: Essential statistics. Boston, USA: Springer; 1989:24-38.
31. Zhu Z, Ong Y, Dash M. Markov blanket-embedded genetic algorithm for gene selection. Pattern Recognition 2007 Nov;40(11):3236-3248. [doi: 10.1016/j.patcog.2007.02.007]

32. Gene enrichment Profiler. Center for Computational and Integrative Biology. URL: http://xavierlab2.mgh.harvard.edu/EnrichmentProfiler/help.html [accessed 2022-05-10]

Abbreviations

ALL-AML: acute lymphoblastic leukemia-acute myeloid leukemia
ASW: average silhouette width

Edited by A Mavragani; submitted 02.06.21; peer-reviewed by D Sengupta, R Zhang, M Hetti Arachchilage; comments to author 28.08.21; revised version received 10.05.22; accepted 29.05.22; published 17.06.22.

Please cite as:
Choudhury J, Ashraf FB
An Analysis of Different Distance-Linkage Methods for Clustering Gene Expression Data and Observing Pleiotropy: Empirical Study
JMIR Bioinform Biotech 2022;3(1):e30890
URL: https://bioinform.jmir.org/2022/1/e30890
doi:10.2196/30890
PMID:

Original Paper

Monitoring Physical Behavior in Rehabilitation Using a Machine Learning–Based Algorithm for Thigh-Mounted Accelerometers: Development and Validation Study

Frederik Skovbjerg[1], MSc; Helene Honoré[1], MSc; Inger Mechlenburg[2], DMSc; Matthijs Lipperts[3], MSc; Rikke Gade[4], PhD; Erhard Trillingsgaard Næss-Schmidt[1,2], PhD

[1]Research Unit, Hammel Neurorehabilitation Centre & University Research Clinic, Hammel, Denmark

[2]Department of Clinical Medicine, Aarhus University, Aarhus, Denmark

[3]Department of Medical Information and Communication Technology, St. Anna Hospital, Geldrop, Netherlands

[4]Section of Media Technology, Aalborg University, Aalborg, Denmark

Corresponding Author:
Frederik Skovbjerg, MSc
Research Unit
Hammel Neurorehabilitation Centre & University Research Clinic
Voldbyvej 15
Hammel, 8450
Denmark
Phone: 45 28739264
Email: freskv@rm.dk

Abstract

Background: Physical activity is emerging as an outcome measure. Accelerometers have become an important tool in monitoring physical behavior, and newer analytical approaches of recognition methods increase the degree of details. Many studies have achieved high performance in the classification of physical behaviors through the use of multiple wearable sensors; however, multiple wearables can be impractical and lower compliance.

Objective: The aim of this study was to develop and validate an algorithm for classifying several daily physical behaviors using a single thigh-mounted accelerometer and a supervised machine-learning scheme.

Methods: We collected training data by adding the behavior classes—running, cycling, stair climbing, wheelchair ambulation, and vehicle driving—to an existing algorithm with the classes of sitting, lying, standing, walking, and transitioning. After combining the training data, we used a random forest learning scheme for model development. We validated the algorithm through a simulated free-living procedure using chest-mounted cameras for establishing the ground truth. Furthermore, we adjusted our algorithm and compared the performance with an existing algorithm based on vector thresholds.

Results: We developed an algorithm to classify 11 physical behaviors relevant for rehabilitation. In the simulated free-living validation, the performance of the algorithm decreased to 57% as an average for the 11 classes (F-measure). After merging classes into sedentary behavior, standing, walking, running, and cycling, the result revealed high performance in comparison to both the ground truth and the existing algorithm.

Conclusions: Using a single thigh-mounted accelerometer, we obtained high classification levels within specific behaviors. The behaviors classified with high levels of performance mostly occur in populations with higher levels of functioning. Further development should aim at describing behaviors within populations with lower levels of functioning.

(JMIR Bioinform Biotech 2022;3(1):e38512) doi:10.2196/38512

KEYWORDS

activity recognition; random forest; acquired brain injury; biometric monitoring; machine learning; physical activity

Introduction

Physical behavior (PB) includes both physical activity (PA) and inactivity, which are both topics of increasing interest in health care. The health benefits associated with PA are well-established [1], which has resulted in the use of PA as prevention and a part of treatment and rehabilitation [2]. The prescription of PA has evolved within a wide range of diseases with long-term health impacts such as diabetes, cardiovascular diseases, obstructive pulmonary diseases, and rheumatoid arthritis [2-6]. Many such subgroups in our societies will continue to need rehabilitation to promote functional recovery, reduce the risk of comorbidities, and prevent the secondary effects of disease [7,8].

In the field of physical and rehabilitation medicine (PRM), functional outcomes and capabilities are of great interest. Today, the International Classification of Functioning, Disability and Health (ICF) is the conceptual foundation of physical and rehabilitation medicine as a biopsychosocial framework for clinicians, researchers, and policy makers [9]**.** Rehabilitation interventions often target functional abilities and limitations to promote physical and cognitive functioning, participation, and the modification of personal and environmental factors [9,10]. These functional aims in daily living require measurement properties that can identify such factors in a meaningful way. Outcome measures used in rehabilitation research are often subjective or self-reported measures [11], which are associated with various limitations such as information bias, intrusiveness, and timeliness [12-14]**,** and more objective measures are warranted. The use of wearable technologies offers an objective and complementary insight to subjective measures. The objective classification and quantification of activities such as standing, sitting, wheelchair ambulation, walking, or running can provide information on changes in functional disability. Additionally, it can indicate changes in more holistic measures, referred to as ICF-related items on activity and participation levels, contextual factors, or transport options such as stair climbing, cycling, and vehicle driving. The development of wearable sensor technologies, such as accelerometers, has added the possibility of monitoring PB continuously for longer periods, making it opportune to investigate the changes and habitual patterns of PB [15,16].

The emerging analytical approaches of raw signal processing use pattern recognition to classify functional activities. Threshold-based algorithms have contributed beneficial frameworks with high accuracies [17]. However, machine-learning techniques have proven useful [18], and many studies have achieved high performance in the classification of physical behaviors through the use of multiple wearable sensors [19-22]. Multiple wearables can be impractical and lead to low compliance [23]; it is necessary to investigate classification potentials that only use 1 sensor device [21,22]. Therefore, the purpose of this study was to further develop and validate a machine learning–based algorithm for thigh-mounted accelerometers. We specifically intended to add the following classes of PB to an existing algorithm: running, cycling, stair climbing, wheelchair ambulation, and vehicle driving.

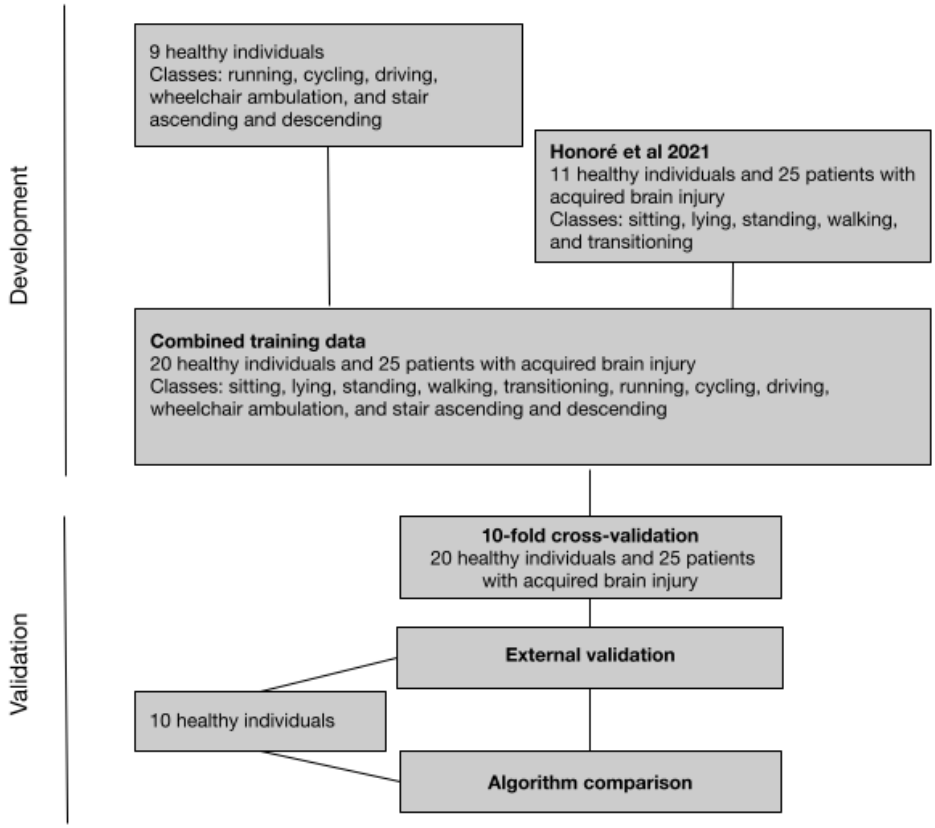
Methods

Design

This study was a development and validation study in 2 phases. For a study overview, see Figure 1.

The application of our algorithm was aimed at patients undergoing neurorehabilitation, and the training data collected in the development phase of this study were combined with the training data from a previous study [24], collected in a population of both healthy people and patients with acquired brain injury. The following method section only describes the data collected in this study. The validation phase describes the algorithm developed based on the combined training data from both studies. Due to ethical considerations, the algorithm was validated in a new cohort of healthy individuals, and performance was compared to another algorithm based on vector thresholds [17].

Figure 1. Study overview.



Instrumentation

A triaxial accelerometer (AX3; Axivity) was mounted on the dominant leg, on the lateral part of the thigh approximately 10 cm above the apex patella. The x-axis was oriented toward the floor in the standing position, as implied by the downward position of the USB port and stated by the visible written information on the device. The accelerometers were programmed with a sampling frequency at 100 Hz, consistent with the method of Honoré et al [24].

Development Phase

A pragmatic data collection method was applied. A protocol described the positioning, direction, and attachment of the accelerometer. We used 3 taps directly on the accelerometer as a data marker for the start and stop of the recording of behaviors. The participants were asked to perform a minimum of 10 minutes of continuous activity for each PB with the exception of stair climbing. Whenever possible, the behaviors were performed at locations of the participants' choosing or alternatively, at locations proposed by FS. Instructions were given immediately before each performed behavior, and data were extracted immediately after. Participants contributed the behaviors of convenience and provided information on gender, age, and height (Table 1).

Table 1. Description and characteristics of the participants (N=9) contributing training data. The total amount of training data for all participants and the distribution within each activity are reported.

Class	Gender (male, female), n	Age (year), mean (SD)	Height (cm), mean (SD)	Total duration[a] (h, min)
All participants	4, 5	36.1 (13.4)	176.7 (5.7)	21, 27
Running	4, 2	30.8 (13.0)	179.8 (4.5)	4, 52
Cycling	4, 2	42.8 (14.6)	179.5 (5.2)	6, 10
Stair climbing				
Ascending	3, 2	31.4 (12.6)	178.8 (4.7)	0, 10
Descending	3, 2	31.4 (12.6)	178.8 (4.7)	0, 9
Driving	2, 2	40 (16.7)	179.2 (5.3)	5, 53
Wheelchair ambulation	3, 2	33 (7.7)	176.8 (5.0)	4, 13

[a]Total duration describes the total amount of training data.

Data Preprocessing and Learning Scheme

Each activity sequence containing 1 PB was manually identified by the data markers and extracted from the original data file using OMGUI configuration and analysis tool (V43 ; Open Movement). The raw accelerometer data was processed in a custom-made MATLAB script (R2020b; MathWorks) for the manual label annotation of each sample period of 1 second with a sample overlap of 0.5 seconds. All manual annotation and classification were done by FS. For all accelerometer axes, we extracted the features of 1-second samples. Based on the findings of Yan et al [25], a preselected subset of features was used (Textbox 1). To model baseline PB classifications, we used the nonlinear classifier random forest with default hyperparameters in Weka software (version 3.8.4; University of Waikato) [26,27].

Textbox 1. Features used.

Features

- Mean values
- SDs
- Root mean square values
- Maximum number of peaks
- Highest value of axes
- Lowest value of axes
- Number of distinctive points
- Pearson correlation between axes

Validation Phase

The validation phase consisted of a k-fold cross-validation, an external validation, and an algorithm comparison procedure. To evaluate the potential of the algorithm, we initially performed a stratified 10-fold cross-validation on the training data collected from 9 healthy individuals and the data from Honoré et al [24] from 11 healthy individuals and 25 patients, and the subsets were randomly split. In the external validation, 10 healthy individuals who did not contribute to the training data were asked to participate in the external validation protocol. The protocol consisted of a semistandardized session, where the participants were instructed to carry out a protocol of PBs at a self-determined level of pace, duration, and order, in a setup that enabled the performance of all behaviors. Throughout the session, the participants wore an accelerometer on the thigh and a chest-mounted GoPro camera was used to identify the ground truth of the PBs performed. The video recording was time-synchronized with the accelerometer data using ELAN tool (version 6.4; Max Planck Institute for Psycholinguistics) [28] and was then manually labeled by FS as a criterion measure. Data collected through the external validation protocol were then used as a test set and a second-by-second analysis was conducted by testing the performance of the algorithm in the validation data.

The algorithm for comparison was chosen based on previous use by research institutions in the central regions of Jutland, Denmark [29-33]. We compared the performance of the algorithm by Lipperts et al [17] and our algorithm by analyzing the data collected in the external validation protocol with both algorithms. We reported the results on a total time basis compared to the ground truth and through confusion matrices for both algorithms. In accounting for differences in the available classes between the algorithms, we adjusted our algorithm to only include classes comparable to the classes by Lipperts et al [17]. Therefore, we excluded the implemented wheelchair ambulation and vehicle driving classes, and similarly, we excluded the data parts containing wheelchair ambulation and vehicle driving from the validation sessions. To create a fair basis for comparison, we merged the relevant classes, sitting and lying, to account for sedentary behavior. Additionally, we merged walking, stair climbing, and transitioning under the walking class, corresponding to the walking class by Lipperts et al [17].

Statistics

For evaluating the performance of the algorithm, we presented confusion matrices for the developed models. We interchangeably used the term performance to refer to the main evaluation metric: F-measure [34,35]. We calculated the F-measure as the harmonic mean between the positive predictive value and sensitivity [36]. In the algorithm comparison, we reported mean errors in durations as calculated by $(|duration^{Alg} - duration^{GT}|) / duration^{GT}$, where $duration^{Alg}$ is the total duration of all correctly classified seconds of either algorithm and $duration^{GT}$ is the duration of the ground truth.

Ethical Considerations

The study was conducted in accordance with the Helsinki Declaration of 2008 [37], and the General Data Protection Regulation was followed. This study did not require approval from the regional ethics committee, as noninterventional studies do not need approval by the Region Committee on Biomedical Research Ethics in Denmark. We only recruited healthy participants, and written informed consent was obtained from all participants.

Results

Participants and Training Data

The data gathering and preprocessing resulted in no missing or exclusion of data. In total, 9 healthy participants contributed data for training the algorithm. Participants of various ages, heights, and gender were included. We strived to accumulate >4 hours of running, cycling, driving, and wheelchair ambulation and 10 sessions of ascending and descending stair climbing (Table 1).

K-fold Cross-validation

By combining data from Honoré et al [24] with the training data in this study, the algorithm constituted 11 classes of PBs. The initial evaluation by a stratified 10-fold cross-validation (Table 2) showed strong agreement between the labels and the classifications performed by the algorithm, with an average F-measure of 92.8% for all classified PBs—a performance strong enough to be tested in simulated free-living conditions. The performance in classifying running and cycling showed high agreement by reaching F-measures of 100 and 99.6%, respectively. The classification of stair climbing likewise showed promising results by reaching F-measures of 91.4% and 90.2% for ascending and descending stairs, respectively. In discriminating between the 4 behaviors involving similar inactive lower extremity postures, the algorithm showed an F-measure of 92.7% for sitting and 92.3% for lying, whereas driving and wheelchair ambulation reached 99.4% and 98.9%, respectively. Walking and standing yielded F-measures of 89% and 96.3%, respectively. Transitioning resulted in the lowest F-measure of 72.5%.

Table 2. Confusion matrix from stratified 10-fold cross-validation. Correctly and incorrectly classified seconds of physical behavior by the algorithm (columns) and the ground truth (rows). Seconds overlap by 0.5 second.

Ground truth	Algorithm										
	Sitting	Transitioning	Walking	Standing	Lying	Ascending stairs	Cycling	Descending stairs	Running	Driving	Wheelchair ambulation
Sitting	2236	59	0	0	68	0	2	0	0	64	106
Transitioning	27	1683	286	46	32	5	104	5	0	92	163
Walking	0	220	3103	21	0	13	15	15	0	0	0
Standing	0	48	4	1688	5	0	3	0	0	0	0
Lying	17	48	0	0	1935	0	0	0	0	64	51
Ascending stairs	0	7	63	0	0	1060	31	24	6	0	0
Cycling	0	36	20	1	0	19	44,280	7	0	1	53
Descending stairs	0	0	105	0	0	30	12	979	7	0	0
Running	0	0	7	0	0	1	1	8	34,985	0	0
Driving	4	44	0	0	20	0	3	0	0	42,148	109
Wheelchair ambulation	5	52	0	0	19	0	28	0	0	80	30,134

External Validation

The external validation protocol resulted in 10 sessions of PB monitoring, which included all the behaviors of interest performed by 10 healthy participants recruited at Hammel Neurorehabilitation Center and University Research Clinic, Denmark. Participant characteristics are described in Table 3. The performance of the algorithm in the validation data showed moderate agreement between the ground truth and the classifications by the algorithm with 57% as the average F-measure for all classifications (Table 4). The performance in classifying running and cycling remained high by reaching 88.7% and 87.1%, respectively. The classification of stair climbing decreased to an F-measure of 44.8% for ascending and 25.5% for descending stair climbing. In discriminating between the 4 behaviors involving inactive lower extremity postures, the algorithm showed an F-measure of 63.7% for sitting, 66.8% for lying, 77.1% for driving, and 31% for wheelchair ambulation. Walking, standing, and transitioning were classified with F-measures of 55%, 67.1%, and 20%, respectively.

Table 3. Characteristics of participants contributing data from the external validation.

Characteristic	Value
Participants, n	10
Gender (male, female), n	5, 5
Age (year), mean	43.6
Height (cm), mean	174.4
Duration[a] (min, sec), mean	12, 58

[a]Duration describes the average time taken to complete the validation session.

Table 4. Confusion matrix from the external validation. Correctly and incorrectly classified seconds of physical behavior by the algorithm (columns) and the ground truth (rows). Seconds overlap by 0.5 second.

Ground truth	Algorithm										
	Sitting	Transitioning	Walking	Standing	Lying	Ascending stairs	Cycling	Descending stairs	Running	Driving	Wheelchair ambulation
Sitting	746	28	15	0	236	2	10	5	0	163	151
Transitioning	1	131	64	0	10	4	66	14	7	35	30
Walking	5	253	1178	72	0	60	108	191	89	50	69
Standing	1	190	118	589	1	31	47	23	16	8	3
Lying	208	58	4	0	746	0	0	0	0	54	136
Ascending stairs	0	8	143	29	0	184	40	38	42	0	0
Cycling	0	57	57	6	0	19	1673	13	12	5	18
Descending stairs	0	17	520	26	0	30	7	162	12	0	0
Running	0	13	28	7	0	4	5	18	1014	0	0
Driving	23	50	31	0	34	0	4	15	4	3124	542
Wheelchair ambulation	1	140	52	0	0	4	23	16	2	830	453

Algorithm Comparison

To compare the performance of the 2 algorithms, noncomparable classes were excluded. The validation sessions subsequently averaged 7.21 minutes and included the behaviors lying, sitting, standing, transitioning, walking, stair climbing, running, and cycling. The results of the merged algorithm showed high performance by reaching an averaging F-measure of 85.3% for all classes in the external validation data (Table 5). In comparison, Lipperts et al's [17] algorithm showed an average F-measure of 81.1% (Table 6). Table 7 shows the mean error by the algorithms for each behavior class across the 10 validation sessions. The results indicated high agreement between the ground truth and both algorithms when classifying sedentary behavior, walking, running, and cycling, whereas both algorithms showed poor performance in classifying standing. The mean error for Lipperts et al's [17] algorithm varied between 13.6% to 72.8%, consequently overestimating sedentary and standing behavior, and was hardly influenced by not detecting running and cycling in 2 and 1 sessions of validation, respectively. The mean error for our algorithm varied between 7.9% to 41.7%, consequently underestimating all classes.

Table 5. Confusion matrix from the adjusted algorithm in external validation data. Correctly and incorrectly classified seconds of physical behavior by the algorithm (columns) and the ground truth (rows). Seconds overlap by 0.5 second.

Ground truth	Algorithm				
	Sedentary	Walking	Standing	Cycling	Running
Sedentary	2046	143	0	11	0
Walking	10	2381	95	122	95
Standing	0	359	568	40	16
Cycling	0	191	6	1631	8
Running	0	66	7	6	1010

Table 6. Confusion matrix for Lipperts et al's [17] algorithm in the external validation data. Correctly and incorrectly classified seconds of physical behavior by the algorithm (columns) and the ground truth (rows). Seconds overlap by 0.5 second.

Ground truth	Algorithm				
	Sedentary	Walking	Standing	Cycling	Running
Sedentary	2124	4	72	0	0
Walking	219	1999	443	12	30
Standing	28	156	776	12	11
Cycling	0	122	205	1491	7
Running	0	203	43	0	834

Table 7. Mean error, SD, and range of output duration parameters for analyzing the external validation data by the 2 algorithms. We calculated the mean error, SD, and minimum and maximum error percentage across the 10 validation sessions within each activity class.

Algorithm, parameter	Activities				
	Sedentary	Standing	Walking	Running	Cycling
Lipperts et al [17]					
Mean error (%)	13.6	72.8	14.5	27.2	21.8
SD (%)	7.2	72.8	6.2	40.9	29.8
Minimum error (%)	6.4	22.2	2.9	1.6	1.3
Maximum error (%)	28.6	267	22.2	100	100
Skovbjerg et al					
Mean error (%)	7.9	41.7	12.4	10	8.1
SD (%)	4	14.1	7	15.6	5.3
Minimum error (%)	2.4	19	2.8	0	0
Maximum error (%)	13.9	59.1	23	51.5	16.6

Discussion

Principal Findings

We developed an algorithm to classify 11 PBs related to daily living in rehabilitation. The cross-validation demonstrated high performance (93%), and the validation of the algorithm in a free-living setting was reasonable. The algorithm showed moderate performance (57%) when applied to simulated free-living data. The algorithm performed well in classifying cycling and running, whereas an acceptable level of performance was found in classifying driving. In classifying the remaining behaviors, the algorithm showed low to moderate performance ranging from 20% to 67%. In comparison to a validated algorithm by Lipperts et al [17], our adjusted algorithm showed equally strong performance and high agreement with ground truth annotations after merging relevant classes. The significant performance decrease between cross-validation and external validation may be explained by the fact that in the cross-validation, different samples from the same individual were included in both training and test splits. In the external validation, the individuals and their specific motion pattern were not included in the training data.

Discriminating Rehabilitation Relevant Physical Behaviors

The behaviors classifiable by the algorithm were based on the rationale and aims of rehabilitation. Our results showed lower performance in discriminating behaviors performed in sitting postures, which can be explained by their similar body positioning and behavioral characteristics. Although discriminating these behaviors is important when considering activity and participation from an ICF perspective, the differences within sitting, wheelchair ambulation, and driving might be clinically irrelevant from a perspective of monitoring PA and energy expenditure at a body function and anatomy level. In a visual inspection of accelerometer data, signals from the 3 behaviors revealed only insignificant differences. Likewise, the algorithm had difficulties discriminating between the PBs by the accessible features. Overall, the algorithm performed better in discriminating behaviors with larger variations in body position and movement trajectories, mostly occurring in patients with higher levels of functioning.

Comparison to Existing Literature

Pavey et al [38] achieved a 93% overall accuracy for classifying the PBs—sedentary, stationary, walking, and running—using a wrist-worn accelerometer with the random forest classifier in laboratory settings among 21 healthy participants, evaluated using leave-one-subject-out cross-validation. A back validation in free-living using activPAL as a reference standard for

stepping versus nonstepping showed high agreement. Alber et al [39] used a waist-worn accelerometer for classifying lying, standing, sitting, walking, wheelchair ambulation, and stair climbing among 13 subjects with incomplete spinal cord injury, using a support vector machine (SVM) classifier. Their laboratory-based algorithm decreased from 92% to 55% when tested on home-based data, whereas their home-based algorithm reached 86%, evaluated using within-subject cross-validation.

When focusing on single thigh-mounted accelerometry, Awais et al [20] reached a mean F-measure ranging from 68% to 76% with different combinations of features, using SVM classifier in identifying sitting, lying, standing, and walking among 20 older people in free-living conditions evaluated using leave-one-subject-out cross-validation. Likewise, Tang et al [22] investigated the number of sensors and found a mean F-measure of 76% using a single thigh-worn accelerometer and SVM classifier in identifying sitting, lying, and standing among 42 healthy participants in semistandardized laboratory settings, evaluated using leave-one-subject-out cross-validation. In comparison to Tang et al [22] and Awais et al [20], we reached an F-measure of 57%, evaluated using simulated free-living conditions with 11 classes of PB. For the abovementioned studies, they all use fewer classes of activities, which expectedly will increase the performance of an algorithm and might explain why our algorithm does not reach their level. As indicated in the algorithm comparison, the level of performance required for valid estimation can be obtained by merging relevant classes. It will compromise the degree of details but simultaneously add the possibility of adjusting the measures of PB in relation to the aims.

Algorithm for Patients With Acquired Brain Injury

Our algorithm was aimed at patients undergoing neurorehabilitation. Classifying behaviors within subgroups potentially exposed to characteristic movement patterns, the behavior classes—sitting, lying, standing, walking, and transitioning—were partly based on training data from the population of interest [24]. Some specific PBs or movement patterns such as transitioning and walking may be more influenced by disease-specific characteristics than others. Similarly, some PBs can be less prone to disease-specific characteristics depending on functional level or disease severity. Using healthy individuals for training the algorithm relies on the rationale that a higher functional level is required to perform PB, such as running, and hence is associated with a movement pattern comparable to movement patterns in healthy individuals. Adversely, PBs, such as wheelchair ambulation, may be independent of specific movement characteristics. In principle, the training data should be gathered in the target population to capture complex movements influenced by disabilities, although it can be argued that activities less prone to disease-specific characteristics can be gathered in healthy populations due to ethical considerations.

Limitations

The training data for this study was collected in a setup similar to a laboratory setting. Although the PBs were performed in a free-living setting, only 1 PB was recorded in each session, and therefore, the composition of PBs in free-living was not reflected in the training data. Our training data were probably influenced by a severe class imbalance between the newly gathered classes and the classes gathered in Honoré et al [24], which might have affected the performance of the algorithm in the validation data. Less available training data decrease the performance by reducing the ability of a classifier to generalize patterns not seen before. Balancing minority classes through supplementary data gathering might be advantageous in future work. We did not include a free-living validation but designed a semistandardized session aimed at simulating free-living. All validation sessions were conducted in the same environment—they only lasted 10-20 minutes, and the participants were enforced to perform PBs corresponding to the classes of the algorithm. Variation between sessions consisted of the order and duration of the behaviors. We used video recordings as a criterion measure for labeling accelerometer signals and further merged annotation definitions with Honoré et al [24] to align the labeling protocol, thus the ground truth labeling was only performed by FS and the reliability was not evaluated. The algorithm comparison procedure might have been influenced by differences in annotation definitions, leading to an underestimation of the performance by Lipperts et al's [17] algorithm. Likewise, the cropping procedure have introduced minor differences in the data analyzed by each algorithm.

Clinical Implications

The algorithm comparison revealed that our merged algorithm, constituting 5 classes, reached an acceptable level of agreement with both the algorithm of Lipperts et al [17] and the ground truth. However, the 11-class algorithm did not show acceptable levels of performance within all classes, indicating that the number of behavior classes and similarities between classes may influence the obtainable level of performance. To monitor physical behavior within various functional levels of patients undergoing neurorehabilitation, further research and changes in the monitor setup are required to attain the desired levels, especially within wheelchair ambulation. Furthermore, this study provided an external validation performed in a simulated free-living setting, which constitutes an estimate of the algorithm's performance in clinical settings.

Conclusion

We developed an algorithm for classifying rehabilitation-relevant physical behaviors. We successfully added the classes of running and cycling, which were classified with high performance in a simulated free-living setting. Furthermore, we added stair climbing, wheelchair ambulation, and vehicle driving, which showed high performance in the 10-fold cross-validation on training data, but low to moderate performance in the free-living setting for new individuals. Increasing the implications for rehabilitation use might be done by focusing on the performance in classifying behaviors within populations with lower levels of functioning and within transport ambulation and the use of assistive devices.

Acknowledgments

The authors wish to thank all participants who helped facilitate the data collection.

Conflicts of Interest

None declared.

References

1. Warburton DER, Bredin SSD. Health benefits of physical activity: a systematic review of current systematic reviews. Curr Opin Cardiol 2017 Sep;32(5):541-556. [doi: 10.1097/HCO.0000000000000437] [Medline: 28708630]
2. Ruegsegger GN, Booth FW. Health benefits of exercise. Cold Spring Harb Perspect Med 2018 Jul 02;8(7):a029694 [FREE Full text] [doi: 10.1101/cshperspect.a029694] [Medline: 28507196]
3. Teich T, Zaharieva DP, Riddell MC. Advances in exercise, physical activity, and diabetes mellitus. Diabetes Technol Ther 2019 Feb;21(S1):S112-S122. [doi: 10.1089/dia.2019.2509] [Medline: 30785316]
4. Elagizi A, Kachur S, Carbone S, Lavie CJ, Blair SN. A review of obesity, physical activity, and cardiovascular disease. Curr Obes Rep 2020 Dec;9(4):571-581. [doi: 10.1007/s13679-020-00403-z] [Medline: 32870465]
5. Rabe KF, Watz H. Chronic obstructive pulmonary disease. Lancet 2017 May 13;389(10082):1931-1940. [doi: 10.1016/S0140-6736(17)31222-9] [Medline: 28513453]
6. Katz P, Andonian BJ, Huffman KM. Benefits and promotion of physical activity in rheumatoid arthritis. Curr Opin Rheumatol 2020 May;32(3):307-314. [doi: 10.1097/BOR.0000000000000696] [Medline: 32141951]
7. Cieza A, Causey K, Kamenov K, Hanson SW, Chatterji S, Vos T. Global estimates of the need for rehabilitation based on the Global Burden of Disease study 2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2021 Dec 19;396(10267):2006-2017 [FREE Full text] [doi: 10.1016/S0140-6736(20)32340-0] [Medline: 33275908]
8. European Physical and Rehabilitation Medicine Bodies Alliance. White Book on Physical and Rehabilitation Medicine in Europe. Chapter 2. Why rehabilitation is needed by individual and society. Eur J Phys Rehabil Med 2018 Apr;54(2):166-176 [FREE Full text] [doi: 10.23736/S1973-9087.18.05145-6] [Medline: 29565103]
9. European Physical and Rehabilitation Medicine Bodies Alliance. White Book on Physical and Rehabilitation Medicine (PRM) in Europe. Chapter 1. Definitions and concepts of PRM. Eur J Phys Rehabil Med 2018 Apr;54(2):156-165 [FREE Full text] [doi: 10.23736/S1973-9087.18.05144-4] [Medline: 29565102]
10. Stucki G, Cieza A, Melvin J. The International Classification of Functioning, Disability and Health (ICF): a unifying model for the conceptual description of the rehabilitation strategy. J Rehabil Med 2007 May;39(4):279-285 [FREE Full text] [doi: 10.2340/16501977-0041] [Medline: 17468799]
11. Wade DT, Smeets RJEM, Verbunt JA. Research in rehabilitation medicine: methodological challenges. J Clin Epidemiol 2010 Jul;63(7):699-704. [doi: 10.1016/j.jclinepi.2009.07.010] [Medline: 19788953]
12. Sember V, Meh K, Sorić M, Starc G, Rocha P, Jurak G. Validity and reliability of international physical activity questionnaires for adults across EU countries: systematic review and meta analysis. Int J Environ Res Public Health 2020 Sep 30;17(19):7161 [FREE Full text] [doi: 10.3390/ijerph17197161] [Medline: 33007880]
13. Kjeldsen SS, Brodal L, Brunner I. Activity and rest in patients with severe acquired brain injury: an observational study. Disabil Rehabil 2022 Jun;44(12):2744-2751. [doi: 10.1080/09638288.2020.1844317] [Medline: 33161752]
14. Gebruers N, Vanroy C, Truijen S, Engelborghs S, De Deyn PP. Monitoring of physical activity after stroke: a systematic review of accelerometry-based measures. Arch Phys Med Rehabil 2010 Feb;91(2):288-297. [doi: 10.1016/j.apmr.2009.10.025] [Medline: 20159136]
15. Shephard R, Tudor-Locke C, editors. The Objective Monitoring of Physical Activity: Contributions of Accelerometry to Epidemiology, Exercise Science and Rehabilitation. Cham, Switzerland: Springer; 2016.
16. Westerterp KR. Assessment of physical activity: a critical appraisal. Eur J Appl Physiol 2009 Apr;105(6):823-828. [doi: 10.1007/s00421-009-1000-2] [Medline: 19205725]
17. Lipperts M, van Laarhoven S, Senden R, Heyligers I, Grimm B. Clinical validation of a body-fixed 3D accelerometer and algorithm for activity monitoring in orthopaedic patients. J Orthop Translat 2017 Oct;11:19-29 [FREE Full text] [doi: 10.1016/j.jot.2017.02.003] [Medline: 29662766]
18. Farrahi V, Niemelä M, Kangas M, Korpelainen R, Jämsä T. Calibration and validation of accelerometer-based activity monitors: a systematic review of machine-learning approaches. Gait Posture 2019 Feb;68:285-299 [FREE Full text] [doi: 10.1016/j.gaitpost.2018.12.003] [Medline: 30579037]
19. Sasaki JE, Hickey AM, Staudenmayer JW, John D, Kent JA, Freedson PS. Performance of activity classification algorithms in free-living older adults. Med Sci Sports Exerc 2016 May;48(5):941-950 [FREE Full text] [doi: 10.1249/MSS.0000000000000844] [Medline: 26673129]
20. Awais M, Chiari L, Ihlen EAF, Helbostad JL, Palmerini L. Physical activity classification for elderly people in free-living conditions. IEEE J Biomed Health Inform 2019 Jan;23(1):197-207. [doi: 10.1109/JBHI.2018.2820179] [Medline: 29994291]

21. Trost SG, Cliff DP, Ahmadi MN, Tuc NV, Hagenbuchner M. Sensor-enabled activity class recognition in preschoolers: hip versus wrist data. Med Sci Sports Exerc 2018 Mar;50(3):634-641. [doi: 10.1249/MSS.0000000000001460] [Medline: 29059107]
22. Tang QU, John D, Thapa-Chhetry B, Arguello DJ, Intille S. Posture and physical activity detection: impact of number of sensors and feature type. Med Sci Sports Exerc 2020 Aug;52(8):1834-1845 [FREE Full text] [doi: 10.1249/MSS.0000000000002306] [Medline: 32079910]
23. Sliepen M, Lipperts M, Tjur M, Mechlenburg I. Use of accelerometer-based activity monitoring in orthopaedics: benefits, impact and practical considerations. EFORT Open Rev 2019 Dec;4(12):678-685 [FREE Full text] [doi: 10.1302/2058-5241.4.180041] [Medline: 32010456]
24. Honoré H, Gade R, Nielsen JF, Mechlenburg I. Developing and validating an accelerometer-based algorithm with machine learning to classify physical activity after acquired brain injury. Brain Inj 2021 Mar 21;35(4):460-467. [doi: 10.1080/02699052.2021.1880026] [Medline: 33599161]
25. Yan N, Chen J, Yu T. A feature set for the similar activity recognition using smartphone. 2018 Dec 03 Presented at: 2018 10th International Conference on Wireless Communications and Signal Processing (WCSP); October 18-20, 2018; Hangzhou, China p. 1-6. [doi: 10.1109/wcsp.2018.8555704]
26. Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. 2nd ed. San Francisco, CA: Morgan Kaufmann; 2005.
27. Biau G, Scornet E. A random forest guided tour. Test 2016 Apr 19;25(2):197-227. [doi: 10.1007/s11749-016-0481-7]
28. ELAN. The Language Archive. URL: https://archive.mpi.nl/tla/elan [accessed 2022-07-11]
29. Næss-Schmidt E, Pedersen A, Christiansen D, Andersen N, Brincks J, Grimm B, et al. Daily activity and functional performance in people with chronic disease: a cross-sectional study. Cogent Med 2020 Jan 9;7(1). [doi: 10.1080/2331205x.2020.1713280]
30. Sandell Jacobsen J, Thorborg K, Hölmich P, Bolvig L, Storgaard Jakobsen S, Søballe K, et al. Does the physical activity profile change in patients with hip dysplasia from before to 1 year after periacetabular osteotomy? Acta Orthop 2018 Dec 18;89(6):622-627 [FREE Full text] [doi: 10.1080/17453674.2018.1531492] [Medline: 30334645]
31. Hjorth MH, Mechlenburg I, Soballe K, Jakobsen SS, Roemer L, Stilling M. Physical activity is associated with the level of chromium but not with changes in pseudotumor size in patients with metal-on-metal hip arthroplasty. J Arthroplasty 2018 Sep;33(9):2932-2939. [doi: 10.1016/j.arth.2018.04.039] [Medline: 29807790]
32. Daugaard R, Tjur M, Sliepen M, Lipperts M, Grimm B, Mechlenburg I. Are patients with knee osteoarthritis and patients with knee joint replacement as physically active as healthy persons? J Orthop Translat 2018 Jul;14:8-15 [FREE Full text] [doi: 10.1016/j.jot.2018.03.001] [Medline: 30035028]
33. Kierkegaard S, Dalgas U, Lund B, Lipperts M, Søballe K, Mechlenburg I. Despite patient-reported outcomes improve, patients with femoroacetabular impingement syndrome do not increase their objectively measured sport and physical activity level 1 year after hip arthroscopic surgery. results from the HAFAI cohort. Knee Surg Sports Traumatol Arthrosc 2020 May 6;28(5):1639-1647. [doi: 10.1007/s00167-019-05503-5] [Medline: 31062043]
34. Trevethan R. Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice. Front Public Health 2017 Nov 20;5:307 [FREE Full text] [doi: 10.3389/fpubh.2017.00307] [Medline: 29209603]
35. Hossin M, Sulaiman MN. A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process 2015 Mar 31;5(2):01-11. [doi: 10.5121/ijdkp.2015.5201]
36. Zhang E, Zhang Y. F-Measure. In: Liu L, Özsu MT, editors. Encyclopedia of Database Systems. Boston, MA: Springer; 2009:1147.
37. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013 Nov 27;310(20):2191-2194. [doi: 10.1001/jama.2013.281053] [Medline: 24141714]
38. Pavey TG, Gilson ND, Gomersall SR, Clark B, Trost SG. Field evaluation of a random forest activity classifier for wrist-worn accelerometer data. J Sci Med Sport 2017 Jan;20(1):75-80. [doi: 10.1016/j.jsams.2016.06.003] [Medline: 27372275]
39. Albert MV, Azeze Y, Courtois M, Jayaraman A. In-lab versus at-home activity recognition in ambulatory subjects with incomplete spinal cord injury. J Neuroeng Rehabil 2017 Feb 06;14(1):10 [FREE Full text] [doi: 10.1186/s12984-017-0222-5] [Medline: 28166824]

Abbreviations

ICF: International Classification of Functioning, Disability and Health
PA: physical activity
PB: physical behavior
SVM: support vector machine

Edited by A Mavragani; submitted 06.04.22; peer-reviewed by H Li, M Albert; comments to author 17.05.22; revised version received 24.06.22; accepted 07.07.22; published 26.07.22.

Please cite as:
Skovbjerg F, Honoré H, Mechlenburg I, Lipperts M, Gade R, Næss-Schmidt ET
Monitoring Physical Behavior in Rehabilitation Using a Machine Learning–Based Algorithm for Thigh-Mounted Accelerometers: Development and Validation Study
JMIR Bioinform Biotech 2022;3(1):e38512
URL: https://bioinform.jmir.org/2022/1/e38512
doi:10.2196/38512
PMID:

Original Paper

Reducing Crowding in Emergency Departments With Early Prediction of Hospital Admission of Adult Patients Using Biomarkers Collected at Triage: Retrospective Cohort Study

Ann Corneille Monahan[1], MSHI, PhD; Sue S Feldman[2], MEd, RN, PhD; Tony P Fitzgerald[3,4], ScD

[1]University of Alabama at Birmingham, Birmingham, AL, United States

[2]Department of Health Services Administration, University of Alabama at Birmingham, Birmingham, AL, United States

[3]School of Mathematical Sciences, University College Cork, Cork, Ireland

[4]School of Public Health, University College Cork, Cork, Ireland

Corresponding Author:
Ann Corneille Monahan, MSHI, PhD
University of Alabama at Birmingham
1720 University Blvd
Birmingham, AL, 35294
United States
Phone: 1 2056174780
Email: monahanannc@gmail.com

Abstract

Background: Emergency department crowding continues to threaten patient safety and cause poor patient outcomes. Prior models designed to predict hospital admission have had biases. Predictive models that successfully estimate the probability of patient hospital admission would be useful in reducing or preventing emergency department "boarding" and hospital "exit block" and would reduce emergency department crowding by initiating earlier hospital admission and avoiding protracted bed procurement processes.

Objective: To develop a model to predict imminent adult patient hospital admission from the emergency department early in the patient visit by utilizing existing clinical descriptors (ie, patient biomarkers) that are routinely collected at triage and captured in the hospital's electronic medical records. Biomarkers are advantageous for modeling due to their early and routine collection at triage; instantaneous availability; standardized definition, measurement, and interpretation; and their freedom from the confines of patient histories (ie, they are not affected by inaccurate patient reports on medical history, unavailable reports, or delayed report retrieval).

Methods: This retrospective cohort study evaluated 1 year of consecutive data events among adult patients admitted to the emergency department and developed an algorithm that predicted which patients would require imminent hospital admission. Eight predictor variables were evaluated for their roles in the outcome of the patient emergency department visit. Logistic regression was used to model the study data.

Results: The 8-predictor model included the following biomarkers: age, systolic blood pressure, diastolic blood pressure, heart rate, respiration rate, temperature, gender, and acuity level. The model used these biomarkers to identify emergency department patients who required hospital admission. Our model performed well, with good agreement between observed and predicted admissions, indicating a well-fitting and well-calibrated model that showed good ability to discriminate between patients who would and would not be admitted.

Conclusions: This prediction model based on primary data identified emergency department patients with an increased risk of hospital admission. This actionable information can be used to improve patient care and hospital operations, especially by reducing emergency department crowding by looking ahead to predict which patients are likely to be admitted following triage, thereby providing needed information to initiate the complex admission and bed assignment processes much earlier in the care continuum.

(JMIR Bioinform Biotech 2022;3(1):e38845) doi:10.2196/38845

KEYWORDS

emergency care; prehospital; emergency; information system; crowding; boarding; exit block; medical informatics; application; health service research; personalized medicine; predictive medicine; model; probabilistic; polynomial model; decision support technique; decision support; evidence-based health care; management information systems; algorithm; machine learning; model; predict; risk

Introduction

Overview

The problem of emergency department (ED) crowding is well known in health care as a complex, multi-dimensional problem that threatens patient safety and care quality and has remained largely unresolved for over 20 years. Despite ED efficiency interventions [1,2] and government policy [3] aimed at reducing crowding, it continues to threaten patient safety and contribute to poor patient outcomes [4-6]. ED crowding occurs when ED demand exceeds the staff's ability to provide quality care in a reasonable time frame [7,8]. The main causes of crowding are ED "boarding" [9-12] (eg, when an ED bed is occupied by a patient due to be admitted to the hospital, but the patient remains in the ED because no inpatient bed has been assigned) and hospital "exit block" [6] (eg, when patients are delayed or blocked from transitioning out of the ED and into the hospital in a reasonable time frame).

Although some recent literature has attributed ED boarding to insufficient hospital bed capacity [13-17], this description of the situation belies boarding's complex roots and suggests that hospitals simply do not have inpatient beds available because they are all occupied by patients. In fact, this is rarely the case, as occupancy rates in most US hospitals average 40% for rural hospitals and 65% for urban hospitals [18-20]; these rates have been slowly declining for decades [18-21]. Instead, insufficient bed capacity in most hospitals refers to a shortage of available beds for ED admission. Reasons for this "shortage" include existing bed reservations, which can be for elective surgery patients who might require admission [9,10], for transfer patients from other hospitals, and for geographic bed plans that assign beds to specialties (eg, orthopedics) to keep relevant patients and providers close together [14]. There are positive logistical and care-quality reasons for these bed reservations, but there are also financial reasons that may benefit the hospital yet contribute to ED boarding. For example, reserving beds for highly reimbursable elective procedures that might not be utilized [9], instead of opening the beds for the immediate needs of ED admissions, increases boarding.

Securing hospital beds for ED patients is a time-intensive, interdepartmental negotiation requiring multiple approvals before an ED patient can be transitioned into an inpatient hospital bed. Larger hospitals have bed managers dedicated to effectively utilizing each hospital bed and the patient support services each requires. Much like air traffic controllers, bed managers are the conductors of a complex series of interdependent processes and activities. Bed management involves assessing bed availability throughout the hospital, assessing whether unit resources are in place to enable a particular bed to be filled by a particular patient, identifying additional unit resources that are required to fill a particular bed, determining whether sufficient resources are available to care for specialty patients (eg, cardiac patients) in a general medicine unit, identifying the required resources and available staff (eg, who is at the hospital and who is on call), identifying which beds are reserved for urgent postoperation surgical cases and which are reserved for elective surgeries, and possessing knowledge of matters spanning multiple departments with a multitude of players. This complex and important process may be unnecessarily convoluted in hospitals that have grown in size and responsibility and have become incongruent with effective organizational management. Examinations of clinical workflows for admitting hospital patients from the ED have revealed processes layered with cultural and organizational factors that exacerbate an already inefficient process. There can be 50 to 75 steps between a bed request and time to admission orders, and staff have reported they believe the process is excessively complex, redundant, and in some respects, unsafe [22]. The earlier bed managers have information about a patient who will likely be admitted, the earlier they can begin the bed assignment process, and the earlier the patient can move out of the ED to a hospital bed. This often dysfunctional and protracted hospital transition and bed assignment process blocks patients from transitioning out of the ED to inpatient hospital care (ie, exit block), resulting in the patient waiting long periods of time in the ED for an inpatient hospital bed assignment (ie, boarding).

Boarding negatively affects hospital operations, causing resource strain due to boarded patients' continued consumption of nurse and physician resources. It precludes the ability to see more patients, because the boarder is occupying an ED bed when an ED level of care is likely not needed [23]. This strain results in a ripple effect throughout the ED that limits all patients' access to timely emergency care [24] and further impacts the emergency medical services system by increasing ambulance diversion and patient offload time (the time paramedics spend waiting for an ED bed to become available, after which they are able to return to service) [25]. Reduction or removal of the exit block that causes boarding would dramatically reduce the duration of patient boarding. Thus, removal of the 2 main causes of ED crowding (ie, exit block and boarding) would greatly reduce crowding and increase access to care.

Easing ED crowding requires a multifaceted approach. That is to say that there is not one single solution, but rather multiple solutions applied at various points of care that hold promise in easing ED crowding. This paper will report on the use of biomarkers, measured very early in the continuum of care, as a mechanism to predict admission, thereby enabling hospitals to initiate the complex and time-consuming bed management process early and reduce ED crowding.

Background and Rationale

Prior Interventions to Address Boarding and Crowding

Hospitals have implemented a variety of hospital-level and ED-level interventions to reduce boarding and crowding. In terms of hospital-level interventions, some hospitals have reportedly reduced boarding by improving inpatient bed availability, for example by shortening patient stays through better management of hospital services that are not continuously available (eg, catheterizations [26,27]), moving discharged patients awaiting transportation or nonacute care services to "discharge lounges" [28,29], managing discharges in a more timely manner and expediting discharges [30-32], and managing bed cleaning turnaround more efficiently [33-35]. Because ED crowding does not have a singular cause, it also does not have a singular solution. As such, these measures may only address part of the issue; they do not address boarding's root cause—exit block. Rather, these efficiency measures contribute to process improvement and operational efficiency, because they successfully and sensibly increase hospital bed availability and streamline and simplify processes for providers and administrators, saving them time and improving flow. Thus, they have the potential to reduce boarding and consequent crowding. Hospitals could benefit from early intelligence about demand for beds, and the need to ensure that the right types of beds are available for ED patients.

ED-level interventions have primarily focused on reducing crowding through improvements in ED flow and throughput, such as by using fast tracking [36-38], split-flow processing [39-41], rapid assessment zones [42-44], team triage [45], triage nurse ordering [46], triage standing orders [47], bedside registration [48], physician scribes [49-51], ED flow coordinators [52], point-of-care testing [53-56], and physical expansion of the ED [57]. While some of these measures may make positive contributions to ED efficiency and flow, they are not unlike the hospital-level interventions, which contribute to solving parts of the problem but do not address exit block [58]. Instead, these measures primarily promote efficiency in subsections of the ED care continuum and move patients more quickly toward upstream bottlenecks in the ED process. Even a dedicated ED flow coordinator who successfully increases flow throughout the ED will see much of that benefit lost if improvements are not also implemented outside the ED [52].

Interdepartmental Interventions to Address Exit Block

Interventions that have included interdepartmental collaboration with hospital management support have made positive steps toward reducing exit block. Such interventions have resulted in a 68% reduction in the time from inpatient bed requests to receipt of inpatient admission orders (210 minutes to 75 minutes; this does not mean a bed has been assigned, only that the order has been created) and a 25% reduction in the time from inpatient bed requests to patient departure from the ED (360 minutes to 270 minutes) [22]. These interventions are primarily viewed through a process improvement lens, in terms of bed management strategy. For example, a study by Barrett et al [59] reported a 52% reduction in "hold time" (the time from admission decision to departure from the ED) when full-time bed managers were able to identify and assign patients to beds within 15 minutes of a bed request. Another such study, by Howell et al [60], reported a 90-minute reduction between ED patient registration and patient physical departure from the ED for admitted patients when a dedicated "bed traffic controller" was used. The difference between the Barrett et al [59] study and the Howell et al [60] study is that the former employed resources at the micro or patient level, whereas the latter employed resources at the macro or process level, presumably with a top-down view of "traffic."

The process improvements and bed management strategies reported by Barrett et al [59] and Howell et al [60] also demonstrated how real-time ED data on congestion, flow, and patient admissions can be used by hospital staff outside of the ED, such as the hospital bed manager, to prepare for and manage admissions and bed demand. With reliable information to predict the likeliness of being admitted, effective bed management strategies could be deployed earlier in the admission continuum cycle.

Predictive Modeling in Health Care

A variety of models have been used to estimate the risk of hospital admission from the ED, including logistic regression and machine learning, and a variety of predictor variables have been used, including the primary complaint, prior ED visits, referral source, medical history, and mode of arrival. However, model reliance on information that is not readily available (such as patient records) or is inaccurate (such as patient reports of medical history) can be problematic for the application and operation of hospital-admission prediction models. Immediately available point-of-care information from patient biomarkers, such as age, gender, vital signs, and acuity level, offer an advantage over previously collected information [61].

Currently, there is no benchmark to compare hospital admission prediction models. This was evidenced by the authors' previous systematic review and critical assessment study of models predicting hospital admission, which found that all had potential biases [61].

Biomarker Indicators of Admission

The word "biomarker," short for "biological marker," refers to a broad category of objective indicators of medical state that can be measured accurately and reproducibly [62]. Examples of biomarkers are age, x-ray images, vital signs, genes, alleles, gender, cognitive state, and acuity level. Vital signs are the most essential biomarker for monitoring hospitalized patients and are the simplest, least expensive, most readily available, and probably the most important information gathered on patients [63]. They are especially useful in the ED environment, which is populated by patients with a variety of symptoms and conditions, challenging care providers to assess patients quickly. Failure to recognize patient severity or acuity can be detrimental or fatal in the ED. Vital signs that are assessed in real time provide an opportunity to avert this risk to patients, because changes in vital signs have been shown to occur several hours before serious adverse events [64-68]. As such, vital signs can be used to identify ED patients at risk of deterioration [67-72].

The purpose of this study was to report on the development of a model that used patient biomarkers collected at triage (the 5

vital signs and age, gender, and acuity level) for the early prediction of the risk of imminent hospital admission or transfer from the ED for adult patients.

Methods

Study Design

This retrospective cohort study evaluated 1 year of consecutive data events for adult patients admitted to the ED and developed an algorithm to predict which patients would require imminent hospital admission. Eight variables collected at triage were evaluated for their role in the outcome of the patient ED visit. Logistic regression was used to model the study data.

Study Setting, Data Source, and Population

The sample population of deidentified data was drawn from 1 year (January 1, 2019, through December 31, 2019) of consecutive ED admissions to an academic medical center and were queried from its Informatics for Integrating Biology to the Bedside (i2b2) database [73], part of the National Institutes of Health–funded National Centers for Biomedical Computing. Transfer patients (ie, ED patients requiring inpatient hospital admission who were transferred to other hospitals for clinical reasons, such as to receive specialty care) were grouped with admitted patients, because their clinical presentation and reasons for transfer to other facilities for inpatient admission were clinically identical to those of admitted patients [74]. The academic medical center ED is a 48-bed, level-1 adult trauma center with an average daily ED census of 300 patients. The hospital has 1157 beds enterprise-wide.

All clinical data were collected at triage by nurses and entered into the electronic medical record at the point of care. Standardized methods were used to collect vital signs and acuity level.

Inclusion and Exclusion Criteria

All adult (ie, age ≥18 years), nonpsychiatric, nonobstetric, fully triaged patients admitted to the ED (including those transferred to other hospitals for admission) and subsequently admitted to the hospital or discharged from the ED were included in this study. Psychiatric, pediatric (ie, age <18 years), and obstetric patients were excluded. Psychiatric and obstetric patients were excluded because these populations' symptomology, and the clinical variables that are evaluated to determine their course of treatment, are significantly clinically different from the general-medicine population [75,76]. Pediatric patients were excluded because the threshold for admission for these patients is lower than for adults [77], and their inclusion would result in overly sensitive inclusion criteria for adults.

Selection of Variables for Measurement

The choice of variables evaluated (Table 1) for model development was derived from a systematic review of studies evaluating models designed to predict hospital admission, which suggested variables that were most valuable for patient admission or discharge [61]. Predictors were selected a priori by expert knowledge. Although the literature suggests that SpO_2 [78], level of consciousness [79], and mode of arrival [78,80-85] are important variables for consideration [61], the available data were too inconsistent and, therefore, were not included in this model.

Table 1. Eight variables were analyzed for their utility in predicting hospital admission and discharge.

Variables	Means of collection
Predictor variables	
Age	Provided by patient (or family or friend if patient was unable to report)
Acuity	The standardized 5-level Emergency Severity Index [86] was used by the triage nurse to categorize patient acuity from most urgent (level 1) to least urgent (level 5)
Systolic blood pressure; diastolic blood pressure; heart rate, respiration rate; temperature	Typical, standardized methods were used to collect vital signs; blood pressure was the only variable collected by 2 methods: manual (the primary method) and automated
Gender	Provided by patient (or family or friend if patient was unable to report); if the patient was unaccompanied, clinicians determined gender by visual inspection
Outcome variable	
Admitted or discharged	Determined by physician

Study Protocol and Data Management

The data were exported from i2b2, imported into an Excel table for review and cleaning, then exported to the Stata statistical package (version 14.1; StataCorp) for analysis. The data were evaluated for missing values.

Data Analysis and Model Development

Data distribution was investigated with summary statistics and histograms. We examined the univariate associations of age, systolic blood pressure (BP), diastolic BP, heart rate, acuity, and gender with the probability of admission using a logistic regression. Traditional logistic regression assumes that the association between continuous risk factors and the probability of admission is linear on a log-odds scale. We considered a more flexible model in which continuous risk factors were included as fractional polynomials, a model-building technique that allows for nonlinear associations [87], and temperature, respiration, and acuity were included as categorical risk factors.

We performed a multivariable fractional polynomial (MFP) analysis that included all risk factors. Temperature values were tightly clustered, with 97% of values between 36.1 °C and 37.8 °C, with a wide spread in values above and below these values. For the purposes of modeling, we created a modified

temperature variable where values less than 36.1 °C or greater than 37.8 °C were truncated. Fractional polynomials of the modified temperature were combined with dummy variables indicating high (>37.8 °C) and low (<36.1 °C) temperatures. Respiration was included as a categorical variable due to difficulty in modeling the association between respiration as a continuous variable and the probability of admission. The MFP analysis included nonlinear relationships if they were sufficiently supported by the data. The fit of a second order fractional polynomial was compared to that of the null model, the linear model, and finally to the optimal first-order polynomial. Convergence was achieved when the functional forms did not change. The significance level for the comparison of fractional polynomial models was set equal to 0.01. As some subjects visited the ED more than once, we considered a robust MFP analysis that allowed for correlation between repeat observations of the same subject. Descriptive statistics considered each patient visit as unique. Patient numbers refer to the number of ED encounters.

Model performance was assessed by discrimination and calibration. Discrimination, the model's ability to accurately distinguish between admission and nonadmission [88], was measured with the area under the receiver operating characteristics curve (AUROC) [89]. To assess potential overoptimism, we also calculated a 10-fold cross-validated AUROC. Calibration, the extent to which the model-predicted probabilities agree with observed binary outcomes [90], is a more appropriate gauge of model performance [91] and was measured by Hosmer-Lemeshow goodness of fit and evaluated graphically using a "calibration belt" [92] for internal validation. The calibration belt methodology formulated the relationship between the predictions and the true probabilities of admission with a second logit regression model based on a polynomial transformation of the predictions. The degree of the polynomial was forwardly selected, beginning with the second order on the basis of a sequence of likelihood-ratio tests [91].

The model was designed to be hospital-specific with application to a particular ED population. As such, we did not measure external validity. Risk factors were evaluated for extreme values, resulting in the loss of less than 2% of patient events overall.

Ethical Considerations

Ethical approval to conduct the investigation was obtained from The University of Alabama at Birmingham Institutional Review Board (IRB-300007437). This study was conducted on a data set that was void of any protected health information.

Results

Descriptive Data

The population consisted of 93,847 adults (age ≥18 years) who were fully triaged, general-medicine (ie, nonpsychiatric and nonobstetric) patients admitted to the ED from January 1, 2019, through December 31, 2019, and subsequently discharged from the ED or admitted to the hospital. The mean age of the 93,847 patients was 46.3 years; 55.6% (52,147) were female; 56.4% (52,974) had acuity level 3; mean systolic BP was 139 mmHg; mean diastolic BP was 84 mmHg; mean heart rate was 87.2 beats/minute; mean respiration rate was 17.7 breaths/minute; and mean temperature was 36.8 °C (Table 2). Temperature and respiration rate had long-tailed, tightly clustered distributions. Temperature ranged from 27 °C to 40.3 °C with only 1% (938) of values less than 36 °C and 1% (938) greater than 38.4 °C. Respiration rate was recorded as a whole number and was clustered at even numbers, with 26% (24,400), 40% (37,539), and 12% (11,262) of subjects having respiration rates of 16, 18, and 20 breaths per minute, respectively, ranging from 10 to 40, with 1% (938) of values less than 14 and 1% (938) greater than 26. Compared to those not admitted, admitted patients were more likely to be male; be older; have lower systolic BP and lower diastolic BP; have higher heart rate, respiration rate, and temperature; and be more acute, as indicated by a lower Emergency Severity Index (ESI) [86] level. This index has a scale of 1 to 5, with 1 being most urgent and 5 being least urgent. Of those admitted, 45% (5779/12,711) had acuity scores less than 3, compared to only 8% (6426/81,136) of those not admitted.

Table 2. Values for predictor variables by admission status.

Variables	Not admitted (N=81,136)	Admitted (N=12,711)	Total (N=93,847)
Age (years), mean (SD)	44.8 (17.4)	55.8 (16.9)	46.3 (17.3)
Gender, n (%)			
Male	35,169 (43.3)	6531 (51.4)	41,700 (44.4)
Female	45,967 (56.7)	6180 (48.6)	52,147 (55.6)
Systolic blood pressure (mm Hg), mean (SD)	139.1 (23.2)	137.9 (28.5)	139.0 (50)
Diastolic blood pressure (mm Hg), mean (SD)	84.4 (13.5)	81.2 (16)	84. (13.9)
Heart rate (beats/minute), mean (SD)	86.3 (15.5)	92.9 (18.6)	87.2 (16.1)
Respiration rate (breaths/minute), mean (SD)	17.6 (1.8)	18.6 (3.1)	17.7 (2.1)
Temperature (°C), mean (SD)	36.8 (0.4)	36.8 (0.6)	36.8 (0.4)
Emergency Severity Index level, n (%)[a]			
1	23 (0)	571 (4.5)	594 (0.6)
2	6403 (7.9)	5208 (41)	11,611 (12.4)
3	46,454 (57.3)	6520 (51.3)	52,974 (56.4)
4	26,128 (32.2)	385 (3)	26,513 (28.3)
5	2128 (2.6)	27 (0.2)	2155 (2.3)

[a]Ranges from most urgent (1) to least urgent (5).

Probability of Admission

Male admission rates were higher, at 15.7% (6531/41,700) compared to 11.9% (6180/52,147) for women. There was a strong association between acuity and admission, with the probability of admission falling sharply from 96.1% (571/594) for the most urgent patients (ESI 1) to 1.1% (23/2155) for the least urgent patients (ESI 5) (Table 3). Figure 1 shows variable distributions and their associations with the probability of admission, shown on a logit scale. Results are based on second order fractional polynomials. There was a clear, nonlinear association for all continuous variables except age. The probability of admission increased until age 80 and then leveled off. For systolic BP, diastolic BP, heart rate (Figure 1), respiration rate, and temperature (Figure 2), the association was nonlinear, with the probability of admission lowest at the center of the distribution and higher at the extremes. For example, for systolic BP, the probability of admission was lowest, at 15%, for values between 120 mm Hg and 150 mm Hg, and the probability of admission continued to rise at values outside this range, reaching 95% for values <75 mm Hg and a probability of 40% at values >235 mm Hg. Similarly, at a temperature of 36.7 °C, the probability of admission was lowest, at 17%, increasing to 50% at 39.4 °C and to 80% at 35 °C.

The MFP logistic regression showed significant associations with acuity, respiration, and gender, first-order fractional polynomials for age, and second-order fractional polynomials for systolic BP, diastolic BP, temperature, and heart rate.

Table 3. Probability of hospital admission by acuity level. The most acute patients are at Emergency Severity Index levels 1 and 2, with the least urgent at level 5. A total of 12,711 of 93,847 (13.5%) patients were admitted.

Emergency Severity Index level	Total patients, n	Admitted patients, n	Probability of admission, %
1	594	571	96.1
2	11,611	5208	44.9
3	52,974	6520	12.3
4	26,513	385	1.5
5	2155	27	1.1

Figure 1. Histograms of variable distributions overlaid with graphs showing the probability of admission on a log-odds scale. The shaded areas around the curves represent the 95% CI. BP: blood pressure.

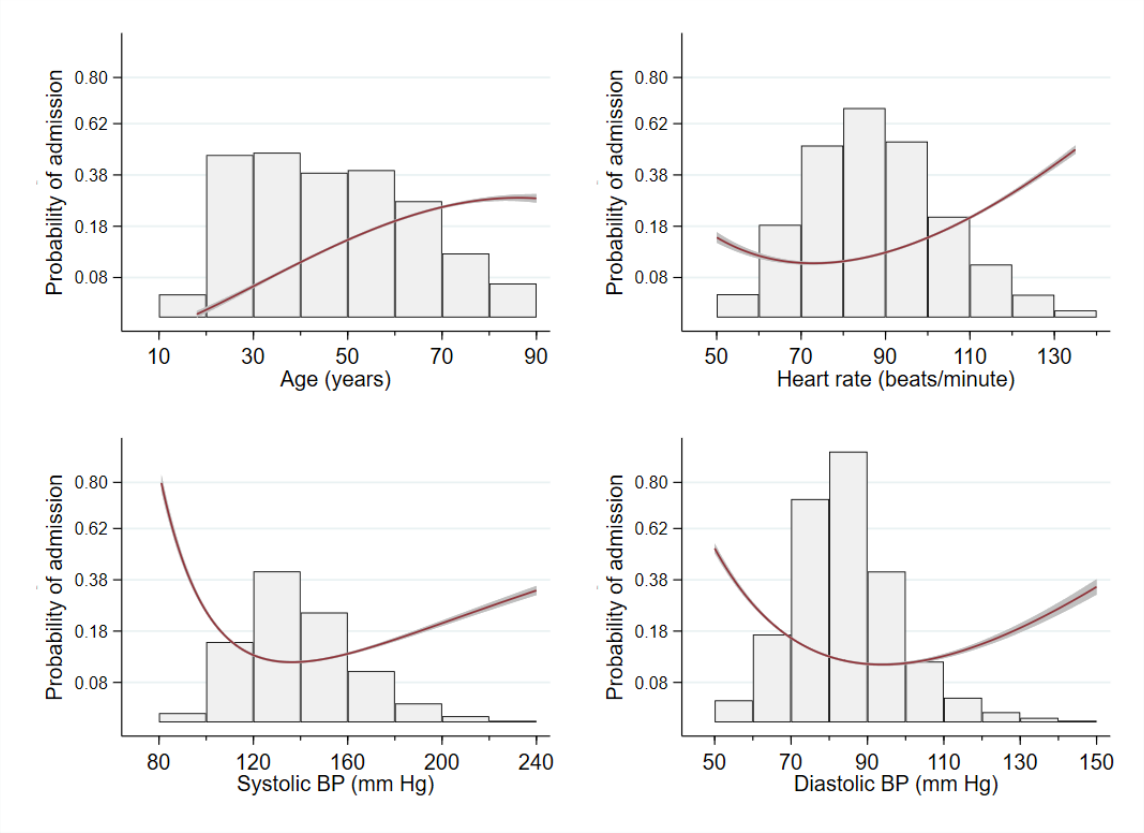
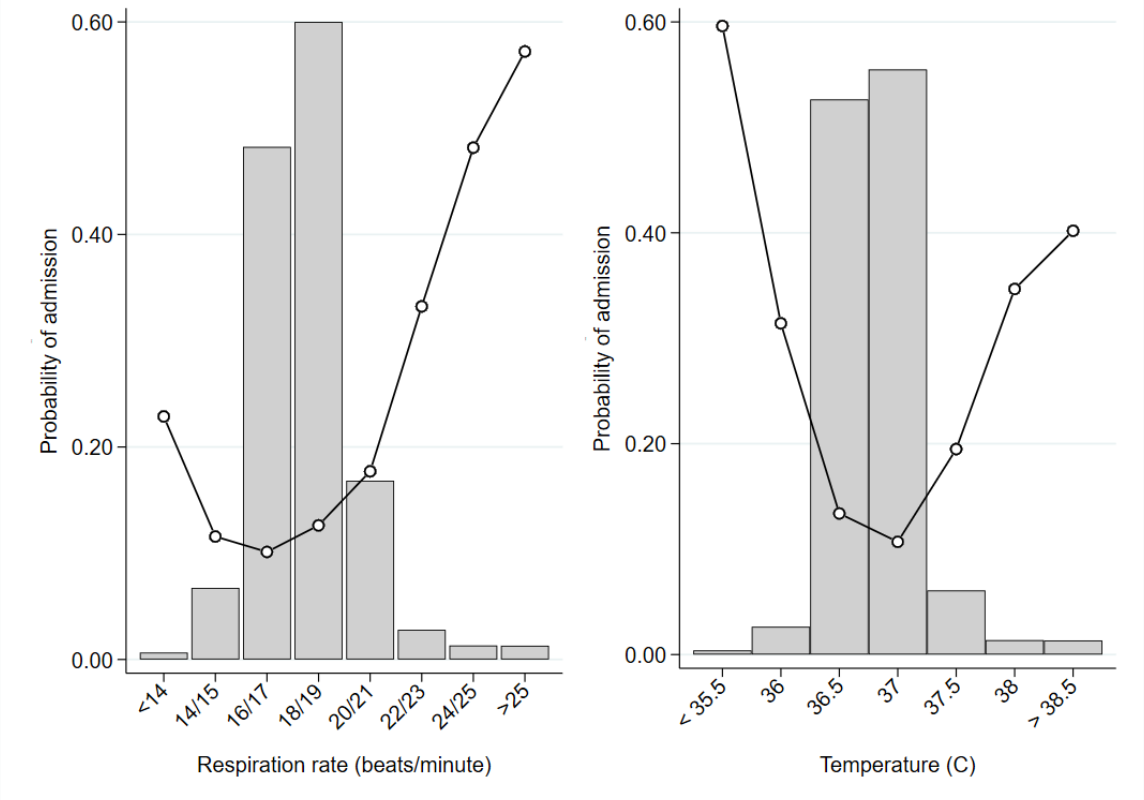


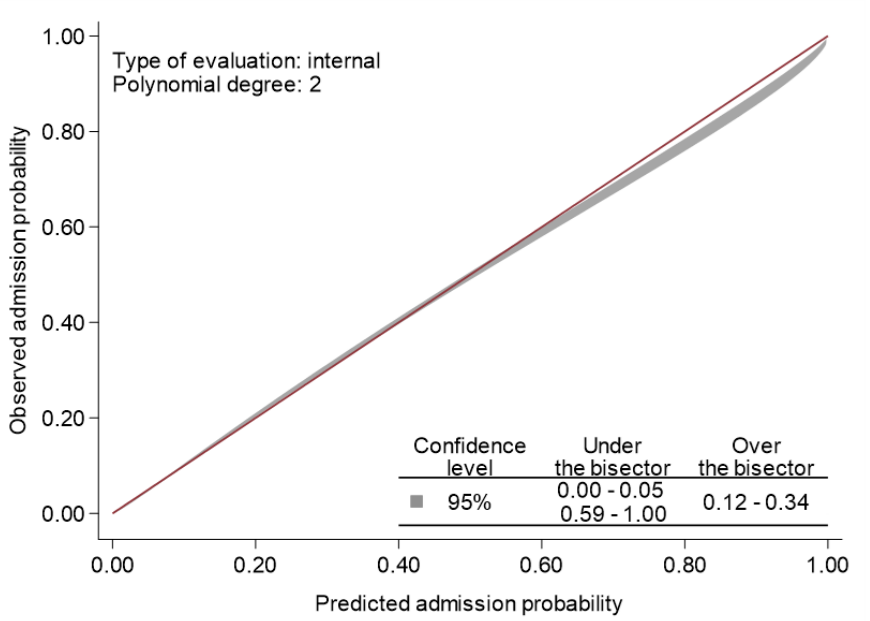
Figure 2. Histograms of variable distributions overlaid with graphs showing the observed admission rates.



Fit and Calibration

In Figure 3, the observed admission rate is plotted against the predicted admission rate based on the results of our final logistic regression model. There was excellent agreement between observed probabilities and predicted probabilities based on the model. The 95% CI fell below the identity line at the high end, which indicated that the model slightly overpredicted risk for patients who had a probability of admission over 0.59. This difference in probabilities was less than 0.04. For admission probabilities less than 0.59, the bias was less than 0.01.

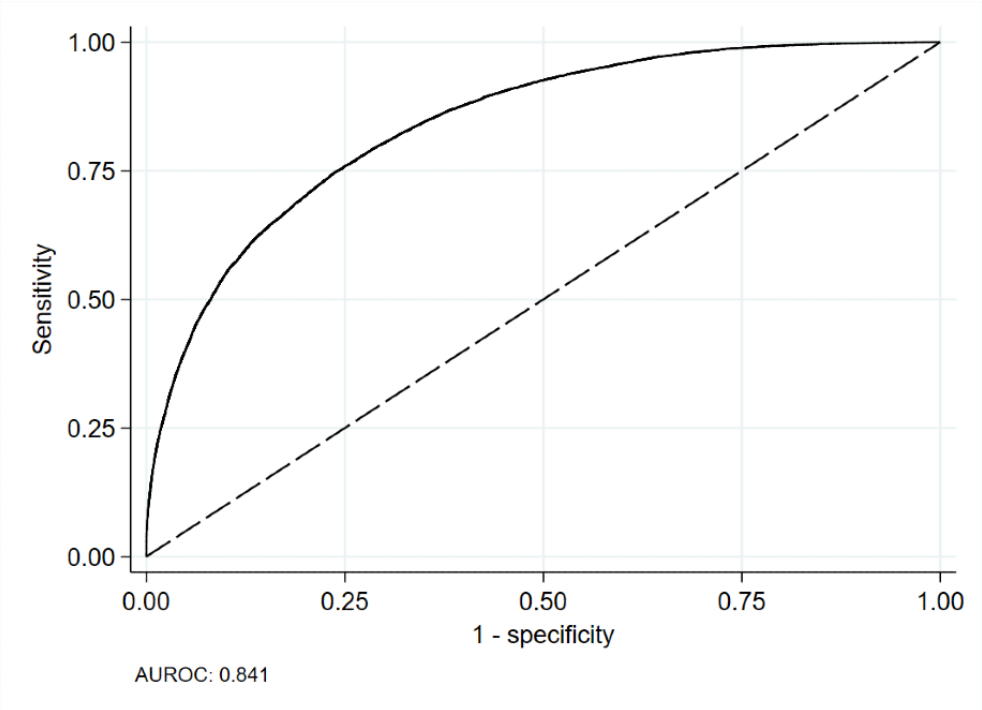
Figure 3. Calibration belt of observed versus predicted admission probabilities. The bisector is the line of perfect calibration. The calibration belt (shown in gray) represents the 95% confidence level calibration of the model.



Model Discrimination

Model discrimination was measured with the AUROC (Figure 4). The AUROC was 0.841, indicating that the model had good ability to discriminate between patients who would and would not be admitted [93]. The 10-fold cross-validated AUROCs ranged from 0.839 to 0.842.

Figure 4. Area under the receiver operating characteristics curve for predicting admission. AUROC: Area under the receiver operating characteristics curve.



Missing Values

Age and gender values were complete for all 113,739 patients. Missing values for all other variables were very low, ranging from 1.5% to 1.9% (1650 to 2143), except for temperature and acuity, which were missing for 6.8% (7685) and 9.2% (10,419) of patients, respectively. As 83% (8,647/10,419) of cases missing acuity had been admitted, this group was identified as "missing not at random." We found no evidence to suggest that temperature was not "missing at random," and the degree of missingness was low enough that we did not expect it to bias results. Cases with missing values were excluded.

Discussion

Principal Findings

We have illustrated the application of sophisticated fractional polynomials to identify and model nonlinear associations between risk factors and the probability of admission using immediately available patient biomarkers (age, systolic BP, diastolic BP, heart rate, respiration rate, temperature, gender, and acuity level) collected at triage. The resulting prediction model exhibited excellent calibration with good agreement between observed and predicted admissions at all risk-of-admission levels. The model showed good ability to discriminate between patients who would and would not be admitted. Methodological techniques promoted internal validity and mitigated against overfitting and endogeneity, which can arise when predictor variables are correlated with the outcome due to their relationship with variables not in the model [94]. Given our large sample size, a priori inclusion of risk factors, and predictor selection based on topic knowledge, the risk of variable omissions was reduced, and the risk of overfitting due to the use of optimal fractional polynomials was not a concern. A 10-fold cross-validation yielded almost identical AUROC values.

Categorization of respiration rate and truncation of temperature could cause loss of relevant information [95]. However, our transformations were informed by the data: our cut-off points for respiration rate reflected how it was recorded (ie, responses were usually 1 of 3 values and showed a preference for even numbers), and although temperature was truncated, less than 4% (3753/93,847) of observations were affected, indicator variables for low and high temperature were included, and temperature was included as a continuous variable.

Anecdotal information suggests that in the practice of a busy ED, failure to record information deemed nonessential can occur when a patient is already scheduled for hospital admission, and this is most likely to occur when the patient is receiving life-saving care. These cases are very acute and are likely to be admitted. The patients with a missing acuity level tended to be very acute (ie, most cases with missing acuity were admitted) and we were comfortable excluding them, because they were not the patient group that the model aimed to identify as requiring admission; these patients likely had already been identified as urgently needing care and were already likely to be admitted. This is a site-specific model designed to operate in a test environment and show proof of concept; generalizability is not assumed. Because this model uses standardized biomarker data and not data specific to the study environment, it is possible that with recalibration, this model could be useful outside the study environment. It is worth noting, however, that in addition to the real-time availability of electronic patient data, a requirement of model implementation is an application to retrieve the data and apply it to the model algorithm to produce a patient's likelihood of admission, then provide the information to bed managers to begin securing patient beds early.

This model, as proposed, has real-world utility for those involved in the patient admission continuum, because it allows patients to be moved out of the ED sooner, thereby easing exit block and benefiting patient care and hospital operations [59,60]. The model's reliance on biomarkers that are routinely collected at the initial point of care (ie, ED triage) and have standardized definitions, measurements, and interpretations [62] is advantageous for a model that can be used very early in the patient care continuum. That, however, does not imply that model development and implementation within a setting is easy. Rather, the data coming from electronic medical records may require labor-intensive preparation to make it suitable for model development and implementation.

This model also showed that a hospital can develop a system for identification of patients at high risk of admission for use in resolving problems such as exit block. The model can be adapted to other ED environments using each ED's individual data.

Comparison With Prior Work

Addressing ED overcrowding and exit block cannot be accomplished by applying a one-size-fits-all solution. The most recent prior work in this area centers around different methods and models for addressing the same problem—ED crowding. For example, Acuna et al [96] optimized ED crowding by creating an ambulance allocation model that led to a 31% improvement in ED crowding, and Isfahani et al [97] used a computer simulation model to assess the effect of ED discharge lounges, finding there was a 5% reduction in admission waiting times. In terms of applying algorithms, Brink et al [98] developed an 8-variable model to predict hospital admission for elderly patients and Marcusson et al [99] developed a 38-variable model to predict hospital admission for elderly patients; both models aimed at helping patients receive care sooner.

Limitations

The applicability of this study should be understood in the context of its limitations. As mentioned earlier, this study was performed as a proof of concept in a large academic medical center and may lack generalizability to other environments. If this model were applied in a setting where complex processes to secure inpatient beds are not undertaken by the hospital (involving, for example, providers, equipment, and other specialized resources for different patient conditions), or where securing beds does not require a large amount of time, then the time-saving advantages of this model would not be realized by bed managers. Additionally, there may have been confounding factors that were mediating or moderating factors in our model and outside the scope of this study. Lastly, while exit block and

ED boarding have been reported internationally, this study was conducted in a US hospital, and is not internationally generalizable. Perhaps a similar, recalibrated version of our derived model would have applications in divergent ED settings. However, we recommend that other hospitals develop hospital-specific models using the MFP modeling techniques presented here.

Conclusion

This primary data study illustrates the application of a site-specific risk prediction model to reduce ED crowding due to exit block. MFPs were used to predict the probability of admission based on 8 biomarkers (5 vital signs and age, gender, and acuity level) and to generate variables utilized by the logistic regression model to produce a site-specific formula to analyze future input data. This intervention can reduce ED exit block, the known source of ED boarding and crowding, by enabling the hospital to seek and requisition hospital beds earlier and transition ED patients into those beds earlier. This intervention requires interdepartmental collaboration with the support of hospital management to be successfully implemented into hospital structures and processes. Compared to other interventions in the hospital admission and bed assignment process that have successfully reduced crowding [22,59,60], this model goes a step further by looking ahead to predict which patients will be admitted, thereby providing the needed information to initiate admission and bed assignment processes much earlier in the care continuum. The model's prediction of patient admissions combined with the utility of real-time hospital data to improve congestion, flow, and patient admissions [59,60] results in a powerful tool to impact the ED crowding crisis.

Conflicts of Interest

None declared.

References

1. Stead LG, Decker WW. The International Journal of Emergency Medicine: successes of the first year. Int J Emerg Med 2009 Apr 15;2(1):1-2 [FREE Full text] [doi: 10.1007/s12245-009-0102-2]
2. Kauppila T, Seppänen K, Mattila J, Kaartinen J. The effect on the patient flow in a local health care after implementing reverse triage in a primary care emergency department: a longitudinal follow-up study. Scand J Prim Health Care 2017 Jun 08;35(2):214-220 [FREE Full text] [doi: 10.1080/02813432.2017.1333320] [Medline: 28593802]
3. Hospital Compare. Centers for Medicare and Medicaid Services. URL: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/HospitalCompare [accessed 2022-02-05]
4. Carter EJ, Pouch SM, Larson EL. The relationship between emergency department crowding and patient outcomes: a systematic review. J Nurs Scholarsh 2014 Mar;46(2):106-115 [FREE Full text] [doi: 10.1111/jnu.12055] [Medline: 24354886]
5. Reznek MA, Murray E, Youngren MN, Durham NT, Michael SS. Door-to-Imaging Time for Acute Stroke Patients Is Adversely Affected by Emergency Department Crowding. Stroke 2017 Jan;48(1):49-54. [doi: 10.1161/STROKEAHA.116.015131] [Medline: 27856953]
6. Richards JR, van der Linden MC, Derlet RW. Providing Care in Emergency Department Hallways: Demands, Dangers, and Deaths. Advances in Emergency Medicine 2014 Dec 25;2014:1-7 [FREE Full text] [doi: 10.1155/2014/495219]
7. Sinclair D. Emergency department overcrowding - implications for paediatric emergency medicine. Paediatr Child Health 2007 Jul;12(6):491-494 [FREE Full text] [doi: 10.1093/pch/12.6.491] [Medline: 19030415]
8. American College of Emergency Physicians (ACEP). Crowding. Policy statement. Ann Emerg Med 2013 Jun;61(6):726-727. [doi: 10.1016/j.annemergmed.2013.03.037] [Medline: 23684339]
9. Hospital emergency departments: Crowding continues to occur, and some patients wait longer than recommended time frames. Government Accountability Office. 2009. URL: https://www.gao.gov/products/gao-09-347 [accessed 2022-07-18]
10. Institute of Medicine. Hospital-Based Emergency Care: At the Breaking Point. Washington, DC: The National Academies Press; 2007.
11. Higginson I. Emergency department crowding. Emerg Med J 2012 Jun;29(6):437-443. [doi: 10.1136/emermed-2011-200532] [Medline: 22223713]
12. Mason S, Knowles E, Boyle A. Exit block in emergency departments: a rapid evidence review. Emerg Med J 2017 Jan;34(1):46-51. [doi: 10.1136/emermed-2015-205201] [Medline: 27789568]
13. Cowan RM, Trzeciak S. Clinical review: Emergency department overcrowding and the potential impact on the critically ill. Crit Care 2005 Jun;9(3):291-295 [FREE Full text] [doi: 10.1186/cc2981] [Medline: 15987383]
14. Rabin E, Kocher K, McClelland M, Pines J, Hwang U, Rathlev N, et al. Solutions to emergency department 'boarding' and crowding are underused and may need to be legislated. Health Aff (Millwood) 2012 Aug;31(8):1757-1766. [doi: 10.1377/hlthaff.2011.0786] [Medline: 22869654]
15. Derlet RW, Richards JR. Ten solutions for emergency department crowding. West J Emerg Med 2008 Jan;9(1):24-27 [FREE Full text] [Medline: 19561699]
16. Handel DA, Hilton JA, Ward MJ, Rabin E, Zwemer FL, Pines JM. Emergency department throughput, crowding, and financial outcomes for hospitals. Acad Emerg Med 2010 Aug;17(8):840-847 [FREE Full text] [doi: 10.1111/j.1553-2712.2010.00814.x] [Medline: 20670321]

17. Rutherford PA, Kotagal U, Luther K, Provost L, Ryckman F, Taylor J. Achieving Hospital-wide Patient Flow. Boston, MA: Institute for Healthcare Improvement; 2020.
18. AHA's Annual Survey of Hospitals 2016. American Hospital Association. 2016. URL: https://www.ahadata.com/aha-annual-survey-database [accessed 2022-02-05]
19. Protect Public Data Hub: Inpatient Bed Dashboard. U.S. Department of Health & Human Services. 2022. URL: https://public-data-hub-dhhs.hub.arcgis.com/pages/Hospital%20Utilization [accessed 2022-02-05]
20. March 2019 Report to the Congress: Medicare Payment Policy. The Medicare Payment Advisory Commission. 2019. URL: https://www.medpac.gov/document/march-2019-report-to-the-congress-medicare-payment-policy/ [accessed 2022-02-05]
21. June 2014 Report to the Congress: Medicare and the Health Care Delivery System. The Medicare Payment Advisory Commission. 2014. URL: https://www.medpac.gov/document/http-www-medpac-gov-docs-default-source-reports-jun14_entirereport-pdf/ [accessed 2021-02-05]
22. Amarasingham R, Swanson TS, Treichler DB, Amarasingham SN, Reed WG. A rapid admission protocol to reduce emergency department boarding times. Qual Saf Health Care 2010 Jun;19(3):200-204. [doi: 10.1136/qshc.2008.031641] [Medline: 20142408]
23. Bukata R. Why is ED holding still an issue? Emergency Physicians Monthly. 2017. URL: http://epmonthly.com/article/why -is-ed-holding-still-an-issue/ [accessed 2019-05-01]
24. Nicks BA, Manthey DM. The impact of psychiatric patient boarding in emergency departments. Emerg Med Int 2012;2012:360308-360305 [FREE Full text] [doi: 10.1155/2012/360308] [Medline: 22888437]
25. Geiderman JM, Marco CA, Moskop JC, Adams J, Derse AR. Ethics of ambulance diversion. Am J Emerg Med 2015 Jun;33(6):822-827. [doi: 10.1016/j.ajem.2014.12.002] [Medline: 25616586]
26. Levin SR, Dittus R, Aronsky D, Weinger MB, Han J, Boord J, et al. Optimizing cardiology capacity to reduce emergency department boarding: a systems engineering approach. Am Heart J 2008 Dec;156(6):1202-1209. [doi: 10.1016/j.ahj.2008.07.007] [Medline: 19033021]
27. Levin S, Dittus R, Aronsky D, Weinger M, France D. Evaluating the effects of increasing surgical volume on emergency department patient access. BMJ Qual Saf 2011 Feb;20(2):146-152. [doi: 10.1136/bmjqs.2008.030007] [Medline: 21209127]
28. Smith K. Implementation of the Discharge Hospitality Center to Reduce Emergency Department Boarding: A Quality Improvement Project. The Free Library. 2018. URL: https://www.thefreelibrary.com/Implementation+of+the+Discharge+Hospitality+Center+to+Reduce...-a0568974204 [accessed 2022-07-18]
29. Franklin BJ, Vakili S, Huckman RS, Hosein S, Falk N, Cheng K, et al. The Inpatient Discharge Lounge as a Potential Mechanism to Mitigate Emergency Department Boarding and Crowding. Ann Emerg Med 2020 Jun;75(6):704-714. [doi: 10.1016/j.annemergmed.2019.12.002] [Medline: 31983501]
30. Johnson M, Sensei L, Capasso V. Improving patient flow through a better discharge process. J Healthc Manag 2012;57(2):89-93. [Medline: 22530290]
31. Powell ES, Khare RK, Venkatesh AK, Van Roo BD, Adams JG, Reinhardt G. The relationship between inpatient discharge timing and emergency department boarding. J Emerg Med 2012 Feb;42(2):186-196. [doi: 10.1016/j.jemermed.2010.06.028] [Medline: 20888163]
32. El-Eid GR, Kaddoum R, Tamim H, Hitti EA. Improving hospital discharge time: a successful implementation of Six Sigma methodology. Medicine (Baltimore) 2015 Mar;94(12):e633 [FREE Full text] [doi: 10.1097/MD.0000000000000633] [Medline: 25816029]
33. Kehoe B. From good to great: 2010 ES department of the year: Doylestown Hospital. Health Facilities Management. 2010. URL: https://www.hfmmagazine.com/articles/1147-from-good-to-great [accessed 2022-08-18]
34. Tortorella F, Ukanowicz D, Douglas-Ntagha P, Ray R, Triller M. Improving bed turnover time with a bed management system. J Nurs Adm 2013 Jan;43(1):37-43. [doi: 10.1097/NNA.0b013e3182785fe7] [Medline: 23232178]
35. VA National Bed Control System. Department of Veterans Affairs. 2021. URL: https://catalog.data.gov/dataset/va-national-bed-control-system [accessed 2022-01-29]
36. Cheney C. Adopt this 5-part process to reduce ER length of stay. Health Leaders. 2019. URL: https://www.healthleadersmedia.com/clinical-care/adopt-5-part-process-reduce-er-length-stay [accessed 2019-05-05]
37. Celona C, Amaranto A, Ferrer R, Wieland M, Abrams S, Obusan F, et al. Interdisciplinary Design to Improve Fast Track in the Emergency Department. Adv Emerg Nurs J 2018;40(3):198-203. [doi: 10.1097/TME.0000000000000199] [Medline: 30059375]
38. Chrusciel J, Fontaine X, Devillard A, Cordonnier A, Kanagaratnam L, Laplanche D, et al. Impact of the implementation of a fast-track on emergency department length of stay and quality of care indicators in the Champagne-Ardenne region: a before-after study. BMJ Open 2019 Jun 19;9(6):e026200 [FREE Full text] [doi: 10.1136/bmjopen-2018-026200] [Medline: 31221873]
39. Bish P, McCormick M, Otegbeye M. Ready-JET-Go: Split Flow Accelerates ED Throughput. J Emerg Nurs 2016 Mar;42(2):114-119. [doi: 10.1016/j.jen.2015.06.003] [Medline: 26264788]
40. Garrett JS, Berry C, Wong H, Qin H, Kline JA. The effect of vertical split-flow patient management on emergency department throughput and efficiency. Am J Emerg Med 2018 Sep;36(9):1581-1584. [doi: 10.1016/j.ajem.2018.01.035] [Medline: 29352674]

41. Wallingford G, Joshi N, Callagy P, Stone J, Brown I, Shen S. Introduction of a Horizontal and Vertical Split Flow Model of Emergency Department Patients as a Response to Overcrowding. J Emerg Nurs 2018 Jul;44(4):345-352. [doi: 10.1016/j.jen.2017.10.017] [Medline: 29169818]
42. Bullard MJ, Villa-Roel C, Guo X, Holroyd BR, Innes G, Schull MJ, et al. The role of a rapid assessment zone/pod on reducing overcrowding in emergency departments: a systematic review. Emerg Med J 2012 May;29(5):372-378. [doi: 10.1136/emj.2010.103598] [Medline: 21515880]
43. Anderson J, Burke R, Augusto K, Beagan B, Rodrigues-Belong M, Frazer L, et al. The Effect of a Rapid Assessment Zone on Emergency Department Operations and Throughput. Ann Emerg Med 2020 Feb;75(2):236-245. [doi: 10.1016/j.annemergmed.2019.07.047] [Medline: 31668573]
44. Chartier L, Josephson T, Bates K, Kuipers M. Improving emergency department flow through Rapid Medical Evaluation unit. BMJ Qual Improv Rep 2015;4(1):u206156.w2663 [FREE Full text] [doi: 10.1136/bmjquality.u206156.w2663] [Medline: 26734447]
45. Ming T, Lai A, Lau P. Can Team Triage Improve Patient Flow in the Emergency Department? A Systematic Review and Meta-Analysis. Adv Emerg Nurs J 2016;38(3):233-250. [doi: 10.1097/TME.0000000000000113] [Medline: 27482995]
46. Rowe BH, Guo X, Villa-Roel C, Schull M, Holroyd B, Bullard M, et al. The role of triage liaison physicians on mitigating overcrowding in emergency departments: a systematic review. Acad Emerg Med 2011 Feb;18(2):111-120 [FREE Full text] [doi: 10.1111/j.1553-2712.2010.00984.x] [Medline: 21314769]
47. Hwang CW, Payton T, Weeks E, Plourde M. Implementing Triage Standing Orders in the Emergency Department Leads to Reduced Physician-to-Disposition Times. Advances in Emergency Medicine 2016 Jun 15;2016:1-6 [FREE Full text] [doi: 10.1155/2016/7213625]
48. McHugh M, Van Dyke KJ, Howell E, Adams F, Moss D, Yonek J. Changes in patient flow among five hospitals participating in a learning collaborative. J Healthc Qual 2013;35(1):21-29. [doi: 10.1111/j.1945-1474.2011.00163.x] [Medline: 22092988]
49. Heaton H, Nestler D, Lohse C, Sadosty A. Impact of scribes on emergency department patient throughput one year after implementation. Am J Emerg Med 2017 Feb;35(2):311-314. [doi: 10.1016/j.ajem.2016.11.017] [Medline: 27856140]
50. Walker K, Ben-Meir M, Dunlop W, Rosler R, West A, O'Connor G, et al. Impact of scribes on emergency medicine doctors' productivity and patient throughput: multicentre randomised trial. BMJ 2019 Jan 30;364:l121 [FREE Full text] [doi: 10.1136/bmj.l121] [Medline: 30700408]
51. Heaton HA, Schwartz EJ, Gifford WJ, Koch KA, Lohse CM, Monroe RJ, et al. Impact of scribes on throughput metrics and billing during an electronic medical record transition. Am J Emerg Med 2020 Aug;38(8):1594-1598. [doi: 10.1016/j.ajem.2019.158433] [Medline: 31522929]
52. Murphy S, Barth B, Carlton E, Gleason M, Cannon C. Does an ED flow coordinator improve patient throughput? J Emerg Nurs 2014 Nov;40(6):605-612. [doi: 10.1016/j.jen.2014.03.007] [Medline: 24974359]
53. Jarvis P, Davies M, Mitchell K, Taylor I, Baker M. Can the Introduction of Point-of-Care Testing for Renal Function in the Emergency Department Reduce Overcrowding? Point Care 2015;14:42-44. [doi: 10.1097/POC.0000000000000048]
54. Kankaanpää M, Raitakari M, Muukkonen L, Gustafsson S, Heitto M, Palomäki A, et al. Use of point-of-care testing and early assessment model reduces length of stay for ambulatory patients in an emergency department. Scand J Trauma Resusc Emerg Med 2016 Oct 18;24(1):125 [FREE Full text] [doi: 10.1186/s13049-016-0319-z] [Medline: 27756354]
55. Li L, McCaughey E, Iles-Mann J, Sargeant A, Westbrook J, Georgiou A. Does Point-of-Care Testing Impact Length of Stay in Emergency Departments (EDs)?: A Before and After Study of 26 Rural and Remote EDs. Stud Health Technol Inform 2018;252:99-104. [Medline: 30040690]
56. Singer AJ, Taylor M, LeBlanc D, Meyers K, Perez K, Thode HC, et al. Early Point-of-Care Testing at Triage Reduces Care Time in Stable Adult Emergency Department Patients. J Emerg Med 2018 Aug;55(2):172-178. [doi: 10.1016/j.jemermed.2018.04.061] [Medline: 29887410]
57. Mumma BE, McCue JY, Li C, Holmes JF. Effects of emergency department expansion on emergency department patient flow. Acad Emerg Med 2014 May 19;21(5):504-509 [FREE Full text] [doi: 10.1111/acem.12366] [Medline: 24842500]
58. Henderson K, Boyle A. Exit block in the emergency department: recognition and consequences. Br J Hosp Med (Lond) 2014 Nov 02;75(11):623-626 [FREE Full text] [doi: 10.12968/hmed.2014.75.11.623] [Medline: 25383431]
59. Barrett L, Ford S, Ward-Smith P. A bed management strategy for overcrowding in the emergency department. Nurs Econ 2012;30(2):82-5, 116. [Medline: 22558725]
60. Howell E, Bessman E, Kravet S, Kolodner K, Marshall R, Wright S. Active bed management by hospitalists and emergency department throughput. Ann Intern Med 2008 Dec 02;149(11):804-811. [doi: 10.7326/0003-4819-149-11-200812020-00006] [Medline: 19047027]
61. Monahan AC, Feldman SS. Models Predicting Hospital Admission of Adult Patients Utilizing Prehospital Data: Systematic Review Using PROBAST and CHARMS. JMIR Med Inform 2021 Sep 16;9(9):e30022 [FREE Full text] [doi: 10.2196/30022] [Medline: 34528893]
62. Strimbu K, Tavel JA. What are biomarkers? Current Opinion in HIV and AIDS 2010;5(6):463-466. [doi: 10.1097/coh.0b013e32833ed177]
63. Kellett J, Sebat F. Make vital signs great again - A call for action. Eur J Intern Med 2017 Nov;45:13-19. [doi: 10.1016/j.ejim.2017.09.018] [Medline: 28941841]

64. Kause J, Smith G, Prytherch D, Parr M, Flabouris A, Hillman K, Intensive Care Society (UK), AustralianNew Zealand Intensive Care Society Clinical Trials Group. A comparison of antecedents to cardiac arrests, deaths and emergency intensive care admissions in Australia and New Zealand, and the United Kingdom--the ACADEMIA study. Resuscitation 2004 Sep;62(3):275-282. [doi: 10.1016/j.resuscitation.2004.05.016] [Medline: 15325446]
65. Buist M, Bernard S, Nguyen TV, Moore G, Anderson J. Association between clinically abnormal observations and subsequent in-hospital mortality: a prospective study. Resuscitation 2004 Aug;62(2):137-141. [doi: 10.1016/j.resuscitation.2004.03.005] [Medline: 15294398]
66. Hillman KM, Bristow PJ, Chey T, Daffurn K, Jacques T, Norman SL, et al. Antecedents to hospital deaths. Intern Med J 2001 Aug;31(6):343-348. [doi: 10.1046/j.1445-5994.2001.00077.x] [Medline: 11529588]
67. Henriksen DP, Brabrand M, Lassen AT. Prognosis and risk factors for deterioration in patients admitted to a medical emergency department. PLoS One 2014;9(4):e94649 [FREE Full text] [doi: 10.1371/journal.pone.0094649] [Medline: 24718637]
68. Barfod C, Lauritzen MMP, Danker JK, Sölétormos G, Forberg JL, Berlac PA, et al. Abnormal vital signs are strong predictors for intensive care unit admission and in-hospital mortality in adults triaged in the emergency department - a prospective cohort study. Scand J Trauma Resusc Emerg Med 2012 Apr 10;20:28 [FREE Full text] [doi: 10.1186/1757-7241-20-28] [Medline: 22490208]
69. Ljunggren M, Castrén M, Nordberg M, Kurland L. The association between vital signs and mortality in a retrospective cohort study of an unselected emergency department population. Scand J Trauma Resusc Emerg Med 2016 Mar 03;24:21 [FREE Full text] [doi: 10.1186/s13049-016-0213-8] [Medline: 26940235]
70. Farrohknia N, Castrén M, Ehrenberg A, Lind L, Oredsson S, Jonsson H, et al. Emergency department triage scales and their components: a systematic review of the scientific evidence. Scand J Trauma Resusc Emerg Med 2011 Jun 30;19:42 [FREE Full text] [doi: 10.1186/1757-7241-19-42] [Medline: 21718476]
71. Pedersen NE, Oestergaard D, Lippert A. End points for validating early warning scores in the context of rapid response systems: a Delphi consensus study. Acta Anaesthesiol Scand 2016 May;60(5):616-622. [doi: 10.1111/aas.12668] [Medline: 26708159]
72. Cardona-Morrell M, Prgomet M, Lake R, Nicholson M, Harrison R, Long J, et al. Vital signs monitoring and nurse-patient interaction: A qualitative observational study of hospital practice. Int J Nurs Stud 2016 Apr;56:9-16. [doi: 10.1016/j.ijnurstu.2015.12.007] [Medline: 26775214]
73. i2b2: Informatics for Integrating Biology and the Bedside. i2b2 tranSMARTFoundation. URL: https://www.i2b2.org/about/index.html [accessed 2021-12-20]
74. EMTALA Fact Sheet. American College of Emergency Physicians. URL: https://www.acep.org/life-as-a-physician/ethics--legal/emtala/emtala-fact-sheet/ [accessed 2022-01-20]
75. Care of the Psychiatric Patient in the Emergency Department – A Review of the Literature. American College of Emergency Physicians Emergency Medicine Practice Committee. 2014. URL: https://tinyurl.com/y7zpxc4a [accessed 2022-07-18]
76. Triage of the Obstetrics Patient in the Emergency Department: Is There Only One Patient? Pennsylvania Patient Safety Advisor. 2008 May. URL: http://patientsafety.pa.gov/ADVISORIES/Pages/200809_85.aspx [accessed 2022-08-18]
77. Paediatric emergency triage, assessment and treatment. World Health Organization. 2016. URL: https://apps.who.int/iris/bitstream/handle/10665/204463/9789241510219_eng.pdf [accessed 2022-02-05]
78. Lucke JA, de Gelder J, Clarijs F, Heringhaus C, de Craen AJM, Fogteloo AJ, et al. Early prediction of hospital admission for emergency department patients: a comparison between patients younger or older than 70 years. Emerg Med J 2018 Jan 16;35(1):18-27. [doi: 10.1136/emermed-2016-205846] [Medline: 28814479]
79. Burch VC, Tarr G, Morroni C. Modified early warning score predicts the need for hospital admission and inhospital mortality. Emerg Med J 2008 Oct 01;25(10):674-678. [doi: 10.1136/emj.2007.057661] [Medline: 18843068]
80. Parker CA, Liu N, Wu SX, Shen Y, Lam SSW, Ong MEH. Predicting hospital admission at the emergency department triage: A novel prediction model. Am J Emerg Med 2019 Aug;37(8):1498-1504. [doi: 10.1016/j.ajem.2018.10.060] [Medline: 30413365]
81. Peck J, Benneyan J, Nightingale D, Gaehde S. Predicting emergency department inpatient admissions to improve same-day patient flow. Acad Emerg Med 2012 Sep;19(9):E1045-E1054 [FREE Full text] [doi: 10.1111/j.1553-2712.2012.01435.x] [Medline: 22978731]
82. Kraaijvanger N, Rijpsma D, Roovers L, van Leeuwen H, Kaasjager K, van den Brand L, et al. Development and validation of an admission prediction tool for emergency departments in the Netherlands. Emerg Med J 2018 Aug 07;35(8):464-470. [doi: 10.1136/emermed-2017-206673] [Medline: 29627769]
83. Cameron A, Rodgers K, Ireland A, Jamdar R, McKay GA. A simple tool to predict admission at the time of triage. Emerg Med J 2015 Mar 13;32(3):174-179 [FREE Full text] [doi: 10.1136/emermed-2013-203200] [Medline: 24421344]
84. Cagle R, Darling E, Kim B. Identifying healthcare activities using a real-time location system. J Med Pract Manage 2014;30(2):128-130. [Medline: 25807605]
85. Fihn SD, Francis J, Clancy C, Nielson C, Nelson K, Rumsfeld J, et al. Insights from advanced analytics at the Veterans Health Administration. Health Aff (Millwood) 2014 Jul;33(7):1203-1211. [doi: 10.1377/hlthaff.2014.0054] [Medline: 25006147]

86. Emergency Severity Index (ESI): A Triage Tool for Emergency Departments. Agency for Healthcare Research Quality. 2020. URL: https://www.ahrq.gov/patient-safety/settings/emergency-dept/esi.html [accessed 2021-10-12]
87. Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. Int J Epidemiol 1999 Oct 01;28(5):964-974. [doi: 10.1093/ije/28.5.964] [Medline: 10597998]
88. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010 Jan;21(1):128-138 [FREE Full text] [doi: 10.1097/EDE.0b013e3181c30fb2] [Medline: 20010215]
89. Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. Biometrics 2005 Mar;61(1):92-105. [doi: 10.1111/j.0006-341X.2005.030814.x] [Medline: 15737082]
90. Waljee A, Higgins P, Singal A. A primer on predictive models. Clin Transl Gastroenterol 2014 Jan 02;5:e44 [FREE Full text] [doi: 10.1038/ctg.2013.19] [Medline: 24384866]
91. Moran JL, Santamaria JD, Duke GJ. Modelling hospital outcome: problems with endogeneity. BMC Med Res Methodol 2021 Jun 21;21(1):124 [FREE Full text] [doi: 10.1186/s12874-021-01251-8] [Medline: 34154530]
92. Nattino G, Lemeshow S, Phillips G, Finazzi S, Bertolini G. Assessing the Calibration of Dichotomous Outcome Models with the Calibration Belt. Stata J 2018 Jan 01;17(4):1003-1014 [FREE Full text] [doi: 10.1177/1536867X1801700414]
93. Tape TG. Interpreting diagnostic tests: The area under an ROC curve. University of Nebraska Medical Center. URL: http://gim.unmc.edu/dxtests/roc3.htm [accessed 2020-02-10]
94. Mennemeyer ST. Can econometrics rescue epidemiology? Ann Epidemiol 1997 May;7(4):249-250. [doi: 10.1016/s1047-2797(97)00021-5] [Medline: 9177106]
95. Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med 2014 Oct;11(10):e1001744 [FREE Full text] [doi: 10.1371/journal.pmed.1001744] [Medline: 25314315]
96. Acuna J, Zayas-Castro J, Charkhgard H. Ambulance allocation optimization model for the overcrowding problem in US emergency departments: A case study in Florida. Socio-Econ Plan Sci 2020 Sep;71:100747 [FREE Full text] [doi: 10.1016/j.seps.2019.100747]
97. Nasr Isfahani M, Davari F, Azizkhani R, Rezvani M. Decreased Emergency Department Overcrowding by Discharge Lounge: A Computer Simulation Study. Int J Prev Med 2020;11:13 [FREE Full text] [doi: 10.4103/ijpvm.IJPVM_582_18] [Medline: 32175053]
98. Brink A, Alsma J, Brink HS, de Gelder J, Lucke JA, Mooijaart SP, et al. Prediction admission in the older population in the Emergency Department: the CLEARED tool. Neth J Med 2020 Dec;78(6):357-367 [FREE Full text] [Medline: 33380533]
99. Marcusson J, Nord M, Dong H, Lyth J. Clinically useful prediction of hospital admissions in an older population. BMC Geriatr 2020 Mar 06;20(1):95 [FREE Full text] [doi: 10.1186/s12877-020-1475-6] [Medline: 32143637]

Abbreviations

AUROC: area under the receiver operating characteristics curve
BP: blood pressure
ED: emergency department
ESI: Emergency Severity Index
MFP: multivariable fractional polynomial

Edited by A Mavragani; submitted 21.04.22; peer-reviewed by X Ma, S Nagavally, K Mortey, H Musawir; comments to author 31.05.22; revised version received 05.07.22; accepted 17.07.22; published 13.09.22.

Please cite as:
Monahan AC, Feldman SS, Fitzgerald TP
Reducing Crowding in Emergency Departments With Early Prediction of Hospital Admission of Adult Patients Using Biomarkers Collected at Triage: Retrospective Cohort Study
JMIR Bioinform Biotech 2022;3(1):e38845
URL: https://bioinform.jmir.org/2022/1/e38845
doi:10.2196/38845
PMID:

Original Paper

Treatment Discontinuation Prediction in Patients With Diabetes Using a Ranking Model: Machine Learning Model Development

Hisashi Kurasawa[1*], PhD; Kayo Waki[2*], MPH, MD, PhD; Akihiro Chiba[1,3], PhD; Tomohisa Seki[2], MD, PhD; Katsuyoshi Hayashi[1], PhD; Akinori Fujino[1], PhD; Tsuneyuki Haga[1,4], PhD; Takashi Noguchi[5], MD, PhD; Kazuhiko Ohe[2], MD, PhD

[1]Nippon Telegraph and Telephone Corporation, Tokyo, Japan

[2]Department of Healthcare Information Management, The University of Tokyo Hospital, Tokyo, Japan

[3]NTT DOCOMO, INC, Tokyo, Japan

[4]NTT-AT IPS Corporation, Kanagawa, Japan

[5]National Center for Child Health and Development, Tokyo, Japan

[*]these authors contributed equally

Corresponding Author:
Kayo Waki, MPH, MD, PhD
Department of Healthcare Information Management
The University of Tokyo Hospital
7-3-1 Hongo, Bunkyo-ku
Tokyo, 113-8655
Japan
Phone: 81 3 5800 9077
Email: kwaki-tky@m.u-tokyo.ac.jp

Abstract

Background: Treatment discontinuation (TD) is one of the major prognostic issues in diabetes care, and several models have been proposed to predict a missed appointment that may lead to TD in patients with diabetes by using binary classification models for the early detection of TD and for providing intervention support for patients. However, as binary classification models output the probability of a missed appointment occurring within a predetermined period, they are limited in their ability to estimate the magnitude of TD risk in patients with inconsistent intervals between appointments, making it difficult to prioritize patients for whom intervention support should be provided.

Objective: This study aimed to develop a machine-learned prediction model that can output a TD risk score defined by the length of time until TD and prioritize patients for intervention according to their TD risk.

Methods: This model included patients with diagnostic codes indicative of diabetes at the University of Tokyo Hospital between September 3, 2012, and May 17, 2014. The model was internally validated with patients from the same hospital from May 18, 2014, to January 29, 2016. The data used in this study included 7551 patients who visited the hospital after January 1, 2004, and had diagnostic codes indicative of diabetes. In particular, data that were recorded in the electronic medical records between September 3, 2012, and January 29, 2016, were used. The main outcome was the TD of a patient, which was defined as missing a scheduled clinical appointment and having no hospital visits within 3 times the average number of days between the visits of the patient and within 60 days. The TD risk score was calculated by using the parameters derived from the machine-learned ranking model. The prediction capacity was evaluated by using test data with the C-index for the performance of ranking patients, area under the receiver operating characteristic curve, and area under the precision-recall curve for discrimination, in addition to a calibration plot.

Results: The means (95% confidence limits) of the C-index, area under the receiver operating characteristic curve, and area under the precision-recall curve for the TD risk score were 0.749 (0.655, 0.823), 0.758 (0.649, 0.857), and 0.713 (0.554, 0.841), respectively. The observed and predicted probabilities were correlated with the calibration plots.

Conclusions: A TD risk score was developed for patients with diabetes by combining a machine-learned method with electronic medical records. The score calculation can be integrated into medical records to identify patients at high risk of TD, which would be useful in supporting diabetes care and preventing TD.

(JMIR Bioinform Biotech 2022;3(1):e37951) doi:10.2196/37951

KEYWORDS

machine learning; machine-learned ranking model; treatment discontinuation; diabetes; prediction; electronic health record; EHR; big data; ranking; algorithm

Introduction

Background

Diabetes is a chronic disease requiring both self-management and long-term management. Poor glycemic control increases the risk of complications, including cardiovascular and cerebrovascular diseases as well as macrovascular and microvascular diseases, such as nephropathy, retinopathy, and neuropathy [1-4]. To prevent the progression of these complications, adherence to dietary, exercise, and medication regimens is necessary [5]. Nonadherence has been shown to increase the risk of morbidity [4] and all-cause mortality [6].

Treatment discontinuation (TD), defined as dropping out of regular medical care, is likely to result in the worsening of glycemic control and progression of complications [3,4]. TD rates in patients with diabetes are rather high, ranging from 4% to 19% in the United Kingdom [3,4], 12% to 50% in the United States [7,8], and 13.5% to 56.9% in Japan [9,10]. Furthermore, patients who have previously discontinued treatment have been shown to have a 3-fold higher risk of repeated TD than those who have never done so [11].

Prior Work

Preventing TD is crucial in the management of diabetes, and several studies have statistically analyzed the factors associated with TD [6-8,12]. Previously identified factors include younger age [6,13], smoking [6,14], poor glycemic control [6,13,15,16], high blood pressure [13], obesity [9], medications [12,16], employment status [8,17], region [18], transportation barriers [7,19,20], clinical appointments [20], and complications [21]. The most commonly used statistical hypothesis tests are *t* test and chi-square test. However, a review [22] pointed out a variety of multilevel factors in association with TD with inconsistent findings. It has remained difficult for clinicians to carefully discern each patient's risk of TD.

Machine learning (ML) may be useful for predicting each patient's risk of TD by taking into account a wide variety of factors. Statistics focus on *explaining outcomes with data*, whereas ML focuses on *predicting outcomes with data* [23]. Although ML cannot identify consistent factors, it can inform clinicians about who is a high-risk patient for TD. It could help clinicians shift their time spent on identifying high-risk patients to encouraging them to continue treatment. According to a systematic review by Carreras-García et al [24], most studies designed their model as a binary classification problem [25] that classified scheduled appointments based on whether they were kept or missed. Furthermore, the most commonly used model was logistic regression, and the most frequently used metric was the area under the receiver operating characteristic curve (AUROC). However, as a binary classification outputs the probability of a missed appointment (MA) occurring after a predetermined period, it is limited in its ability to estimate the magnitude of TD risk in patients with inconsistent intervals between appointments. Even if a patient missed an appointment, if the frequency of visits was maintained such that their condition did not worsen thereafter, the TD risk of the patient would be low. An MA is a necessary but not sufficient condition for TD.

Goal of This Study

In this study, we aimed to develop a novel method of calculating TD risk via ML. We designed a prediction model of TD as a ranking problem with imbalanced data to compare patients by length of time until TD. The ranking problem [26] is an application of survival time analysis [27]. Cox regression [28] is generally used in statistical analysis, whereas the ranking model is used in ML [29-31]. Cox regression is a model of the hazard function in which the effects of the explanatory variables on outcomes are predetermined, requiring an assumption that they remain constant over time [28]. In contrast, the ranking model does not require this assumption and makes flexible use of the variables. Furthermore, because there was a concern that the learning model would have a heavier bias toward TD cases than treatment continuation (TC) cases, the sampling was devised on the basis of the findings of the imbalanced data.

The contributions of this work are as follows:

1. This study designed a prediction model of TD as a ranking problem with imbalanced data, which allows for a comparison of patients' risk of TD with the time remaining before TD. This is the first study to use a machine-learned ranking model to predict TD.
2. The mean (95% confidence limits) of the C-index for the TD risk score obtained with the model was 0.749 (0.655, 0.823). This was higher than 0.662 (0.574, 0.748), which was obtained with the Cox regression model; the results for the AUROC and area under the precision-recall curve (AUPRC) were similar.

Methods

Ethics Approval

This study was approved by the research ethics committees of the Graduate School of Medicine and Faculty of Medicine at the University of Tokyo (approval number: 10705) and was conducted in accordance with the Declaration of Helsinki. Informed consent was obtained, and an opportunity to opt out of participation was provided.

Study Population

All data were collected from electronic health records (EHRs) at the University of Tokyo Hospital, which included 7551 patients who visited the hospital after January 1, 2004, and had diagnostic codes indicative of diabetes. Characteristics of patient in the training and test data are shown in Table 1.

Table 1. Characteristics of patients in the training and test data.

Characteristics	Training data (n=6509)		Test data (n=1042)	
Group	TD[a] (n=204, 3.13%)	TC[b] (n=6305, 96.86%)	TD (n=38, 3.65%)	TC (n=1004, 96.35%)
Number of appointments, mean (SD)	4.8 (3.3)	10.4 (5.0)	3.1 (2.6)	5.8 (4.1)
Number of missed appointments, mean (SD)	1.6 (1.2)	1.6 (1.2)	1.2 (0.5)	1.3 (0.7)
Age (years), mean (SD)	62.6 (15.9)	66.0 (12.6)	59.9 (15.0)	61.1 (14.1)
<20, n (%)	0 (0)	3 (0.05)	0 (0)	1 (0.10)
20-30, n (%)	5 (2.50)	45 (0.71)	1 (3)	25 (2.49)
30-40, n (%)	14 (6.90)	204 (3.24)	4 (11)	63 (6.27)
40-50, n (%)	28 (13.70)	452 (7.17)	6 (16)	117 (11.65)
50-60, n (%)	31 (15.20)	883 (14)	6 (16)	188 (18.73)
60-70, n (%)	47 (23)	1950 (30.93)	8 (21)	310 (30.88)
≥70, n (%)	79 (38.70)	2768 (43.90)	13 (34)	300 (29.88)
Sex, n (%)				
Male	127 (63.30)	3777 (59.90)	25 (66)	594 (59.16)
Female	77 (37.70)	2528 (40.10)	13 (34)	410 (40.84)
Hospital visit interval in days, mean (SD)	65.9 (33.1)	57.3 (23.9)	56.2 (65.5)	49.0 (21.0)
<30, n (%)	4 (2)	283 (4.49)	7 (18)	127 (12.65)
30-60, n (%)	72 (35.30)	3237 (51.34)	15 (39)	511 (50.90)
60-90, n (%)	66 (32.30)	2140 (33.94)	3 (8)	177 (17.63)
≥90, n (%)	26 (12.80)	415 (6.58)	2 (5)	39 (3.88)
First visit, n (%)	36 (17.70)	230 (3.65)	11 (29)	150 (14.94)
HbA_{1c}[c] (NGSP[d]),%, mean (SD)	7.1 (1.2)	7.0 (1.0)	7.0 (1.1)	7.0 (1.1)
<6, n (%)	31 (15.20)	770 (12.21)	6 (16)	118 (11.75)
6-7, n (%)	64 (31.40)	2281 (36.18)	12 (32)	382 (38.05)
7-8, n (%)	48 (23.50)	1788 (28.36)	9 (24)	285 (28.39)
≥8, n (%)	33 (16.20)	632 (10.02)	4 (11)	148 (14.74)
Missing value, n (%)	28 (13.70)	834 (13.23)	7 (18)	71 (7.07)
TG[e], mg/dL, mean (SD)	182.2 (167.4)	143.5 (96.5)	199.0 (239.1)	160.5 (120.9)
<30, n (%)	0 (0)	4 (0.06)	0 (0)	0 (0)
30-150, n (%)	91 (44.60)	3601 (57.11)	15 (39)	550 (54.78)
150-300, n (%)	65 (31.90)	1631 (25.87)	10 (26)	291 (28.98)
300-750, n (%)	16 (7.80)	213 (3.38)	3 (8)	72 (7.17)
≥750, n (%)	3 (1.50)	11 (0.17)	1 (3)	6 (0.60)
Missing value, n (%)	29 (14.20)	845 (13.40)	9 (24)	85 (8.47)
HDL[f], mg/dL, mean (SD)	58.6 (15)	60.6 (16.9)	54.4 (20.3)	56.6 (16.8)
<20, n (%)	0 (0)	2 (0.03)	0 (0)	0 (0)
20 to <40, n (%)	15 (7.40)	387 (6.14)	8 (21)	130 (12.95)
40 to <100, n (%)	159 (77.90)	4882 (77.43)	20 (52)	759 (75.60)
≥100, n (%)	3 (1.50)	126 (2)	1 (3)	15 (1.49)
Missing value, n (%)	27 (13.20)	908 (14.40)	9 (24)	100 (9.96)
LDL[g], mg/dL, mean (SD)	121.6 (31.3)	111.6 (26.8)	119.9 (33.7)	113.0 (35.0)

Characteristics	Training data (n=6509)		Test data (n=1042)	
Group	TD[a] (n=204, 3.13%)	TC[b] (n=6305, 96.86%)	TD (n=38, 3.65%)	TC (n=1004, 96.35%)
<60, n (%)	2 (1)	107 (1.70)	1 (3)	26 (2.59)
60-120, n (%)	64 (31.40)	2700 (42.82)	7 (18)	338 (33.67)
120-140, n (%)	36 (17.70)	988 (15.67)	2 (5)	125 (12.45)
≥140, n (%)	32 (15.70)	532 (8.44)	5 (13)	120 (11.95)
Missing value, n (%)	70 (34.30)	1978 (31.37)	23 (61)	395 (39.34)
TCho[h], mg/dL, mean (SD)	201.6 (44.5)	189.5 (32.8)	193.3 (36.6)	192.9 (43.4)
<130, n (%)	2 (1)	152 (2.41)	1 (3)	50 (4.98)
130-220, n (%)	111 (54.40)	4202 (66.65)	20 (53)	650 (64.74)
220-240, n (%)	23 (11.30)	516 (8.18)	6 (16)	97 (9.66)
240-280, n (%)	15 (7.40)	246 (3.90)	1 (3)	77 (7.67)
≥280, n (%)	5 (2.50)	43 (0.68)	0 (0)	29 (2.89)
Missing value, n (%)	48 (23.50)	1146 (18.18)	10 (26)	101 (10.06)

[a]TD: treatment discontinuation.

[b]TC: treatment continuation.

[c]HbA_{1c}: hemoglobin A_{1c}.

[d]NGSP: National Glycohemoglobin Standardization Program.

[e]TG: triglyceride.

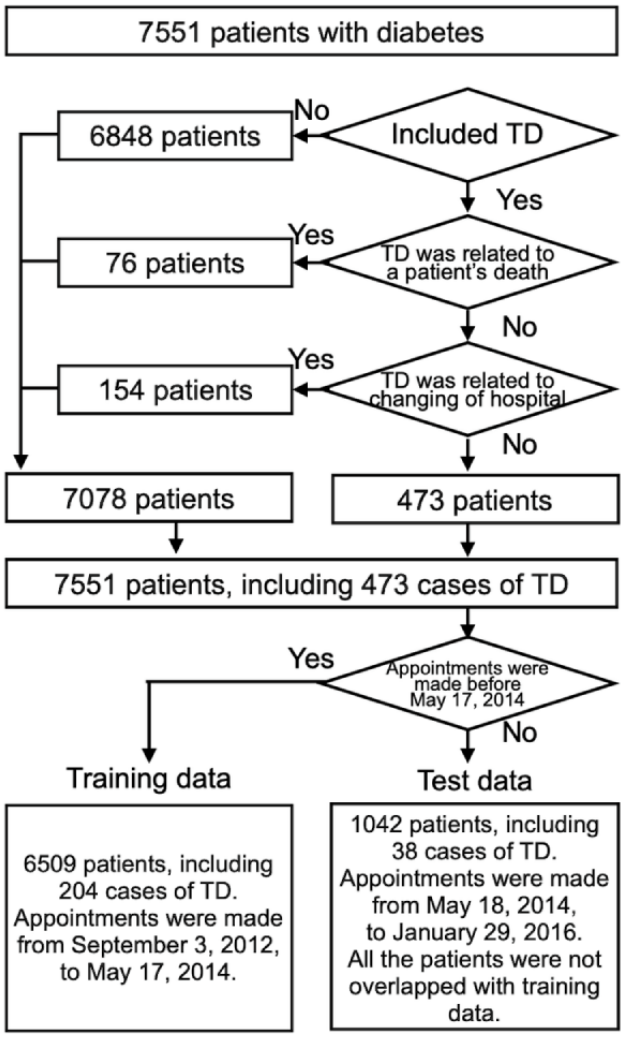
[f]HDL: high-density lipoprotein.

[g]LDL: low-density lipoprotein.

[h]TCho: total choline.

The data were recorded in the EHRs between September 3, 2012, and January 29, 2016. As illustrated in Figure 1, based on the calendar date, two-thirds of the data (days: 828/1243, 66.6%) were used for training (between September 3, 2012, and May 17, 2014) and the remaining one-third (days: 415/1243, 33.4%) was used for testing (between May 18, 2014, and January 29, 2016). The records used for training were not used for testing to ensure that the same patients were not included in both groups. A total of 6509 patients (204 cases of TD) were included in the training group, and 1042 patients (38 cases of TD) were included in the testing group.

Figure 1. Illustration of patient selection and data preprocessing. TD: treatment discontinuation.



Definition of TD

The TD of a patient was defined as missing a scheduled clinical appointment and having no hospital visits within 3 times the average number of days between the visits of the patient and within 60 days. Each patient's average number of days between visits was calculated from the last 3 visit days. In other words, if 3 times the average number of days between visits was greater than 60 days, then 60 days was used as the threshold. Otherwise, 3 times the average number of days between visits was used as the threshold.

Other studies have defined TD as the lack of hospital visits over a particular threshold of time (between 1 day and 6 months) [6-8,12-21]. When the threshold was set at 60 days, 336 cases of TD were detected in the training data and 65 cases of TD were detected in the test data, but there was a trend that patients with longer visit intervals were more likely to be judged as TD cases. It is not easy to set appropriate thresholds for outpatients whose hospital visits are at inconsistent intervals. Next, when the threshold was set to 3 times the average number of days between visits, 218 cases of TD were detected in the training data and 54 cases of TD were detected in the test data, but patients with shorter visit intervals tended to be more likely to be judged as TD cases or judged as having a risk of TD. Therefore, we included both conditions in the definition.

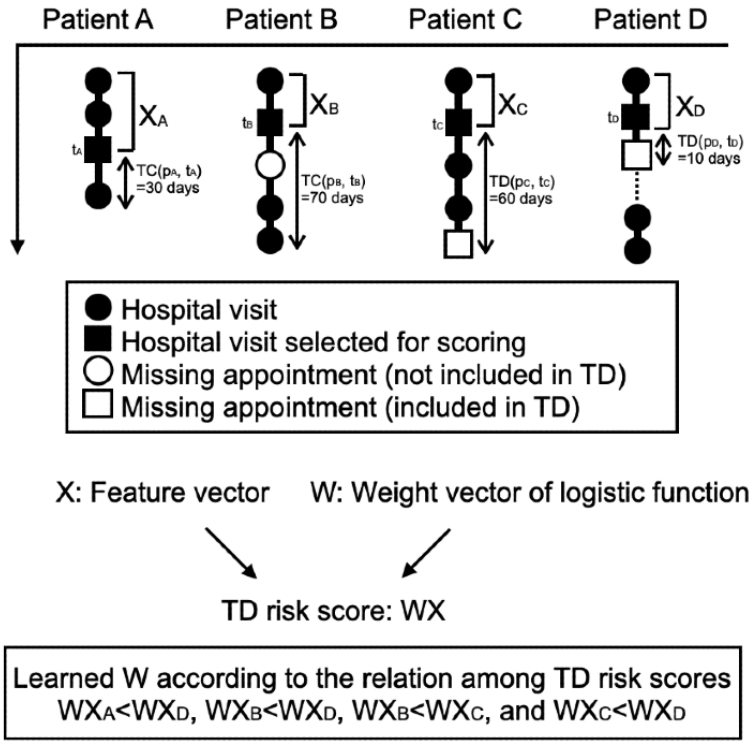
To ensure accurate TD detection, a physician, one of the coauthors, verified that the above definition was met and excluded cases of patient death or changes in care setting.

Length of Treatment Until Discontinuation

Length of treatment was measured in 2 ways. First, TD (p_m, t_m) was defined as the number of days from the date t_m to the missed scheduled clinical appointment associated with TD for the patient p_m who had TD (or possible TD). In the second way, TC (p_n, t_n) was defined as the number of days from the date t_n to the most recently recorded visit for the patient p_n who had no TD.

For example, as shown in Figure 2, in the case of patient A, there were 30 days from t_A to the most recently recorded visit, so TC (p_A, t_A) was set to 30 days. In the case of patient C, there were 60 days from t_C to the missed scheduled clinical appointment associated with TD, so TD (p_C, t_C) was set to 60 days.

Figure 2. Examples of the value of the treatment discontinuation (TD) risk. TC: treatment continuation; W: weight vector; X: feature vector.



Class Design

The classification $y_{m,n}$ was based on the difference between a pair of treatment lengths. Here, $y_{m,n}$=+1 for the pair of TD (p_m, t_m) for the patient p_m and the date t_m and TD (p_n, t_n) for the patient p_n and the date t_n if TD (p_m, t_m) is shorter than TD (p_n, t_n) and the pair of TD (p_m, t_m) and TC (p_n, t_n) if TD (p_m, t_m) is shorter than TC (p_n, t_n). $y_{m,n}$=–1 for the pair of TD (p_m, t_m) and TD (p_n, t_n) if TD (p_m, t_m) is longer than TD (p_n, t_n) and for the pair of TC (p_m, t_m) and TD (p_n, t_n) if TC (p_m, t_m) is longer than TD (p_n, t_n).

The classification was performed only when the patients had different times until TD, or when one patient had TD and the other had TC, where TC (p_n, t_n) was longer than TD (p_m, t_m). The classification was not performed on other occasions because the difference in time until TD between the 2 patients could not be compared. For the examples shown in Figure 2, the classes of the pair of TC (p_A, t_A) and TD (p_D, t_D) and that of TC (p_B, t_B) and TD (p_D, t_D), TC (p_B, t_B) and TD (p_C, t_C), and TD (p_C, t_C) and TD (p_D, t_D) were all set to −1.

Feature Design

To ensure that the factors related to TD were included, we designed a feature vector x_n for patient p_n at time t_n, representing the clinical conditions beginning with the initial visit and lasting until just before t_n. In total, 149,699 features, 51,778 qualitative features and 97,921 quantitative features, were used. Table 2 describes the features used for the prediction.

We designed the features using 3 classes of representation. The first included detailed demographic and clinical conditions (sex, age, previously consulted medical departments, diagnosed diseases, and prescribed medications). These had numerous features, most of which had a 0 value, leading to a very sparse representation.

The second class included changes occurring during the treatment of a patient to identify the risk of TD at each hospital visit. For example, we used the accumulated number of hospital visits, length of prescription time, number of medications prescribed, laboratory results, day of the week an appointment was scheduled, the interval between the date on which a clinical appointment was made and the scheduled appointment date, and the weather conditions on the appointment day. Detailed histories of hospital visits were included because features related to when and how appointments were made influenced the accuracy of the predicted MAs in our previous work [25].

The third class included data from public databases beyond the EHR. For instance, to represent the distance from a patient's home to the hospital, we used a geographic information system and measured the distance and travel time. We also used information regarding patient occupations. The observed values of each quantitative variable, for example, blood test results, were linearly transformed (normalized) to make the variance of each variable equal to 1. The transformed variable was then assigned to the vector.

Table 2. Description of explanatory variables used for prediction.

Primary and secondary categories	Qualitative variables (n=51,778), n (%)	Quantitative variables (n=97,921), n (%)	Characteristic feature (reference)
Attribute			
Sex and age	4 (0.01)	5 (0.01)	Sex and age
Address	492 (0.95)	492 (0.50)	Distance and time duration from the house to the hospital by public transport (geographic information system)
Insurance	67 (0.13)	3 (0)	Business-type category (health insurance societies of companies)
Consultation			
Medical department, outpatient, and inpatient	267 (0.52)	514 (0.52)	Previously and recently consulted medical departments
Subject	8021 (15.49)	13,108 (13.39)	Subject categories of consultation assigned by each medical department
Time	33 (0.06)	105 (0.11)	Late arrival for an appointment
Appointment (intervals and changes)	74 (0.14)	197 (0.20)	Interval between the date on which a clinical appointment was made and scheduled appointment date
Medicine			
Directions of each medicine	10,346 (19.98)	17,678 (18.05)	How many times a day medication is taken
Doses of each medicine	4570 (8.83)	33,403 (34.11)	Total amount of medication per day
Component	2332 (4.50)	5082 (5.19)	Component (medicine code defined by the Ministry of Health, Labor and Welfare)
Medical department, outpatient, and inpatient	324 (0.63)	678 (0.69)	Medication for outpatient to the department of Diabetes and Metabolic Diseases
Disease (recovered from and under treatment)	21,977 (42.44)	22,012 (22.48)	Disease category under care and recovered (ICD-10[a])
Laboratory tests			
Medical department, outpatient, and inpatient	170 (0.33)	357 (0.36)	HbA_{1c}[b], HDL-C[c], LDL-C[d], TG[e], TCho[f], etc
Order, exam and intervals	219 (0.42)	462 (0.47)	Interval between tests
Results	297 (0.57)	658 (0.67)	Categorized result according to the criteria (Diabetes Medical Guideline)
Physiological tests (order, exam, and intervals)	2237 (4.32)	2801 (2.86)	Interval between tests
Surgery (procedure)	336 (0.65)	338 (0.35)	Procedure name
Nutritional guidance (medical department, outpatient, and inpatient)	12 (0.05)	28 (0.03)	Guidance for inpatient to the department of Diabetes and Metabolic Diseases

[a]ICD-10: International Classification of Diseases, Tenth Revision.

[b]HbA_{1c}: hemoglobin A_{1c}.

[c]HDL-C: high-density lipoprotein.

[d]LDL-C: low-density lipoprotein.

[e]TG: triglycerides.

[f]TCho: total choline.

All the features were generated by processing variables obtained from the EHRs. The category with the highest number of variables was medicine. Raw categorical variables such as medicine name, component, units, inpatient and outpatient category, and department that prescribed the medicine were extracted. Raw numerical variables such as amount, dosage, and number of days or times were extracted. In addition, new numerical variables were generated by combining categorical and numeric variables such as pairs of medicine name and amount, pairs of medicine name and dosage, and pairs of medicine name and number of days or times. New categorical variables such as pairs of medicine name and inpatient and outpatient category and pairs of medicine name and department were also generated. The category with the second highest

number of features was disease. Raw categorical variables such as disease name; disease category defined by International Classification of Diseases, Tenth Revision; treatment status (under treatment and recovering); and disease type (primary disease and secondary disease) were extracted. In addition, new categorical variables such as pairs of disease name and treatment status and pairs of disease name and disease type were generated. New numerical variables were also generated by counting the number of diseases that were under treatment and recovered for each disease category. The variables of the other categories were as follows. From the attribute category, categorical variables such as sex, names of regions and cities, insurance categories, and business-type categories were extracted. Numerical variables such as age and copayment rates were extracted. Distance and travel time were generated as new numerical variables using geographic information system from region and city names, as described in the third representation class. From the consultation category, categorical variables such as department, inpatient and outpatient category, and subject name of the reservation slot were extracted. Numerical variables such as time of arrival, appointment, clinic start, and clinic end were extracted. These time intervals were generated as new numerical variables. From the appointment category, categorical variables such as department and appointment status (new, change, and cancellation) were extracted. Numerical variables such as time of registration and reservation were extracted. The new numerical variables were generated, as described in the second representation class. From the laboratory and physiological tests categories, categorical variables such as test name, department, and inpatient and outpatient category were extracted. Numerical variables such as test values were extracted. From the surgery category, categorical variables such as operative name were extracted. From the nutritional guidance category, categorical variables such as department and inpatient and outpatient categories were extracted.

Most features were generated using the following 3-step procedure. First, raw variables were extracted from each category, tied to their recorded times, and classified into categorical variables (eg, names of diagnosed diseases) and numeric variables (eg, number of medicines prescribed). Second, Categorical variables were further classified into raw categorical variables and frequency-transformed categorical variables. Third, the combinations of the raw categorical variables and the statistics of the frequency-transformed categorical variables were computed with varying window sizes to generate qualitative features and quantitative features, respectively. Numeric variables were transformed to linear and logarithmic scales, and their statistics were computed with varying window sizes to generate quantitative features. 4 statistics were used for feature generation: minimum, maximum, mean, and SD. To relate the most recent trends in circumstances to the TD risk score, periods of 3 months, 6 months, and 1 year before the target time were used as window sizes. A categorical variable was also added to indicate missing data if a feature was present for a shorter time than the window size.

For example, from the attribute category, the features sex, age, address, and insurance were extracted to express demographic conditions. The features of sex consisted of 1 qualitative variable representing male or female, 3 quantitative variables representing its frequencies with the 3 window sizes, and 3 qualitative variables representing their missing values. The frequencies of the sex variable itself have no meaning, but because it is a variable that is always listed in each EHR, it was used to represent the number of EHRs in the window size. The features of age consisted of 2 quantitative variables of linear and logarithmic scales. The features of address consisted of 48 quantitative variables of the 4 statistics of the 2 scales of the distance and travel time from a patient's home to the hospital with 3 window sizes, 48 qualitative variables representing their missing values, and 444 quantitative and qualitative variables representing the names of regions and cities and their frequencies. The features of insurance consisted of 67 qualitative variables representing insurance categories and business-type categories and 3 quantitative variables representing copayment rates.

Model Design

We established a TD risk prediction method based on the parameters of the machine-learned ranking model. There are several objective function designs for ranking models [32,33]. In particular, pointwise [26], pairwise [34-36], and listwise [37,38] approaches have been proposed. Furthermore, several learning algorithms have been developed, including ones that use logistic regression, neural networks [39], and boosting [40].

We designed the model on the basis of the pairwise approach and used logistic regression. The pairwise approach was appropriate as the only rating scale for learning was the TD risk score. Logistic regression was selected because it was the most frequently used approach in related work [24] and because it was used in our previous work [25].

We hypothesized that the risk of TD of patient p_m can be calculated from a feature vector x_m that incorporates a variety of patient information up to time t_m. Therefore, we assumed that the scalar TD risk can be represented by the inner product of a weight vector and the feature vector, that is, $w \cdot x_m$. To obtain the weight vector w, we modeled the probability that patient p_m at time t_m would discontinue treatment earlier than p_n at t_n, with x_m and x_n attributed to $y_{m,n}$ with the logistic regression:

$$P(y_{m,n} \mid x_m, x_n; w) = 1 / \{1 + \exp[-y_{m,n} w (x_m - x_n)]\}$$

The notation $w(x_m\text{-}x_n)$ denotes the scalar product of w and $x_m\text{-}x_n$.

ML Design

The ranking method, based on the pairwise approach, requires pairs of data for optimizing the parameters of the model. In general, $n(n\text{-}1)/2$ pairs can be generated for n records with no censoring. As this study included censored data that were TCs, all pairs for optimization must satisfy the abovementioned combination rule. There was also a concern that the model would have a heavier bias toward TD cases than toward TC cases. According to survey papers [41-43] on biased data, sampling has often been attempted as a way to solve this problem [44,45]. We took the means of sampling 1 record from each patient to prevent biased learning on a small number of patients. When the w estimate was computed, we randomly selected 1 recorded

date of a hospital visit for each patient and used the date t_m or t_n as the starting point of TD or TC to calculate TD (p_m, t_m) or TC (p_n, t_n). The number of all pairs satisfying the abovementioned combination rule with the sampling was 867,574 in the training data and 17,038 in the test data. The computational complexity of pairwise-based ranking learning is $O(n^2)$. The sampling results in a slightly reduced computational cost.

When the training data size, N, is smaller than the dimension of the feature vectors, or when sampling of the training data is biased, a maximum-likelihood estimation often overfits a logistic regression model to the training data, leading the model to rank many new patients inaccurately. We used an L2-norm regularization method [23] to mitigate overfitting and improve the generalizability of the model, as we did in our previous study [25].

Using training data [(x_1, x_2, $y_{1,2}$),..., (x_1, x_N, $y_{1,\ N}$),..., (x_2, x_3, $y_{2,3}$),..., (x_m, x_n, $y_{m,n}$),..., (x_{N-1}, x_N, $y_{N-1,N}$)], we estimated w as follows:

where the squared L2-norm of w,  , is an L2-norm regularizer that acts as a mitigating penalty to provide large absolute weight values only to frequently occurring features in the training data.

The symbol λ is a hyperparameter for regularization and was tuned as follows: the training data were randomly split into 2 sets of data and used in a 2-fold cross-validation test; for each test, the prediction accuracy was evaluated with one set of data for training and the other set of data for testing, with λ set to 0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50, and 100. The value of λ at which the average prediction accuracy of the 2 tests was highest was chosen.

TD Risk Score Design

The TD risk score of patient p_m at time t_m is represented by the logit value $w \cdot x_m$. The higher the value of the TD risk, the earlier TD is predicted to occur. Figure 2 shows an example of the TD risk value.

Statistical Analysis

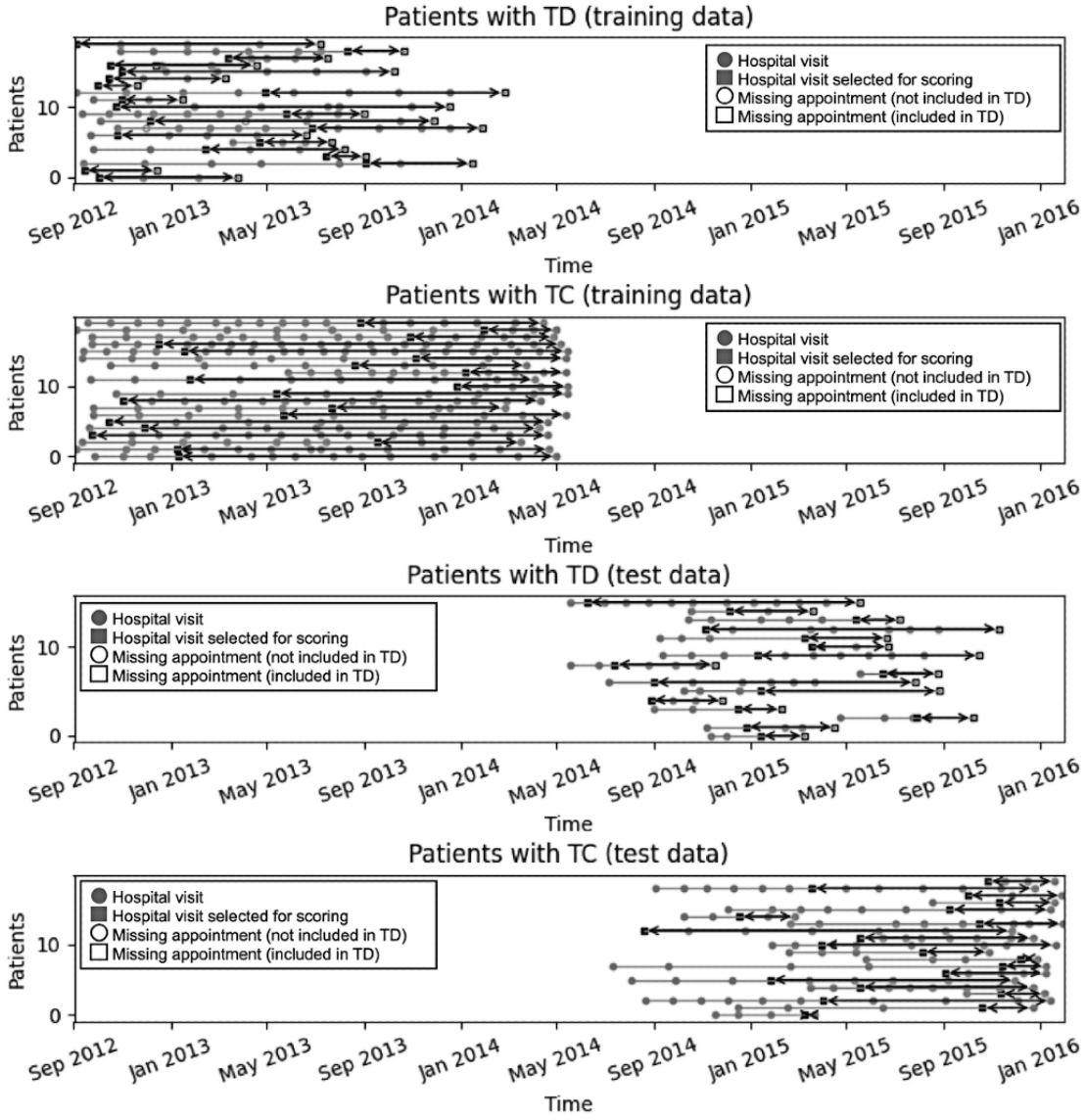
We implemented the model and ML optimization in-house in C and Python 3.7 and used it in all the experiments.

Results

Distribution of TD and TC

The detailed demographic data are shown in Table 1. The average numbers of appointments by patients with TD and TC were 4.8 and 10.4, respectively, in the training group and 3.1 and 5.8, respectively, in the testing group. The difference in distribution was because of the training and test data were classified according to whether or not they had a history of hospital visits before May 17, 2014, and the duration of the training data (828 days) was approximately twice that of the test data (415 days). Furthermore, as shown in Figure 3, the training data included patients who had been attending the hospital since before September 3, 2012, which was the starting point for the experiment; thus, patients with TC in the training data tended to have more appointments. In contrast, patients with TC in the test data tended to have fewer appointments, as these data were limited to patients who had attended the hospital since May 17, 2014. However, the number of appointments for patients with TD was low for both training and test data as patients with TD generally had shorter hospital visits. The average numbers of MAs by patients with TD and TC were 1.6 and 1.6, respectively, in the training group and 1.2 and 1.3, respectively, in the testing group.

Figure 3. Example of distribution of visit and appointment dates. TC: treatment continuation; TD: treatment discontinuation.



Predictive Performance Against TD

The hyperparameter λ of the machine-learned ranking model was tuned with 2 cross-validations, and it was set to 10 in the testing stage. The C-index of the predicted ranking was calculated as the number of correctly ranked pairs divided by the total number of comparable pairs. During testing, the TD risk score generated by the algorithm performed well, with a C-index (95% confidence limits) of 0.749 (0.655, 0.823), and outperformed the Cox regression model, with a C-index (95% confidence limits) of 0.662 (0.574, 0.748). As shown by the Kaplan-Meier curve in Figure 4, it was able to correctly model the population at high risk for TD. 10.3% (36/349) of the patients whose calibrated risk scores were ≥0.5 discontinued treatment within 100 days, whereas 93.9% (651/693) of the patients whose scores were <0.5 continued treatment for over 1 year.

The number of TD cases was much smaller in the data used in this study than the number of patients who did not interrupt their visits. As validation with the C-index alone might not be sufficient to evaluate the performance in the case of imbalanced data [45,46], the AUPRC was used in addition to the AUROC to evaluate whether the risk score could predict TD in a specific period, as shown in Table 3. Both the AUROC and AUPRC of the TD risk score were higher than those of the Cox regression model.

TD prediction within 6 months showed an AUROC (95% confidence limits) of 0.741 (0.641, 0.833) and an AUPRC (95% confidence limits) of 0.335 (0.193, 0.499). These values at 1 year were 0.758 (0.649, 0.857) and 0.713 (0.554, 0.841), respectively.

Subsequently, the TD risk score was converted to a range of 0 to 1 to validate the performance of risk stratification. As shown in the calibration plot using the test data in Figure 5, the observed and predicted TD rates were relatively correlated. These results indicate that the TD risk score can provide clinicians with information about the risk of TD in advance with favorable predictive performance and improve patient outcomes by providing room for interventions to avoid interruptions.

Figure 4. Kaplan-Meier curves displaying the probability of treatment discontinuation (TD) for the 2 groups of test data divided by the median TD risk scores obtained from the training data.

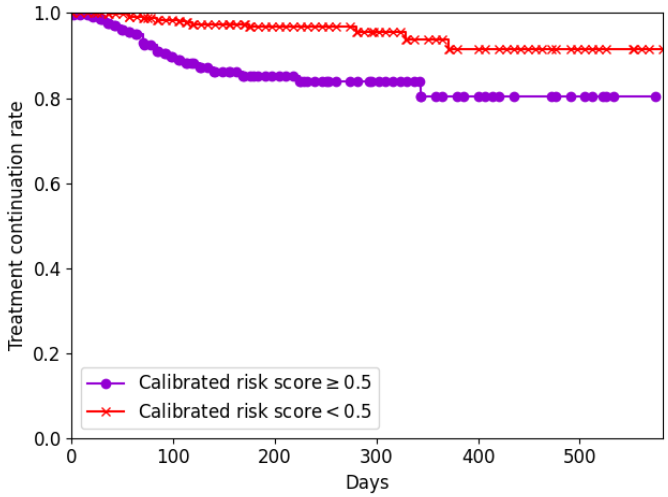


Table 3. Predictive performance against TD[a].

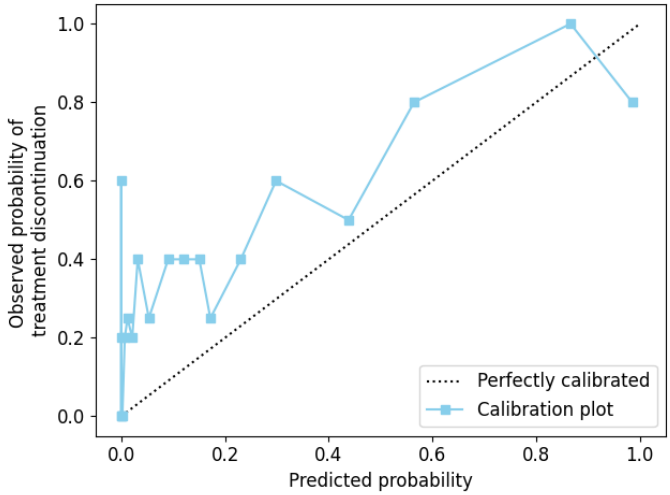
Months	AUROC[b], mean (95% confidence limits)		AUPRC[c], mean (95% confidence limits)	
	Ranking model	Cox model	Ranking model	Cox model
2	0.747 (0.607, 0.868)	0.668 (0.544, 0.787)	0.081 (0.024, 0.299)	0.035 (0.016, 0.071)
3	0.776 (0.666, 0.870)	0.691 (0.581, 0.793)	0.228 (0.090, 0.412)	0.136 (0.052, 0.262)
4	0.748 (0.637, 0.844)	0.641 (0.531, 0.746)	0.290 (0.139, 0.470)	0.156 (0.072, 0.278)
5	0.751 (0.651, 0.843)	0.666 (0.557, 0.768)	0.309 (0.163, 0.483)	0.215 (0.107, 0.360)
6	0.741 (0.641, 0.833)	0.645 (0.533, 0.751)	0.335 (0.193, 0.499)	0.236 (0.127, 0.379)
7	0.746 (0.645, 0.841)	0.660 (0.547, 0.764)	0.414 (0.254, 0.576)	0.308 (0.172, 0.468)
8	0.752 (0.650, 0.846)	0.677 (0.565, 0.781)	0.478 (0.311, 0.635)	0.384 (0.227, 0.544)
9	0.756 (0.654, 0.850)	0.675 (0.561, 0.785)	0.510 (0.337, 0.670)	0.438 (0.269, 0.601)
10	0.750 (0.646, 0.846)	0.691 (0.569, 0.800)	0.570 (0.402, 0.726)	0.562 (0.389, 0.708)
11	0.732 (0.625, 0.830)	0.680 (0.561, 0.793)	0.609 (0.442, 0.757)	0.597 (0.426, 0.742)
12	0.758 (0.649, 0.857)	0.687 (0.569, 0.798)	0.713 (0.554, 0.841)	0.645 (0.485, 0.784)

[a]TD: treatment discontinuation.

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

[c]AUPRC: area under the precision-recall curve.

Figure 5. The distribution of the predicted probability and observed probability of treatment discontinuation is shown in a line chart. Each point represents the observed and predicted probabilities for each of the 20 segments of the test population.



Items With the Largest Coefficient Values

The items with the largest coefficient values were examined to check for leakage, wherein unintended information is used for prediction and degrades the performance of the model. The 5 highest and the 5 lowest items are shown in Table 4. The specific mechanism by which each item contributes to the prediction is difficult to discuss at this time, but there were no items among the top 5 that suggested obvious leakage.

Table 4. Top 5 and bottom 5 explanatory variables obtained from the training set.

Category	Weight size	Feature
Top 1	8.1	Frequency of visits with the reservation at the department of cardiovascular medicine within 3 months
Top 2	5.2	Frequency of visits with no letter of reference within 6 months
Top 3	5.2	Frequency of visits with no letter of reference within 3 months
Top 4	5.2	Frequency of visits with the reservation before an operation in the department of cardiovascular medicine
Top 5	5.2	Frequency of laboratory tests of protein in urine within 6 months
Bottom 1	−28	Frequency of blood pressure tests within 3 months
Bottom 2	−25	Frequency of appointments of carotid artery ultrasound examination within 3 months
Bottom 3	−16	Frequency of carotid echo tests within 3 months
Bottom 4	−15	Frequency of laboratory tests of HbA_{1c}[a] within 6 months
Bottom 5	−15	Frequency of laboratory tests of HbA_{1c} within 1 year

[a]HbA_{1c}: hemoglobin A_{1c}.

Discussion

Principal Findings

In this study, we generated a prediction model for the risk of TD using approximately 150,000 explanatory variables extracted from EHRs and advanced machine-learned techniques. The accuracy of the model's prediction was validated.

Comparison With Prior Work

ML has been used in almost all aspects of diabetic research, especially in biomarker identification and diagnosis prediction [47-50]. The prediction of interruptions in medical visits requires the use of survival time analysis to build a model. However, there are few studies that have used ML for this purpose. In our study, to avoid the proportional hazard assumption of the Cox regression model and learning difficulties because of imbalanced data, we implemented a ranking method and showed that the scores calculated for each patient using the parameters obtained from the training data were useful for predicting TD, as shown in Table 3.

Our method is a novel way of constructing a survival regression model, and our experimental evaluation showed that it outperformed the existing Cox model in terms of the C-index and AUROC and AUPRC measures and that it would be a useful option for imbalanced data such as TD. The obtained level of performance was not significantly superior to that of the Cox regression model with regard to CIs. Nonetheless, it was not inferior. Many prediction tasks in the clinical domain require that imbalanced data be addressed by prediction models using survival time analysis. Our modeling method does not require the proportional hazard assumption of the Cox regression model and avoids the problem of learning from imbalanced data. It

has no variable assumptions, which allowed us to use approximately 150,000 features. Therefore, we believe that our method is a new option for survival regression models in the clinical field.

Limitations

Our study had several important limitations that must be mentioned. First, the data were obtained from just one hospital. In addition, the test data were obtained by splitting up the data from just one hospital. They may not be entirely representative of other regions because of the different implementations and degrees of diabetes care. Consequently, the results of this study are not sufficient to assess the generalizability of our method; a study using more data from different hospitals will be required.

Second, the participants with a history of TD in this study represented only 1 subgroup of patients. Some could have discontinued treatment temporarily, and we were unable to capture these patients in this study. Moreover, if a patient changed clinics without notice and continued treatment elsewhere without any evidence in the EHR, their case would have been judged as TD cases, even if that would not have been accurate. Nonetheless, because this study relied on EHR information, the findings serve the purpose of evaluating the accuracy of the model using real-world data.

Third, our method used a large number of features and optimized them with the L2-norm regularizer, which made it difficult to find features of high importance that contribute to the prediction. In the future, we intend to investigate ways to improve interpretability, such as by using explainable artificial intelligence and Lasso regularization.

Fourth, a large number of features were generated in the predefined procedure, and the inherent trends and meanings of each feature in itself are not adequately considered. The features need to be designed more appropriately to improve the interpretability of the results.

Fifth, our method was superior to the binary classification model in that it could compare a patient's risk of TD with the time remaining until TD. However, it requires $O(n^2)$ pairs to learn the model parameters, whereas a binary classification requires only O(n) records for n training data. We need to reduce the computational cost.

Finally, it should be noted that as ML generally reflects the characteristics of the majority, our results suggest that the predictive performance obtained in this study cannot be applied to a minority of clusters in the population, such as pediatric patients.

Conclusions

We developed a novel prediction model for calculating the TD risk score by applying a machine-learned ranking model to EHR data. This score showed high prediction performance and outperformed the Cox regression model. Our model can alert clinicians about the risk of TD in advance and would be useful in improving patient outcomes by providing room for interventions to avoid interruptions and support diabetes care. In addition to estimating the TD risk score, we are studying ways to predict glycemic control in patients with diabetes to further improve their care.

Acknowledgments

This work was funded by the University of Tokyo and Nippon Telegraph and Telephone Corporation in a joint research program that was conducted at the University of Tokyo Center of Innovation, Sustainable Life Care, and the Ageless Society and dedicated to self-managing health care in the Aging Society of Japan. The funding source had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. The content is solely the responsibility of the authors and does not necessarily represent the official views of the University of Tokyo Center of Innovation.

Data Availability

The data in this study are not openly available because of the restrictions imposed by the research ethics committees that approved this study.

Conflicts of Interest

HK, KH, and AF are employees of the Nippon Telegraph and Telephone Corporation (NTT), Tokyo, Japan. AC was an employee of NTT and is now an employee of NTT DOCOMO, Inc, Tokyo, Japan. TH was an employee of NTT and is now the chief executive officer of the NTT-AT IPS Corporation, Kanagawa, Japan.

References

1. Diabetes Control and Complications Trial Research Group. Effect of intensive diabetes treatment on the development and progression of long-term complications in adolescents with insulin-dependent diabetes mellitus: diabetes control and complications trial. J Pediatrics 1994 Aug;125(2):177-188. [doi: 10.1016/s0022-3476(94)70190-3] [Medline: 8040759]
2. Stratton IM, Adler AI, Neil HA, Matthews DR, Manley SE, Cull CA, et al. Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes (UKPDS 35): prospective observational study. BMJ 2000 Aug 12;321(7258):405-412 [FREE Full text] [doi: 10.1136/bmj.321.7258.405] [Medline: 10938048]

3. Archibald LK, Gill GV. Diabetic clinic defaulters — who are they and why do they default? Pract Diab Int 1992 Jan;9(1):13-14. [doi: 10.1002/pdi.1960090104]
4. Hammersley MS, Holland MR, Walford S, Thorn PA. What happens to defaulters from a diabetic clinic? Br Med J (Clin Res Ed) 1985 Nov 09;291(6505):1330-1332 [FREE Full text] [doi: 10.1136/bmj.291.6505.1330] [Medline: 3933654]
5. American Diabetes Association. Standards of medical care in diabetes—2020 abridged for primary care providers. Clin Diabetes 2020 Jan;38(1):10-38 [FREE Full text] [doi: 10.2337/cd20-as01] [Medline: 31975748]
6. Currie C, Peyrot M, Morgan C, Poole CD, Jenkins-Jones S, Rubin RR, et al. The impact of treatment noncompliance on mortality in people with type 2 diabetes. Diabetes Care 2012 Jun;35(6):1279-1284 [FREE Full text] [doi: 10.2337/dc11-1277] [Medline: 22511257]
7. Graber A, Davidson P, Brown A, McRae J, Woolridge K. Dropout and relapse during diabetes care. Diabetes Care 1992 Nov;15(11):1477-1483. [doi: 10.2337/diacare.15.11.1477] [Medline: 1468274]
8. Gucciardi E, Demelo M, Offenheim A, Stewart DE. Factors contributing to attrition behavior in diabetes self-management programs: a mixed method approach. BMC Health Serv Res 2008 Feb 04;8:33 [FREE Full text] [doi: 10.1186/1472-6963-8-33] [Medline: 18248673]
9. Kawahara R, Amemiya T, Yoshino M, Miyamae M, Sasamoto K, Omori Y. Dropout of young non-insulin-dependent diabetics from diabetic care. Diabetes Res Clin Pract 1994 Jul;24(3):181-185. [doi: 10.1016/0168-8227(94)90114-7]
10. Sone H, Kawai K, Takagi H, Yamada N, Kobayashi M. Outcome of one-year of specialist care of patients with type 2 diabetes: a multi-center prospective survey (JDDM 2). Intern Med 2006;45(9):589-597 [FREE Full text] [doi: 10.2169/internalmedicine.45.1609] [Medline: 16755089]
11. Noda M, Yamazaki K, Hayashino Y, Izumi K, Goto A. Japanese practice guidance to improve patients' adherence to appointments for diabetes care. Human Data. 2019 Jul 15. URL: https://human-data.or.jp/wp/wp-content/uploads/2018/07/dm_jushinchudan_guide43_e.pdf [accessed 2022-01-31]
12. Lee RR, Samsudin MI, Thirumoorthy T, Low LL, Kwan YH. Factors affecting follow-up non-attendance in patients with Type 2 diabetes mellitus and hypertension: a systematic review. Singapore Med J 2019 May;60(5):216-223 [FREE Full text] [doi: 10.11622/smedj.2019042] [Medline: 31187148]
13. Masuda Y, Kubo A, Kokaze A, Yoshida M, Sekiguchi K, Fukuhara N, et al. Personal features and dropout from diabetic care. Environ Health Prev Med 2006 May;11(3):115-119 [FREE Full text] [doi: 10.1265/ehpm.11.115] [Medline: 21432385]
14. Benoit SR, Ji M, Fleming R, Philis-Tsimikas A. Predictors of dropouts from a San Diego diabetes program: a case control study. Prev Chronic Dis 2004 Oct;1(4):A10 [FREE Full text] [Medline: 15670442]
15. Karter AJ, Parker MM, Moffet HH, Ahmed AT, Ferrara A, Liu JY, et al. Missed appointments and poor glycemic control: an opportunity to identify high-risk diabetic patients. Med Care 2004 Feb;42(2):110-115. [doi: 10.1097/01.mlr.0000109023.64650.73] [Medline: 14734947]
16. Díaz EG, Medina DR, López AG, Porras M. Determinants of adherence to hypoglycemic agents and medical visits in patients with type 2 diabetes mellitus. Endocrinol Diabetes Nutr 2017 Dec;64(10):531-538. [doi: 10.1016/j.endinu.2017.08.004] [Medline: 29108925]
17. Rhee MK, Slocum W, Ziemer DC, Culler SD, Cook CB, El-Kebbi IM, et al. Patient adherence improves glycemic control. Diabetes Educ 2005;31(2):240-250. [doi: 10.1177/0145721705274927] [Medline: 15797853]
18. Fullerton B, Erler A, Pöhlmann B, Gerlach FM. Predictors of dropout in the German disease management program for type 2 diabetes. BMC Health Serv Res 2012 Jan 10;12(1):8 [FREE Full text] [doi: 10.1186/1472-6963-12-8] [Medline: 22233930]
19. Buys KC, Selleck C, Buys DR. Assessing retention in a free diabetes clinic. J Nurse Practitioners 2019 Apr;15(4):301-5.e1. [doi: 10.1016/j.nurpra.2018.12.003]
20. Wong M, Haswell-Elkins M, Tamwoy E, McDermott R, d'Abbs P. Perspectives on clinic attendance, medication and foot-care among people with diabetes in the Torres Strait Islands and Northern Peninsula Area. Aust J Rural Health 2005 Jun;13(3):172-177. [doi: 10.1111/j.1440-1854.2005.00678.x] [Medline: 15932487]
21. Gibson DM. Frequency and predictors of missed visits to primary care and eye care providers for annually recommended diabetes preventive care services over a two-year period among U.S. adults with diabetes. Prev Med 2017 Dec;105:257-264. [doi: 10.1016/j.ypmed.2017.09.019] [Medline: 28963006]
22. Sun C, Taylor K, Levin S, Renda SM, Han H. Factors associated with missed appointments by adults with type 2 diabetes mellitus: a systematic review. BMJ Open Diabetes Res Care 2021 Mar 05;9(1):e001819 [FREE Full text] [doi: 10.1136/bmjdrc-2020-001819] [Medline: 33674280]
23. Bishop CM. Pattern Recognition and Machine Learning. New York, NY: Springer; 2006.
24. Carreras-García D, Delgado-Gómez D, Llorente-Fernández F, Arribas-Gil A. Patient no-show prediction: a systematic literature review. Entropy (Basel) 2020 Jun 17;22(6):675 [FREE Full text] [doi: 10.3390/e22060675] [Medline: 33286447]
25. Kurasawa H, Hayashi K, Fujino A, Takasugi K, Haga T, Waki K, et al. Machine-learning-based prediction of a missed scheduled clinical appointment by patients with diabetes. J Diabetes Sci Technol 2016 May;10(3):730-736 [FREE Full text] [doi: 10.1177/1932296815614866] [Medline: 26555782]
26. Liu T. Learning to rank for information retrieval. FNT Inform Retrieval 2009;3(3):225-331. [doi: 10.1561/1500000016]
27. Wang P, Li Y, Reddy CK. Machine learning for survival analysis. ACM Comput Surv 2019 Nov 30;51(6):1-36. [doi: 10.1145/3214306]

28. Cox DR. Regression models and life-tables. J Royal Statistical Soc Series B (Methodological) 2018 Dec 05;34(2):187-202. [doi: 10.1111/j.2517-6161.1972.tb00899.x]
29. Raykar V, Steck H, Krishnapuram B, Dehing-Oberije C, Lambin P. On ranking in survival analysis: bounds on the concordance index. In: Proceedings of the Advances in Neural Information Processing Systems 20 (NIPS 2007). 2007 Presented at: Advances in Neural Information Processing Systems 20 (NIPS 2007); Dec 3-6, 2007; Vancouver, British Columbia. [doi: 10.5555/2981562.2981714]
30. Van Belle V, Pelckmans K, Van Huffel S, Suykens JA. Support vector methods for survival analysis: a comparison between ranking and regression approaches. Artif Intell Med 2011 Oct;53(2):107-118. [doi: 10.1016/j.artmed.2011.06.006] [Medline: 21821401]
31. Chen H, Kodell RL, Cheng KF, Chen JJ. Assessment of performance of survival prediction models for cancer prognosis. BMC Med Res Methodol 2012 Jul 23;12:102 [FREE Full text] [doi: 10.1186/1471-2288-12-102] [Medline: 22824262]
32. Burges C, Ragno R, Le Q. Learning to rank with nonsmooth cost functions. In: Advances in Neural Information Processing Systems 19. Cambridge, Massachusetts, United States: MIT Press; 2006.
33. Donmez P, Svore K, Burges CJ. On the local optimality of LambdaRank. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. 2009 Presented at: SIGIR '09: The 32nd International ACM SIGIR conference on research and development in Information Retrieval; Jul 19 - 23, 2009; Boston MA USA. [doi: 10.1145/1571941.1572021]
34. Cao Z, Qin T, Liu T, Tsai M, Li H. Learning to rank: from pairwise approach to listwise approach. In: Proceedings of the 24th international conference on Machine learning. 2007 Presented at: ICML '07 & ILP '07: The 24th Annual International Conference on Machine Learning held in conjunction with the 2007 International Conference on Inductive Logic Programming; Jun 20 - 24, 2007; Corvalis Oregon USA. [doi: 10.1145/1273496.1273513]
35. Furnkranz J, Hullermeier E. Preference learning and ranking by pairwise comparison. In: Preference Learning. Berlin, Heidelberg: Springer; 2010.
36. Usunier N, Buffoni D, Gallinari P. Ranking with ordered weighted pairwise classification. In: Proceedings of the 26th Annual International Conference on Machine Learning. 2009 Presented at: ICML '09: The 26th Annual International Conference on Machine Learning held in conjunction with the 2007 International Conference on Inductive Logic Programming; Jun 14 - 18, 2009; Montreal Quebec Canada. [doi: 10.1145/1553374.1553509]
37. Xia F, Liu T, Wang J, Zhang W, Li H. Listwise approach to learning to rank: theory and algorithm. In: Proceedings of the 25th international conference on Machine learning. 2008 Presented at: ICML '08: The 25th Annual International Conference on Machine Learning held in conjunction with the 2007 International Conference on Inductive Logic Programming; Jul 5 - 9, 2008; Helsinki Finland. [doi: 10.1145/1390156.1390306]
38. Shi Y, Larson M, Hanjalic A. List-wise learning to rank with matrix factorization for collaborative filtering. In: Proceedings of the fourth ACM conference on Recommender systems. 2010 Presented at: RecSys '10: Fourth ACM Conference on Recommender Systems; Sep 26 - 30, 2010; Barcelona Spain. [doi: 10.1145/1864708.1864764]
39. Burges C, Shaked T, Renshaw E, Lazier A, Deeds M, Hamilton N, et al. Learning to rank using gradient descent. In: Proceedings of the 22nd international conference on Machine learning. 2005 Presented at: ICML '05: Proceedings of the 22nd international conference on Machine learning; Aug 7 - 11, 2005; Bonn Germany. [doi: 10.1145/1102351.1102363]
40. Freund Y, Iyer R, Schapire R, Singer Y. An efficient boosting algorithm for combining preferences. J Mach Learn Res 2003;4:933-969. [doi: 10.5555/945365.964285]
41. He H, Garcia E. Learning from imbalanced data. IEEE Trans Knowl Data Eng 2009 Sep;21(9):1263-1284. [doi: 10.1109/TKDE.2008.239]
42. Sun Y, Wong AK, Kamel MS. Classification of imbalanced data: a review. Int J Patt Recogn Artif Intell 2011 Nov 21;23(04):687-719. [doi: 10.1142/S0218001409007326]
43. Japkowicz N, Stephen S. The class imbalance problem: a systematic study. Intelligent Data Analysis 2002 Nov 15;6(5):429-449. [doi: 10.3233/ida-2002-6504]
44. Branco P, Torgo L, Ribeiro RP. A survey of predictive modeling on imbalanced domains. ACM Comput Surv 2016 Nov 11;49(2):1-50. [doi: 10.1145/2907070]
45. López V, Fernández A, García S, Palade V, Herrera F. An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. Inform Sci 2013 Nov;250:113-141. [doi: 10.1016/j.ins.2013.07.007]
46. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015;10(3):e0118432 [FREE Full text] [doi: 10.1371/journal.pone.0118432] [Medline: 25738806]
47. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016 Dec 13;316(22):2402-2410. [doi: 10.1001/jama.2016.17216] [Medline: 27898976]
48. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J 2017;15:104-116 [FREE Full text] [doi: 10.1016/j.csbj.2016.12.005] [Medline: 28138367]

49. Sudharsan B, Peeples M, Shomali M. Hypoglycemia prediction using machine learning models for patients with type 2 diabetes. J Diabetes Sci Technol 2015 Jan;9(1):86-90 [FREE Full text] [doi: 10.1177/1932296814554260] [Medline: 25316712]
50. 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]

Abbreviations

AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic curve
EHR: electronic health record
HbA_{1c}: hemoglobin A_{1c}
MA: missed appointment
ML: machine learning
NTT: Nippon Telegraph and Telephone Corporation
TC: treatment continuation
TD: treatment discontinuation

Edited by A Mavragani; submitted 28.03.22; peer-reviewed by R Bellazzi, G Nneji, G Lim; comments to author 29.05.22; revised version received 19.06.22; accepted 02.09.22; published 23.09.22.

Please cite as:
Kurasawa H, Waki K, Chiba A, Seki T, Hayashi K, Fujino A, Haga T, Noguchi T, Ohe K
Treatment Discontinuation Prediction in Patients With Diabetes Using a Ranking Model: Machine Learning Model Development
JMIR Bioinform Biotech 2022;3(1):e37951
URL: https://bioinform.jmir.org/2022/1/e37951
doi:10.2196/37951
PMID:

Original Paper

Exploring the Applicability of Using Natural Language Processing to Support Nationwide Venous Thromboembolism Surveillance: Model Evaluation Study

Aaron Wendelboe[1], PhD; Ibrahim Saber[2], MD; Justin Dvorak[1], PhD; Alys Adamski[3], PhD; Natalie Feland[1], RN; Nimia Reyes[3], MD; Karon Abe[3], PhD; Thomas Ortel[2], MD, PhD; Gary Raskob[1], PhD

[1]Department of Biostatistics and Epidemiology, Hudson College of Public Health, University of Oklahoma Health Sciences Center, Oklahoma City, OK, United States

[2]Division of Hematology, Department of Medicine, Duke University, Durham, NC, United States

[3]Division of Blood Disorders, National Center on Birth Defects and Developmental Disabilities, Centers for Disease Control and Prevention, Atlanta, GA, United States

Corresponding Author:
Aaron Wendelboe, PhD
Department of Biostatistics and Epidemiology
Hudson College of Public Health
University of Oklahoma Health Sciences Center
CHB Room 301
801 NE 13th Street
Oklahoma City, OK, 73104
United States
Phone: 1 405 271 2229 ext 57897
Email: Aaron-Wendelboe@ouhsc.edu

Abstract

Background: Venous thromboembolism (VTE) is a preventable, common vascular disease that has been estimated to affect up to 900,000 people per year. It has been associated with risk factors such as recent surgery, cancer, and hospitalization. VTE surveillance for patient management and safety can be improved via natural language processing (NLP). NLP tools have the ability to access electronic medical records, identify patients that meet the VTE case definition, and subsequently enter the relevant information into a database for hospital review.

Objective: We aimed to evaluate the performance of a VTE identification model of IDEAL-X (Information and Data Extraction Using Adaptive Learning; Emory University)—an NLP tool—in automatically classifying cases of VTE by "reading" unstructured text from diagnostic imaging records collected from 2012 to 2014.

Methods: After accessing imaging records from pilot surveillance systems for VTE from Duke University and the University of Oklahoma Health Sciences Center (OUHSC), we used a VTE identification model of IDEAL-X to classify cases of VTE that had previously been manually classified. Experts reviewed the technicians' comments in each record to determine if a VTE event occurred. The performance measures calculated (with 95% CIs) were accuracy, sensitivity, specificity, and positive and negative predictive values. Chi-square tests of homogeneity were conducted to evaluate differences in performance measures by site, using a significance level of .05.

Results: The VTE model of IDEAL-X "read" 1591 records from Duke University and 1487 records from the OUHSC, for a total of 3078 records. The combined performance measures were 93.7% accuracy (95% CI 93.7%-93.8%), 96.3% sensitivity (95% CI 96.2%-96.4%), 92% specificity (95% CI 91.9%-92%), an 89.1% positive predictive value (95% CI 89%-89.2%), and a 97.3% negative predictive value (95% CI 97.3%-97.4%). The sensitivity was higher at Duke University (97.9%, 95% CI 97.8%-98%) than at the OUHSC (93.3%, 95% CI 93.1%-93.4%; $P<.001$), but the specificity was higher at the OUHSC (95.9%, 95% CI 95.8%-96%) than at Duke University (86.5%, 95% CI 86.4%-86.7%; $P<.001$).

Conclusions: The VTE model of IDEAL-X accurately classified cases of VTE from the pilot surveillance systems of two separate health systems in Durham, North Carolina, and Oklahoma City, Oklahoma. NLP is a promising tool for the design and implementation of an automated, cost-effective national surveillance system for VTE. Conducting public health surveillance at

a national scale is important for measuring disease burden and the impact of prevention measures. We recommend additional studies to identify how integrating IDEAL-X in a medical record system could further automate the surveillance process.

(JMIR Bioinform Biotech 2022;3(1):e36877) doi:10.2196/36877

KEYWORDS

venous thromboembolism; public health surveillance; machine learning; natural language processing; medical imaging review; public health

Introduction

Venous thromboembolism (VTE), which includes both deep vein thrombosis (DVT) and pulmonary embolism, is a common yet preventable vascular disease. The disease burden of VTE could be decreased through a coordinated approach to risk assessment, prophylaxis, and treatment [1]. In the United States, 36% to >50% of VTEs are associated with recent hospitalization or surgery and are considered hospital-associated VTEs [2-5]; therefore, hospital systems have the potential to facilitate effective VTE surveillance.

Conducting traditional VTE surveillance by using either active or passive methods is challenging because International Classification of Diseases codes for identifying VTE have been shown to have moderate sensitivity and positive predictive value [6-8], the manual review of medical records is labor intensive, and data entry is subject to human error. In the United States, the majority of newly generated clinical data are stored and analyzed digitally, typically in the form of an electronic medical record (EMR). As of 2017, EMRs are being used by 96% of nonfederal acute care hospitals [9], and EMR use has more than doubled since 2008 [10].

Despite years of progress in developing new database and file formats for medical record keeping, the majority of medical data are stored as unstructured text [3]. Unstructured text is a rich source of data for clinical and translational research [4]. Natural language processing (NLP) tools can be used to overcome the challenges of traditional VTE surveillance, as they can access the critical unstructured text from diagnostic imaging reports (eg, ultrasound and computed tomography [CT] angiography reports) [11], identify patients who meet the VTE case definition, and enter the relevant information into a surveillance database in an efficient amount of time [11-14].

Some of the key features involved with the use of NLP include preprocessing [7], syntactic processing, and concept and named entity recognition [6]. Preprocessing allows an algorithm to remove formatting (including carriage returns and other white-space characters) and then output a single “clean” string of text (free of markup or control characters pertaining to its original source) for later steps. Syntactic processing refers to understanding word order (eg, the subject-verb-object relationship) and references to vague nouns and pronouns, such as *it*. As a result, the algorithm is able to connect elements of complex or coordinated phrases. For example, in the sentence *There is no evidence of a filling defect in the right pulmonary artery*, the keywords that the algorithm needs to detect are *no*, *filling defect*, and *pulmonary artery*. Finally, concept and named entity recognition refer to the ability to identify variations in spelling or wording that relate to a single concept, such as the different ways clinicians may refer to, spell, or misspell *venous thromboembolism*. Linking different textual surface realizations (eg, *thrombus*, *embolism*, and *pulmonary embolism*) to a single conceptual entity (*venous thromboembolism*) facilitates classification and decreases the total number of parameters that need to be estimated in the model training stage.

Although the field of NLP is immense, with an ever-growing range of features and capabilities, the application of NLP in VTE surveillance is narrow. A specific software—IDEAL-X (Information and Data Extraction using Adaptive Learning; Emory University)—was used in a previous study to identify VTE by using the unstructured text from imaging records [14]. IDEAL-X leverages machine learning–based approaches to customize fine-tuned NLP models for various use cases. It analyzes domain-specific terminology and related linguistic features to determine a medical event. The IDEAL-X NLP tool has been applied to different use cases, and its applicability to VTE event identification has been proven by an Emory University pilot study [14]. When the IDEAL-X VTE identification model’s performance in the prefiltering of VTE records was tested in its native clinical setting, it demonstrated a sensitivity of ≥97.2% and a specificity of ≥99.3% [14]. However, since the NLP model was trained based on the records from an individual site, the prefiltering (eg, the identification of cases based on the type and severity of patients) and certain external factors (eg, speech patterns and word choices that are common to a certain clinic or geographic region) may have affected the performance of the NLP tool. Therefore, independent validation is required.

In order to evaluate the robustness and adaptability of our VTE identification model, which we developed based on the machine learning–based NLP tool IDEAL-X, and to determine how the differences among clinical settings can affect its performance (as a proof of concept for applying NLP to national VTE surveillance), we evaluated the accuracy of the VTE model in two independent health care settings—one in Durham, North Carolina, and another in Oklahoma City, Oklahoma.

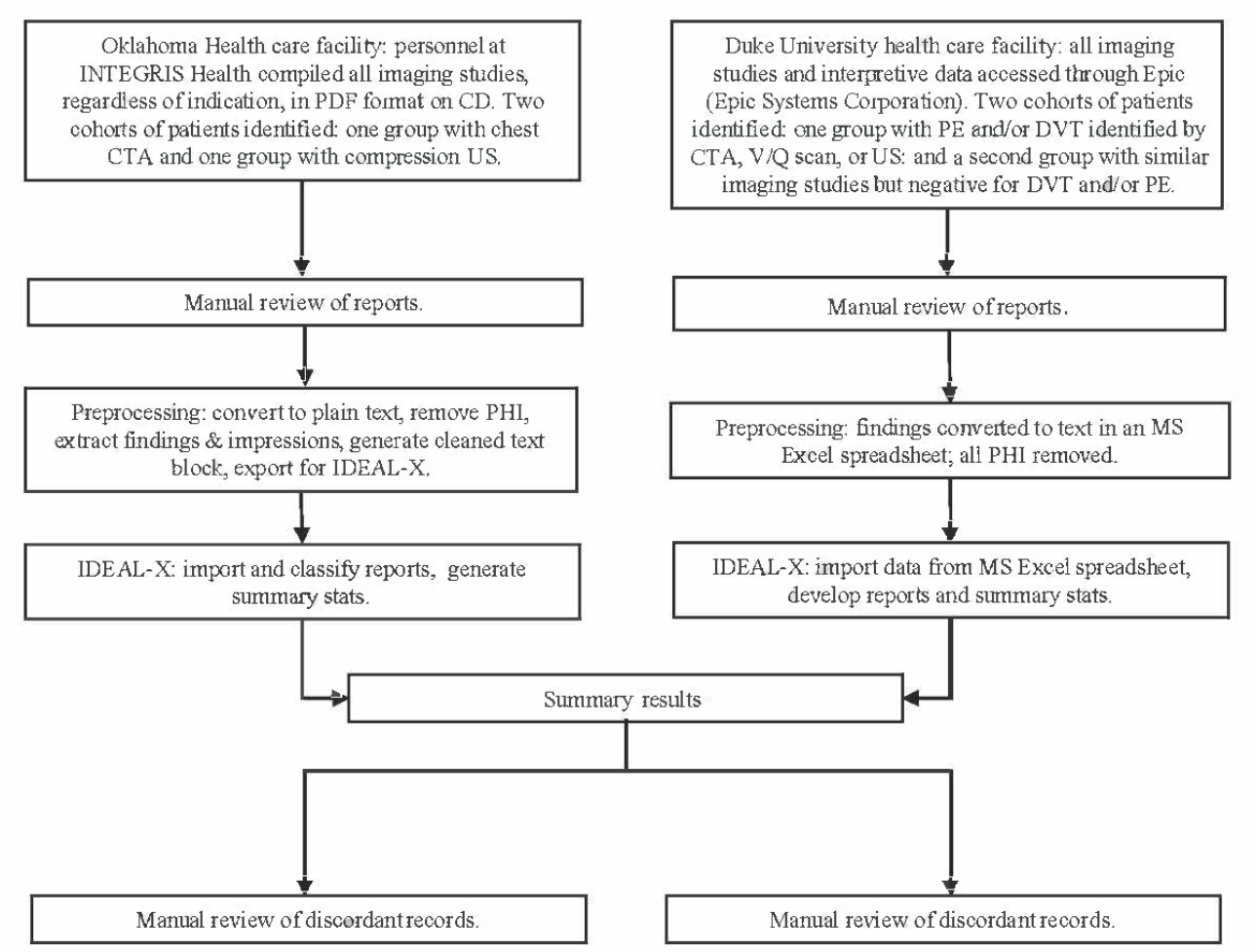
Methods

Study Design

Duke University and the University of Oklahoma Health Sciences Center (OUHSC) collaborated with the Centers for Disease Control and Prevention to establish pilot surveillance systems for VTE [15,16]. The surveillance period (ie, for data collection) for both systems ranged from April 1, 2012, to March 31, 2014 (24 months). We used data from both surveillance systems for this study and evaluation. Members of each

surveillance team served as the gold standard, manually reviewing imaging records and classifying them according to case status. Two investigators from the Duke University study team (IS and TO) and three investigators from the OUHSC study team (AW, NF, and GR) reviewed each record and classified them as positive or negative imaging reports of a DVT or pulmonary embolism. Subsequently, these records were "read" by IDEAL-X, which independently classified them according to case status. We evaluated the performance of the VTE model by comparing the case status results to the gold standard (manual review) findings. Site-specific details are described in the *Participants and Procedures* section, and the data collection and case classification methods are summarized in Figure 1.

Figure 1. Flowchart of information collection and analysis at Duke University and the University of Oklahoma Health Sciences Center. CTA: computed tomography angiography; DVT: deep vein thrombosis; IDEAL-X: Information and Data Extraction Using Adaptive Learning; MS: Microsoft; PE: pulmonary embolism; PHI: personal health information; US: ultrasound; V/Q: ventilation/perfusion.



Ethical Considerations

This study was reviewed by the Duke University Institutional Review Board and the OUHSC Institutional Review Board. Both entities determined that this study did not include research on human subjects and was therefore exempt from institutional review board approval.

Participants and Procedures

Duke University

The investigators at Duke University used the data set generated from the VTE surveillance program at three hospitals in Durham County, North Carolina (Duke University Hospital, Duke Regional Hospital, and the Durham Veterans Affairs Medical Center). The data set included all 818 unique records that were independently positive for the diagnosis of acute DVT, pulmonary embolism, or both (meeting the surveillance system's case definition). To identify a total of 773 unique negative imaging records, the investigators reviewed (1) the negative imaging records from the same cohort of patients who also had a positive imaging study (eg, a negative lower extremity ultrasound from a patient with a positive CT angiogram) and (2) the negative imaging records from patients who were identified through the VTE surveillance program but were determined via the manual evaluation of the records to not have DVT or pulmonary embolism. The Duke University team manually extracted the findings and conclusions or the *Impression* sections from each imaging report to Microsoft Excel, regardless of the terminology or the contextual information. The team excluded additional text that described patient-specific information, the indication for the imaging study, and the type of imaging study used, as well as signature lines.

The radiographic imaging records included in the Duke University data set consisted of (1) ultrasound images of upper extremities, (2) ultrasound images of lower extremities, (3) CT angiography scans of the chest, and (4) ventilation-perfusion scans.

The OUHSC

The investigators at the OUHSC requested all of the imaging records from CT angiograms and compression ultrasounds, regardless of indication, from INTEGRIS Baptist Medical Center and INTEGRIS Southwest Medical Center. To our knowledge, the records were randomly selected and were representative of the patient population. This resulted in a data set with 1487 unique patients. The OUHSC team converted the PDF imaging records (ultrasound and CT records) to plain text format. We then used a search algorithm that was customized to the formatting conventions of the records to automatically locate and demarcate the *Impressions* and *Findings* sections. For each patient, these sections were extracted; cleaned of miscellaneous punctuation, white-space, and formatting characters; and converted into a text field for entry into the IDEAL-X package. Additional text processing was conducted to categorize records according to imaging type. All automated text processing at the OUHSC study site was performed by using Python v3.7.

The IDEAL-X Tool

The VTE identification model of IDEAL-X that we used in this analysis had been used in a previous study at Emory University [14]. In that study, IDEAL-X was used to analyze the radiology reports from the Emory University Orthopedic and Spine Hospital, which were dated from February 1, 2009, to December 9, 2014. The imaging reports included interpretations from ultrasound images of the lower and upper extremities, CT scans of the chest with contrast, and magnetic resonance images of the chest [14]. We applied the VTE identification model developed by the Emory project to our data sets as part of this study, without the further calibration or retraining of the model.

Both study sites (Duke University and the OUHSC) converted their data into the format required by IDEAL-X, which consisted of a Microsoft Excel spreadsheet containing the following four columns for data entry: the *ID*, *Text*, *Manual*, and *System* columns. The *ID* column contained a deidentified record ID that was computed from the PDF image file name by using a cryptographically secure hash function. The *Text* column contained the unstructured text that was extracted from the imaging reports after preprocessing. The *Manual* column contained the gold standard diagnosis for comparison with IDEAL-X results. The *System* column, per the IDEAL-X specification, was left blank and then populated with the automated classification after processing.

Additional aggregate outputs from IDEAL-X included the total number of records, the sensitivity, the specificity, the number of true and false positives, and the number of true and false negatives. Further, 95% CIs were calculated by using the Clopper-Pearson method for binomially distributed data [17]. Chi-square tests of homogeneity were conducted to evaluate differences in performance measures by site, using a significance level of .05. We conducted a post hoc analysis of the false-positive results, in which each coauthor reviewed the text of every false-positive and false-negative result and assigned it to one of the following categories: no evidence for thrombosis, superficial vein thrombosis, chronic or residual vein thrombosis, and indeterminate.

Results

Duke University collected a total of 1591 imaging records (ultrasound images of upper extremities: n=223; ultrasound images of lower extremities: n=729; CT angiography scans of the chest: n=527; ventilation-perfusion scans: n=112). The OUHSC collected a total of 1487 imaging records (compression ultrasound images: n=1333; CT angiography scans of the chest: n=149; ventilation-perfusion scans: n=5). This provided our team with a combined total of 3078 records to be evaluated by IDEAL-X. The number of imaging records that IDEAL-X included or excluded (per the case definition for VTE) and the number of records that were manually reviewed are presented in Table 1 (the combined numbers and the numbers stratified by sites are shown). When both sites were aggregated, there were 1204 true-positive cases, 147 false-positive records, 1681 true-negative records, and 46 false-negative cases. The performance measures of the system are summarized in Table 2. Overall, the VTE model of IDEAL-X achieved over 90% accuracy (93.7%), sensitivity (96.3%), and specificity (92%).

When stratified by site, we found statistically significant differences in the performance measures between Duke University and the OUHSC. The sensitivity was significantly higher at Duke University (P<.001), while specificity was significantly higher at the OUHSC (P<.001). To further investigate differences in specificity, we identified the total number of false-positive results (147/1351, 10.9%). The reasons for the false-positive results are summarized in Table 3. The distribution varied between the two sites, and the categorical reason for false-positive results at Duke University was related to text indicating “there was no evidence for thrombosis” (104/104, 100%). Further, 38 of the 104 (36.5%) false-positive results at Duke University were from reports on ventilation-perfusion scans—an imaging modality that had not been included in the machine learning phase of the VTE identification model of IDEAL-X. The remaining errors occurred with the diagnostic imaging modalities that were previously used with the model (compression ultrasound and CT angiography), and many of the errors in the corresponding imaging reports were due to incorrect line breaks in the original text, which caused the algorithm to interpret the text incorrectly. In contrast, at the OUHSC, the most common reason for a false-positive result was text stating “a blood clot in a superficial vein” (25/43, 58.1%). The 38 false-positive results at Duke University from ventilation-perfusion scans represented 79.2% (38/48) of all ventilation-perfusion scans that were manually interpreted as *negative* at Duke University. In contrast, 20 of the 104 (19.2%) false-positive results at Duke University were from CT angiograms, but these represented only 8.1% (20/248) of all CT angiograms that were manually interpreted as *negative* at Duke University.

Table 1. The distribution of imaging records that the IDEAL-X (Information and Data Extraction Using Adaptive Learning) system identified as meeting the case definition for venous thromboembolism compared to the distribution of those identified via manual review (the gold standard). The combined distributions and the distributions stratified by surveillance site are shown.

Case classification	Classification via manual review								
	Combined			Duke University			The OUHSC[a]		
	Case, n	Noncase, n	Total classifications, N	Case, n	Noncase, n	Total classifications, N	Case, n	Noncase, n	Total classifications, N
Overall classification[b]									
Case identified by IDEAL-X	1204	147	1351	801	104	905	403	43	446
Noncase identified by IDEAL-X	46	1681	1727	17	669	686	29	1012	1041
Total classifications by IDEAL-X	1250	1828	3078	818	773	1591	432	1055	1487
Classification from compression ultrasound records									
Case identified by IDEAL-X	736	85	821	465	46	511	271	39	310
Noncase identified by IDEAL-X	28	1436	1464	10	431	441	18	1005	1023
Total classifications by IDEAL-X	764	1521	2285	475	477	952	289	1044	1333
Classification from chest computed tomography angiogram records									
Case identified by IDEAL-X	403	24	427	274	20	294	129	4	133
Noncase identified by IDEAL-X	15	234	249	5	228	233	10	6	16
Total classifications by IDEAL-X	418	258	676	279	248	527	139	10	149
Classification from ventilation-perfusion scan records									
Case identified by IDEAL-X	65	38	103	62	38	100	3	0	3
Noncase identified by IDEAL-X	3	11	14	2	10	12	1	1	2
Total classifications by IDEAL-X	68	49	117	64	48	112	4	1	5

[a]OUHSC: University of Oklahoma Health Sciences Center.

[b]Includes 112 ventilation-perfusion scans from Duke University and 5 ventilation-perfusion scans from the University of Oklahoma Health Sciences Center.

Table 2. The performance of the IDEAL-X (Information and Data Extraction Using Adaptive Learning) system by surveillance site.

Performance measure	Combined performance, % (95% CI)	Performance at Duke University, % (95% CI)	Performance at the OUHSC[a], % (95% CI)
Overall classification			
Accuracy	93.7 (93.7-93.8)	92.4 (92.3-92.5)	95.2 (95.1-95.2)
Sensitivity	96.3 (96.2-96.4)	97.9 (97.8-98)	93.3 (93.1-93.4)
Specificity	92 (91.9-92)	86.5 (86.4-86.7)	95.9 (95.8-96)
PPV[b]	89.1 (89-89.2)	88.5 (88.4-88.6)	90.4 (90.1-90.5)
NPV[c]	97.3 (97.3-97.4)	97.5 (97.4-97.6)	97.2 (97.1-97.3)
Classification from compression ultrasound records			
Accuracy	95.1 (95-95.1)	94.1 (94-94.2)	95.7 (95.6-95.8)
Sensitivity	96.3 (96.2-96.4)	97.9 (97.7-98)	93.8 (93.5-94)
Specificity	94.4 (94.3-94.5)	90.4 (90.1-90.5)	96.3 (96.2-96.3)
PPV	89.7 (89.5-89.8)	91 (90.8-91.1)	87.4 (87.1-87.7)
NPV	98.1 (98-98.1)	97.7 (97.5-97.9)	98.2 (98.1-98.3)
Classification from chest computed tomography angiogram records			
Accuracy	94.2 (94.1-94.3)	95.3 (95.1-95.4)	90.6 (90-91)
Sensitivity	96.4 (96.2-96.5)	98.2 (97.9-98.4)	92.8 (92.2-93.2)
Specificity	90.7 (90.3-91)	91.9 (91.6-92.2)	60 (53.9-65.4)
PPV	94.4 (94.2-94.5)	93.2 (92.9-93.4)	97 (96.4-97.3)
NPV	94 (93.6-94.2)	97.9 (97.5-98.1)	37.5 (34-41.6)
Classification from ventilation-perfusion scan records			
Accuracy	65 (64.2-65.6)	64.3 (63.5-65)	80 (67.4-87.9)
Sensitivity	95.6 (94.5-96.2)	96.9 (95.7-97.5)	75 (60-85.1)
Specificity	22.5 (21.3-24)	20.8 (19.7-22.4)	100 (47.5-100)
PPV	63.1 (62.3-63.8)	62 (61.2-62.8)	100 (78-100)
NPV	78.6 (73.7-82)	83.3 (77.6-87)	50 (27.5-72.5)

[a]OUHSC: University of Oklahoma Health Sciences Center.

[b]PPV: positive predictive value.

[c]NPV: negative predictive value.

We also reviewed the false-negative results and summarized the findings in Table 3. Some of the potential reasons why IDEAL-X misclassified records could have been that (1) our manual reviewers had a lower threshold for investigating possible cases, such as classifying imaging records indicative of chronic VTE, a partially occluded blood vessel, or a diagnosis of thrombophlebitis as preliminary cases of VTE that would be further investigated and potentially ruled out upon further examination; (2) if the text indicated both evidence for a thrombus in one section and no evidence in another section, IDEAL-X deferred to the section indicating no evidence; and (3) IDEAL-X did not recognize certain misspellings or symbols. However, for 18 of the 46 (39.1%) false-negative cases, it is unclear why IDEAL-X misclassified the records. Of the 6 misclassified results at Duke University, 2 (33%) were from ventilation-perfusion scans.

Table 3. Reasons for discordant records.

Reasons in text	Duke University records, n (%)	OUHSC[a] records, n (%)
False-positive records		
No evidence for thrombosis	104 (100)	4 (9.3)
Superficial vein thrombosis	0 (0)	25 (58.1)
Chronic or residual deep vein thrombosis	0 (0)	13 (30.2)
Indeterminate	0 (0)	1 (2.3)
Subtotal	104 (100)	43 (100)
False-negative records		
Inclusion of questionable cases as "positive"	2 (11.8)	9 (31)
Positive and negative results in same report	2 (11.8)	6 (20.7)
Unrecognized text or symbols, misspellings	7 (41.2)	2 (6.9)
Positive report misclassified	6 (35.3)	12 (41.4)
Subtotal	17 (100)	29 (100)

[a]OUHSC: University of Oklahoma Health Sciences Center.

Discussion

Principal Findings

This study suggests that IDEAL-X is an accurate NLP tool that can be used to identify cases of VTE. This system will likely improve the efficiency of VTE surveillance by automating the identification of VTE cases via accessing information from imaging records—the most reliable data source for VTE diagnosis. Our study results contribute to those published by Dantes et al [14] by broadening the scope of use from a specialty orthopedic hospital and demonstrating IDEAL-X's utility and accuracy in general hospital settings within two different states with radiologists who used somewhat different language, word, and phrase patterns when interpreting imaging studies. In order to examine the robustness of the IDEAL-X VTE model, no additional training was applied subsequent to its configuration by researchers at Emory University [14]. Therefore, this study more fully explores the effect of how differences in hospital systems impact the VTE model's performance.

The performance of such an NLP model was impacted by the imaging modality used. The specificity and positive predictive value for ventilation-perfusion scans, of which 95.7% (112/117) were collected from the Duke University system, were low. The specificity and negative predictive value of chest CT angiograms from the OUHSC were low. These values were likely impacted because we did not receive the requested sample (as demonstrated by only having 10 records from noncases). This resulted in a case prevalence of 93.2% (139/149), which is not representative of the prevalence of pulmonary embolism in the participating health system.

A particular advantage of using NLP to classify cases is the time required for IDEAL-X to classify the records according to case status. The preprocessing time for the OUHSC records (N=1487) was approximately 5 minutes, and the postprocessing time was <1 minute. In contrast, it takes approximately 1 minute per imaging study for a surveillance officer to read the text and classify it according to case status, which translates into potentially 52.5 person-hours for classifying the records used in this study. The time savings become increasingly meaningful when considering implementing surveillance across many facilities for a continuous time frame.

Comparison With Prior Work

IDEAL-X is relatively simple compared to other common NLP tools, including cTAKES (Clinical Text Analysis Knowledge Extraction System), MetaMap, MedLEE (Medical Language Extraction and Encoding System), GATE (General Architecture for Text Engineering), NLTK (Natural Language Toolkit), and OpenNLP. Given that surveillance systems for VTE that use NLP are in a nascent stage of design and implementation, we have not yet included advanced features, such as coreference resolution, relation extraction, and semantic processing. However, these features may be warranted if additional detail is needed to identify physicians' affiliations and organizations' locations or to understand text that is as long as a paragraph (as opposed to 1-2 sentences).

In addition to being used in VTE case identification, IDEAL-X has also been used to extract treatment and prognosis information for patients with non–small cell cancer who are undergoing radiotherapy [18]; cardiac catheterization procedure reports; coronary angiography reports; and reports that contain unstructured text from medical histories, physicals, and hospital discharge summaries [19]. These studies report promising preliminary findings, showing precision values, sensitivity values, and *F* scores of 83% or greater.

Other NLP algorithms have been developed and used to identify cases of VTE. Hinz et al [20] developed an algorithm that reported a positive predictive value of 84.7%, a sensitivity of 95.3%, and an *F* score of 0.897. Gálvez et al [21] developed an NLP tool—Reveal NLP—that identified VTE cases in a pediatric population. The reported sensitivity was 97.2%, and the specificity was 92.5%. Although these previous studies used tools that they had developed, our study implemented IDEAL-X

in institutions with no connection to the software's development, providing additional insight into the usefulness and accuracy of the NLP tool.

Limitations

A primary limitation of IDEAL-X is the lack of integration into an EMR system; IDEAL-X requires personnel to manually pull imaging records—a rate-limiting step. Another limitation is the forced binary options of *case* and *not a case*, such that *indeterminate* was not an option. The observed different distributions of categories of false-positive results by site were attributed to differences in the way records were requested or pulled at each site. Imaging studies from patients with superficial vein thrombosis and chronic or residual DVT were not included in the data set at Duke University. Enabling fast and convenient customization to support various event determination criteria would be a prerequisite for the NLP tool if nationwide deployment is required. In addition, further training is needed, so that IDEAL-X accurately classifies records in a manner that accounts for the patterns detected in false-positive and false-negative records. On the other hand, for surveillance purposes, the VTE case identification criteria also need to be standardized to ensure the consistency of case reporting among different facilities.

Future efforts will be directed at fully automating VTE surveillance. One example of how to better integrate an NLP program, such as IDEAL-X, is to include it in a facility's clinical data process, so that after an imaging report is finalized and sent for billing, it is also run through IDEAL-X (and the associated preprocessing routines). In addition to classifying VTE cases in real time, the next step toward fully automating the process entails collecting demographic, clinical, and risk factor data to facilitate the interpretation of data regarding disease incidence. Other future efforts include implementing machine learning to fine-tune the IDEAL-X algorithm, so that it can "learn" how to more accurately differentiate between cases and noncases. Example text from records that generate false-positive results can be added to further train IDEAL-X and improve its accuracy. Despite the anticipated benefits of using these information extraction software tools, there are certain barriers to implementation. These barriers include the costs of customized deployment and localization and the proprietary nature of the software, as well as having personnel who are responsible for operating and maintaining the system, ensuring health care administrators buy into the benefits, and maintaining compliance with the Health Insurance Portability and Accountability Act and other regulations.

Conclusions and Public Health Impact

The use of machine learning and NLP in disease surveillance is improving the ability to access and analyze unstructured text from EMRs. Their further and extensive use are expected to reduce resource requirements (ie, time and money), while increasing the ability to standardize data collection across sites. By conducting surveillance for VTE, we would have better data for knowing if changes in clinical practice (eg, an increase in the use of direct oral anticoagulants) are reducing the burden of VTE. Enhanced VTE surveillance can improve patient management, care, and safety. Similarly, with the advent of the COVID-19 pandemic, a robust national surveillance system would be instrumental in quickly understanding the association between COVID-19 and VTE [22]. The lessons learned from using NLP in VTE disease surveillance can be extended to improve the surveillance of other hospital-related conditions for which unstructured text from medical records plays a key role in detection and classification.

Acknowledgments

We are grateful to Dr Shuai Zheng at the Centers for Disease Control and Prevention, Division of Health Quality and Promotion, for his expertise in using IDEAL-X (Information and Data Extraction Using Adaptive Learning) in this study and reviewing this manuscript for accuracy. We are also grateful for Heather Hollen's editorial review of the manuscript. We sincerely acknowledge Emory University for allowing us to use IDEAL-X in this project. In addition, we appreciate the collaboration with Lisa Hunter and Lori Black at INTEGRIS Health System in conducting the surveillance for this study. This study was supported by the Centers for Disease Control and Prevention (cooperative agreement number: #5U36OE000002-01). The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Conflicts of Interest

None declared.

References

1. Centers for Disease Control and Prevention (CDC). Venous thromboembolism in adult hospitalizations - United States, 2007-2009. MMWR Morb Mortal Wkly Rep 2012 Jun 08;61(22):401-404 [FREE Full text] [Medline: 22672974]
2. Spencer FA, Emery C, Joffe SW, Pacifico L, Lessard D, Reed G, et al. Incidence rates, clinical profile, and outcomes of patients with venous thromboembolism. The Worcester VTE study. J Thromb Thrombolysis 2009 Nov;28(4):401-409 [FREE Full text] [doi: 10.1007/s11239-009-0378-3] [Medline: 19629642]
3. Maynard G. Preventing hospital-associated venous thromboembolism: a guide for effective quality improvement, 2nd ed. Agency for Healthcare Research and Quality. 2016 Aug. URL: https://www.ahrq.gov/sites/default/files/publications/files/vteguide.pdf [accessed 2021-02-18]

4. Serhal M, Barnes GD. Venous thromboembolism: A clinician update. Vasc Med 2019 Apr;24(2):122-131. [doi: 10.1177/1358863X18821159] [Medline: 30950331]
5. Wendelboe AM, Campbell J, Ding K, Bratzler DW, Beckman MG, Reyes NL, et al. Incidence of venous thromboembolism in a racially diverse population of Oklahoma County, Oklahoma. Thromb Haemost 2021 Jun;121(6):816-825 [FREE Full text] [doi: 10.1055/s-0040-1722189] [Medline: 33423245]
6. Kaafarani HMA, Borzecki AM, Itani KMF, Loveland S, Mull HJ, Hickson K, et al. Validity of selected patient safety indicators: opportunities and concerns. J Am Coll Surg 2011 Jun;212(6):924-934. [doi: 10.1016/j.jamcollsurg.2010.07.007] [Medline: 20869268]
7. Zhan C, Battles J, Chiang YP, Hunt D. The validity of ICD-9-CM codes in identifying postoperative deep vein thrombosis and pulmonary embolism. Jt Comm J Qual Patient Saf 2007 Jun;33(6):326-331. [doi: 10.1016/s1553-7250(07)33037-7] [Medline: 17566542]
8. Fang MC, Fan D, Sung SH, Witt DM, Schmelzer JR, Steinhubl SR, et al. Validity of using inpatient and outpatient administrative codes to identify acute venous thromboembolism: The CVRN VTE study. Med Care 2017 Dec;55(12):e137-e143 [FREE Full text] [doi: 10.1097/MLR.0000000000000524] [Medline: 29135777]
9. Percent of hospitals, by type, that possess certified health IT. Office of the National Coordinator for Health Information Technology. URL: https://dashboard.healthit.gov/quickstats/pages/certified-electronic-health-record-technology-in-hospitals.php [accessed 2021-05-25]
10. Office-based physician electronic health record adoption. Office of the National Coordinator for Health Information Technology. URL: https://www.healthit.gov/data/quickstats/office-based-physician-electronic-health-record-adoption [accessed 2021-02-18]
11. Murff HJ, FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimin K, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA 2011 Aug 24;306(8):848-855. [doi: 10.1001/jama.2011.1204] [Medline: 21862746]
12. Rochefort CM, Verma AD, Eguale T, Lee TC, Buckeridge DL. A novel method of adverse event detection can accurately identify venous thromboembolisms (VTEs) from narrative electronic health record data. J Am Med Inform Assoc 2015 Jan;22(1):155-165 [FREE Full text] [doi: 10.1136/amiajnl-2014-002768] [Medline: 25332356]
13. Tian Z, Sun S, Eguale T, Rochefort CM. Automated extraction of VTE events from narrative radiology reports in electronic health records: A validation study. Med Care 2017 Oct;55(10):e73-e80 [FREE Full text] [doi: 10.1097/MLR.0000000000000346] [Medline: 25924079]
14. Dantes RB, Zheng S, Lu JJ, Beckman MG, Krishnaswamy A, Richardson LC, et al. Improved identification of venous thromboembolism from electronic medical records using a novel information extraction software platform. Med Care 2018 Sep;56(9):e54-e60 [FREE Full text] [doi: 10.1097/MLR.0000000000000831] [Medline: 29087984]
15. Wendelboe AM, Campbell J, McCumber M, Bratzler D, Ding K, Beckman M, et al. The design and implementation of a new surveillance system for venous thromboembolism using combined active and passive methods. Am Heart J 2015 Sep;170(3):447-454.e18 [FREE Full text] [doi: 10.1016/j.ahj.2015.06.004] [Medline: 26385027]
16. Ortel TL, Arnold K, Beckman M, Brown A, Reyes N, Saber I, et al. Design and implementation of a comprehensive surveillance system for venous thromboembolism in a defined region using electronic and manual approaches. Appl Clin Inform 2019 May;10(3):552-562 [FREE Full text] [doi: 10.1055/s-0039-1693711] [Medline: 31365941]
17. Fagerland MW, Lydersen S, Laake P. Recommended tests and confidence intervals for paired binomial proportions. Stat Med 2014 Jul 20;33(16):2850-2875. [doi: 10.1002/sim.6148] [Medline: 24648355]
18. Zheng S, Jabbour SK, O'Reilly SE, Lu JJ, Dong L, Ding L, et al. Automated information extraction on treatment and prognosis for non-small cell lung cancer radiotherapy patients: Clinical study. JMIR Med Inform 2018 Feb 01;6(1):e8 [FREE Full text] [doi: 10.2196/medinform.8662] [Medline: 29391345]
19. Zheng S, Lu JJ, Ghasemzadeh N, Hayek SS, Quyyumi AA, Wang F. Effective information extraction framework for heterogeneous clinical reports using online machine learning and controlled vocabularies. JMIR Med Inform 2017 May 09;5(2):e12 [FREE Full text] [doi: 10.2196/medinform.7235] [Medline: 28487265]
20. Hinz ERM, Bastarache L, Denny JC. A natural language processing algorithm to define a venous thromboembolism phenotype. AMIA Annu Symp Proc 2013 Nov 16;2013:975-983 [FREE Full text] [Medline: 24551388]
21. Gálvez JA, Pappas JM, Ahumada L, Martin JN, Simpao AF, Rehman MA, et al. The use of natural language processing on pediatric diagnostic radiology reports in the electronic health record to identify deep venous thrombosis in children. J Thromb Thrombolysis 2017 Oct;44(3):281-290. [doi: 10.1007/s11239-017-1532-y] [Medline: 28815363]
22. Di Micco P, Russo V, Lodigiani C. Venous thromboembolism and its association with COVID-19: Still an open debate. Medicina (Kaunas) 2020 Sep 27;56(10):506 [FREE Full text] [doi: 10.3390/medicina56100506] [Medline: 32992511]

Abbreviations

CT: computed tomography
cTAKES: Clinical Text Analysis Knowledge Extraction System
DVT: deep vein thrombosis

EMR: electronic medical record
GATE: General Architecture for Text Engineering
IDEAL-X: Information and Data Extraction Using Adaptive Learning
MedLEE: Medical Language Extraction and Encoding System
NLP: natural language processing
NLTK: Natural Language Toolkit
OUHSC: University of Oklahoma Health Sciences Center
VTE: venous thromboembolism

Edited by A Mavragani; submitted 31.01.22; peer-reviewed by S Doan, DW Waqar Ali; comments to author 02.05.22; revised version received 13.06.22; accepted 21.07.22; published 05.08.22.

Please cite as:
Wendelboe A, Saber I, Dvorak J, Adamski A, Feland N, Reyes N, Abe K, Ortel T, Raskob G
Exploring the Applicability of Using Natural Language Processing to Support Nationwide Venous Thromboembolism Surveillance: Model Evaluation Study
JMIR Bioinform Biotech 2022;3(1):e36877
URL: https://bioinform.jmir.org/2022/1/e36877
doi:10.2196/36877
PMID:37206160

Original Paper

The Application of Machine Learning in Predicting Mortality Risk in Patients With Severe Femoral Neck Fractures: Prediction Model Development Study

Lingxiao Xu[1], MM; Jun Liu[1], MM; Chunxia Han[1], MM; Zisheng Ai[1], PhD

Department of Medical Statistics, Tongji University, Shanghai, China

Corresponding Author:
Zisheng Ai, PhD
Department of Medical Statistics
Tongji University
No1239 Siping Road
Shanghai, 200092
China
Phone: 86 13774380743
Email: azs1966@126.com

Abstract

Background: Femoral neck fracture (FNF) accounts for approximately 3.58% of all fractures in the entire body, exhibiting an increasing trend each year. According to a survey, in 1990, the total number of hip fractures in men and women worldwide was approximately 338,000 and 917,000, respectively. In China, FNFs account for 48.22% of hip fractures. Currently, many studies have been conducted on postdischarge mortality and mortality risk in patients with FNF. However, there have been no definitive studies on in-hospital mortality or its influencing factors in patients with severe FNF admitted to the intensive care unit.

Objective: In this paper, 3 machine learning methods were used to construct a nosocomial death prediction model for patients admitted to intensive care units to assist clinicians in early clinical decision-making.

Methods: A retrospective analysis was conducted using information of a patient with FNF from the Medical Information Mart for Intensive Care III. After balancing the data set using the Synthetic Minority Oversampling Technique algorithm, patients were randomly separated into a 70% training set and a 30% testing set for the development and validation, respectively, of the prediction model. Random forest, extreme gradient boosting, and backpropagation neural network prediction models were constructed with nosocomial death as the outcome. Model performance was assessed using the area under the receiver operating characteristic curve, accuracy, precision, sensitivity, and specificity. The predictive value of the models was verified in comparison to the traditional logistic model.

Results: A total of 366 patients with FNFs were selected, including 48 cases (13.1%) of in-hospital death. Data from 636 patients were obtained by balancing the data set with the in-hospital death group to survival group as 1:1. The 3 machine learning models exhibited high predictive accuracy, and the area under the receiver operating characteristic curve of the random forest, extreme gradient boosting, and backpropagation neural network were 0.98, 0.97, and 0.95, respectively, all with higher predictive performance than the traditional logistic regression model. Ranking the importance of the feature variables, the top 10 feature variables that were meaningful for predicting the risk of in-hospital death of patients were the Simplified Acute Physiology Score II, lactate, creatinine, gender, vitamin D, calcium, creatine kinase, creatine kinase isoenzyme, white blood cell, and age.

Conclusions: Death risk assessment models constructed using machine learning have positive significance for predicting the in-hospital mortality of patients with severe disease and provide a valid basis for reducing in-hospital mortality and improving patient prognosis.

(JMIR Bioinform Biotech 2022;3(1):e38226) doi:10.2196/38226

KEYWORDS

machine learning; femoral neck fracture; hospital mortality; hip; fracture; mortality; prediction; intensive care unit; ICU; decision-making; risk; assessment; prognosis

Introduction

Femoral neck fracture (FNF) accounts for approximately 3.58% of all fractures in the entire body [1], exhibiting an increasing trend each year. According to a survey, in 1990, the total number of hip fractures in men and women worldwide was approximately 338,000 and 917,000, respectively [2]. In China, FNFs account for 48.22% of hip fractures [3].

The Medical Information Mart for Intensive Care (MIMIC) III database is a publicly available database commonly used in clinical research [4], which contains medical data on approximately 60,000 patients in the intensive care unit (ICU) at Beth Israel Deaconess Medical Center from 2001 to 2012. The ICU database is more dimensional, dense, and valuable in the field of medicine than the general patient electronic medical record database [5]. The large amount of data recorded from these treatments and examinations is conducive to the close observation of ICU patients to detect physiological changes associated with deterioration and to provide more valuable data for clinical research [6].

Currently, many studies have been conducted on postdischarge mortality and mortality risk in patients with FNF [7-9]. Sheikh et al [8] used backward stepwise likelihood ratio Cox regression model to comprehensively analyze the causes of death in patients with FNF fracture 30 days after surgery, and found that age, admission hemoglobin, and history of myocardial infarction were important influencing factors to increase mortality. Dhingra et al [9] retrospectively analyzed the influencing factors of 1-year postoperative mortality in patients older than 60 years with FNF, and found that smoking, hypertension, diabetes, low hemoglobin, elevated white blood cell count, and surgical delay (>1 week) were significantly associated with higher 1-year postoperative mortality. Frost et al [7] used logistic regression model to determine the risk factors of postoperative nosocomial death in patients with FNF and used a nomogram model to predict the risk of death in a short period of time. Studies showed that age, gender, and complications were the main risk factors for nosocomial death in patients with femoral neck fracture. However, there have been no definitive studies on in-hospital mortality or its influencing factors in such patients with severe FNF admitted to the ICU. Therefore, in this study, we used the electronic case information of FNF patients recorded in the MIMIC database to examine the factors of in-hospital mortality in patients with FNF using a machine learning model to identify indicators that are meaningful for predicting in-hospital mortality and to provide preventive measures to reduce in-hospital mortality in patients as early as possible.

Methods

Data Source

Patient data from MIMIC-III were used for this study, which is a database commonly used in critical care big data studies; it contains clinical information such as demographics, vital signs, laboratory tests, treatment protocols, and diagnostic codes for 46,520 patients in ICU.

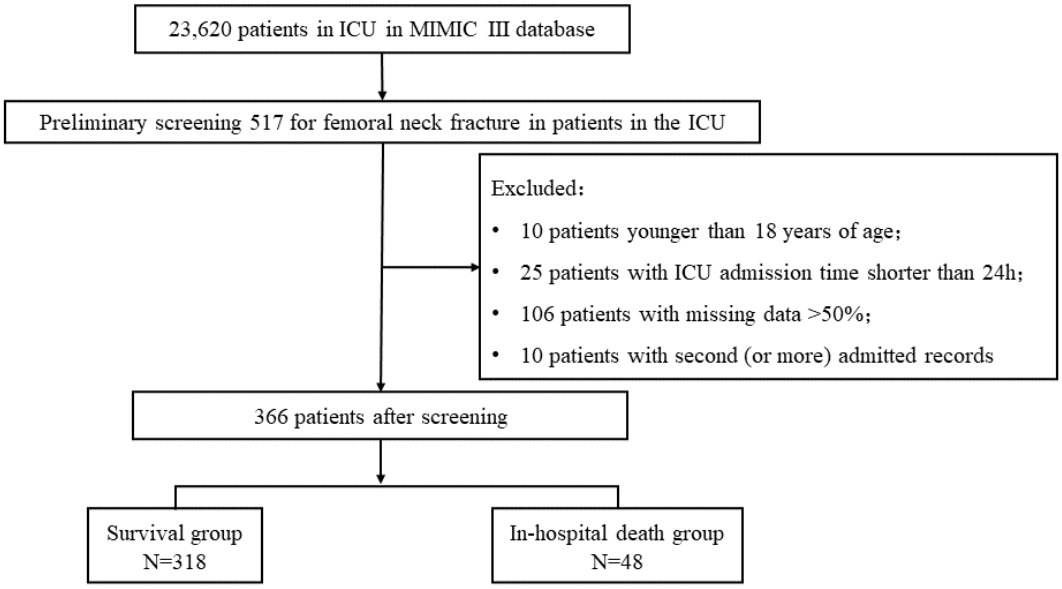
Ethical Considerations

The MIMIC-III database was approved by the Massachusetts Institute of Technology (Cambridge, MA) and Beth Israel Deaconess Medical Center (Boston, MA). The authors have obtained the database download and use right through Protecting Human Research Participants Exam (No. 38335409). Therefore, the ethical approval statement and the need for informed consent were waived for this manuscript.

Inclusion and Exclusion Criteria

In this study, patients admitted to the ICU for FNFs were extracted from the MIMIC-III database according to their diagnosis codes. The case information included in this study was based on the first admission, and data from patients with the first diagnosis code of FNF, including rotator fracture and intertrochanteric fracture, were selected according to the order of diagnosis codes. Patients aged ≤18 years or with ICU length of stay <24 hours were excluded, as were patients with grossly incomplete medical data records (>50% numbers missing). The case screening process is shown in Figure 1.

Figure 1. Case screening flowchart. ICU: intensive care unit; MIMIC: Medical Information Mart for Intensive Care.



Data Collection

Data were collected based on clinical experience, published literature, and data recorded in the MIMIC III database. Data collection for patients with FNFs was performed in the following 3 main areas: (1) demographic information—sex, age, BMI, length of ICU stay, history of previous illness, and Simplified Acute Physiology Score II (SAPS II); (2) physiological and biochemical indices within 24 hours after admission to the ICU—serum calcium, hemoglobin, hematocrit, lactate, cardiac troponin T level, creatine kinase (CK), creatine kinase isoenzyme (CKMB), vitamin D, red blood cells, white blood cells, and creatinine; and (3) outcome—whether in-hospital death occurred after admission to the ICU in patients with critical FNFs.

Data Preprocessing

The variables included in the study were screened to exclude cases with more than 50% missing values. For cases with no more than 50% missing data, random forest (RF) algorithm was used to impute variables containing missing values sequentially in a loop [10]. The common methods for filling missing data are the mean, plurality, median, and fixed value methods, and the RF algorithm is a promising method for filling missing data. The missing values are used as new labels, and the model is built to obtain predicted values for filling. The RF algorithm for filling in missing data is capable of handling mixed types of missing data and has the potential to scale up to big data environments.

Since the outcome labels extracted in this study are unbalanced (48/366, 13.1% cases in the death group and 318/366, 86.9% cases in the survival group), the prediction results of the model trained by the machine learning algorithm are prone to bias for the unbalanced data set; therefore, the original data set needs to be balanced. In this study, the synthetic minority oversampling technique (SMOTE) function in the “imblearn” library of Python (Python Software Foundation) is used to achieve the balanced processing of the data set. The SMOTE algorithm is implemented by randomly selecting a sample y from their k-nearest neighbors for each sample x in a relatively small number of mortality sample sets, and randomly synthesizing a new mortality sample on the x, y line. A total of 48 samples from the original mortality group were analyzed, and then 270 new mortality samples were randomly synthesized and added to the data set to finally obtain a new balanced data set (mortality group: survival group = 1:1).

The linear function normalization method was used in this study to normalize the newly balanced data set. Commonly used methods are linear function normalization (min-max scaling) and 0-mean normalization (*z*-score standardization). The normalization process is used to eliminate the computational errors caused by different data levels and normalize the data to the range of 0-1 to ensure that each feature is treated equally by the classifier.

The normalized data set was randomly assigned to the test set and the training set at a ratio of 7:3. Finally, 445 cases were obtained for training the prediction model, and 191 cases were used to verify the predictive performance of the model.

Model Construction

Currently, logistic regression is one of the commonly used methods for identifying risk factors that predict the occurrence of complications [11,12]. In an open calcaneal fracture study, compared to the traditional logistic regression model, machine learning methods have 30% higher accuracy and are more suitable for clinical applications [13].

RF is an integrated learning algorithm consisting of multiple decision trees formed by randomly adding back resampled samples, which is suitable for problems where the number of samples is much smaller than the number of features [14]. It also has the advantages of robust effect, fast learning speed, strong generalization ability, and good classification performance for missing data and imbalanced data [15].

Backpropagation neural network (BPNN) is a feed-forward, and the most widely used, neural network [16]. The algorithm has high self-learning and self-adaptive ability, strong generalization ability, and good prediction performance for untrained data. At the same time, the BPNN has high fault tolerance; that is, even if the system is damaged locally, it can still work normally [17].

Extreme gradient boosting (XGBoost) algorithm is a mainstream machine learning algorithm based on tree model boosting [18]. It continuously updates the error or residual of the model by adding tree models and then adjusts the weight of the misclassification results so that the model can select samples more intelligently and reduce the errors generated by the model. The XGBoost algorithm has been widely used in clinical studies for predicting the occurrence of diseases and predicting adverse patient outcomes and has been shown to be more effective than other machine learning models in several studies [19-21].

Therefore, in this study, 3 algorithms, namely RF, BPNN, and XGBoost, were used to construct machine learning prediction models (Multimedia Appendix 1).

Statistical Analysis and Model Evaluation

The PostgreSQL database system was used to extract the data. Statistical analysis was performed using SPSS 22.0 (IBM Corp), and data cleaning, model construction, and performance evaluation were performed using Python 3.8. All continuous variables are expressed as medians (quartiles), and count data are expressed as the number of cases (percentages). The Mann-Whitney *U* test was used for univariate analysis of continuous variables, and Fisher exact test was used for univariate analysis of categorical variables. The Pearson ^{2} test was used for the analysis of variance of the machine learning model results. $P<.05$ was considered to be a statistically significant difference.

The model evaluation indices were the area under the receiver operating characteristic curve (AUROC), accuracy, precision, sensitivity, specificity, and F_1-score.

Results

Basic Characteristics of Patients With Severe FNFs

A total of 366 eligible patients with FNF with a mean age of 78 (SD 20.4) years were screened. Compared with surviving patients, in-hospital death occurred in older patients with a mean age of 83 (SD 17.8) years ($P<.05$). The SAPS II score, lactate dehydrogenase level, and creatinine level of patients in the death group were all significantly higher than those in the surviving group ($P<.05$) (Table 1).

Table 1. Baseline data of patients in the intensive care unit (ICU) with a femoral neck fracture.

Characteristics	Patients included (n=366)	Survival patients (n=318)	Death patients (n=48)	*P* value
Male, n (%)	193 (52.7)	172 (54.1)	21 (43.8)	.18
Female, n (%)	173 (47.3)	146 (45.9)	27 (56.2)	.18
Diabetes, n (%)	67 (18.3)	60 (18.9)	7 (14.6)	.47
Hypertension, n (%)	149 (40.7)	130 (40.9)	19 (39.6)	.87
Coronary, n (%)	86 (23.5)	70 (22.0)	16 (33.3)	.09
LOS[a] (h) in ICU (IQR)	2.7 (1.3-4.9)	2.6 (1.4-4.7)	3.0 (1.2-6.1)	.94
BMI (IQR)	25.1 (21.0-31.3)	25.6 (21.1-31.5)	23.9 (20.6-28.6)	.17
Age (years; IQR)	78.0 (58.0-87.0)	76.5 (57.0-86.0)	83.0 (74.5-90.0)	.002
SAPS II[b] score (IQR)	39.0 (27.8-40.0)	36.0 (27.0-45.0)	52.0 (39.5-65.8)	<.001
Calcium (IQR)	1.092 (1.1-1.1)	1.092 (1.1-1.1)	1.094 (1.1-1.1)	.41
Hematocrit (IQR)	22.33 (22.1-22.6)	22.35 (22.1-22.6)	22.25 (22.0-25.1)	.41
Hemoglobin (IQR)	7.610 (7.5-7.9)	7.612 (7.5-7.9)	7.579 (7.5-8.4)	.38
Lactate (IQR)	2.127 (1.8-2.9)	2.095 (1.8-2.8)	2.678 (2.0-4.7)	.001
TnT[c] (IQR)	0.040 (0.0-0.1)	0.041 (0.0-0.1)	0.038 (0.0-0.1)	.69
CK (IQR)	156.5 (64-584.3)	171.0 (63.7-601.3)	133.0 (77.4-445.5)	.60
CKMB (IQR)	5.000 (3.3-12.0)	5.000 (3.3-12.0)	4.925 (3.5-12.6)	.69
Vitamin D (IQR)	218.7 (191.1-246.5)	218.7 (191.6-246.0)	216.1 (189.4-252.7)	.73
Red blood cell (IQR)	3.435 (3.0-3.9)	3.425 (3.0-3.9)	3.470 (3.0-3.9)	.77
White blood cell (IQR)	10.30 (7.4-13.7)	10.25 (7.4-13.7)	11.01 (7.6-14.0)	.67
Creatinine (IQR)	0.90 (0.7-1.3)	0.90 (0.7-1.2)	1.25 (0.7-1.6)	.01

[a]LOS: length of stay.

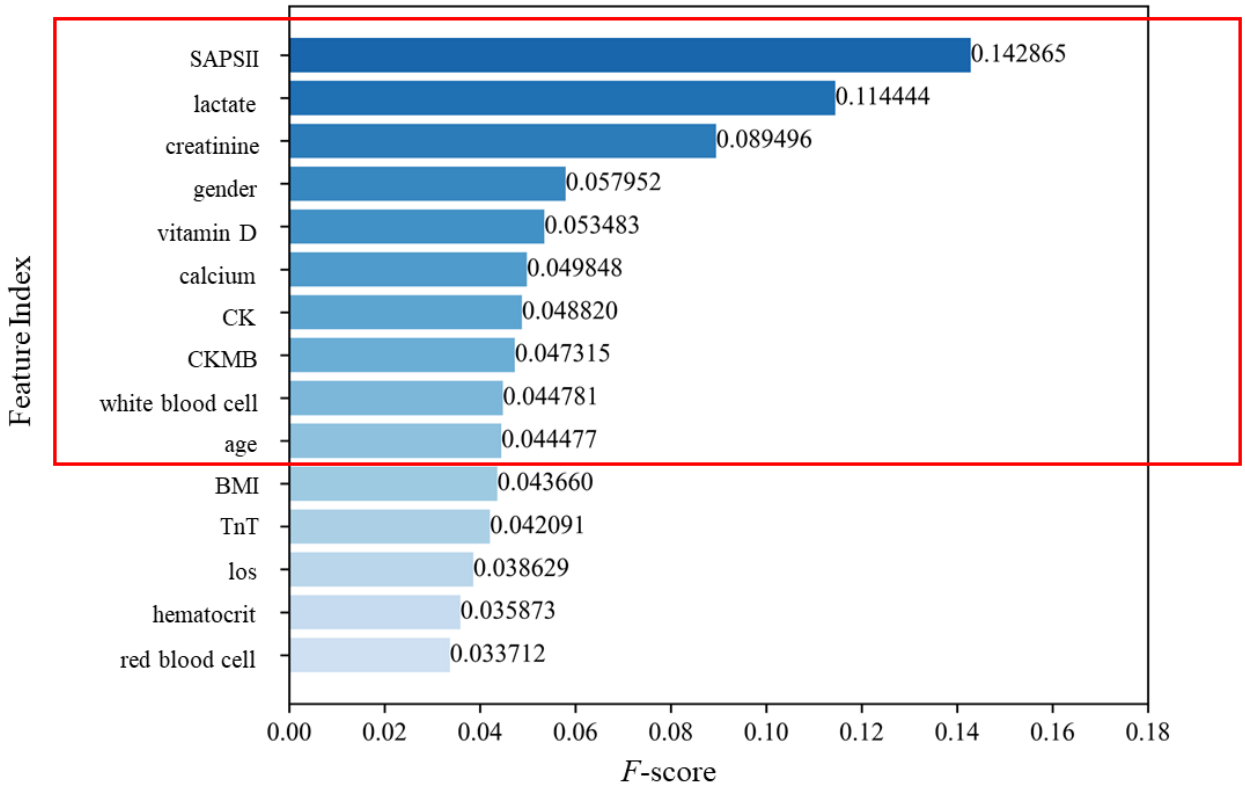
[b]SAPS II: Simplified Acute Physiology Score II.

[c]TnT: troponin T.

Ranking of the Importance of Characteristic Variables

The RF model was used to rank the importance of characteristic variables, and the top 10 variables of characteristic importance (Figure 2) were SAPS II, lactate, creatinine, gender, vitamin D, calcium, CK, CKMB, white blood cell, and age. All biochemical indices were measured within 2 hours after admission to the ICU.

Figure 2. Ranking of important features in the model. CK: creatine kinase; CKMB: creatine kinase isoenzyme; los: length of stay; SAPII: Simplified Acute Physiology Score II; TnT: troponin T.



Model Evaluation

Receiver Operating Characteristic Curve

Three machine learning models and a traditional logistic model were constructed on the training set and verified on the test set. The 3 machine learning models are RF, BPNN, and XGBoost. The receiver operating characteristic curves of the 4 prediction models were obtained, as shown in Figure 3. The AUROCs of the 4 models on the training set were 1.0, 0.99, 1.00, and 0.85, and the AUROCs on the test set were 0.99, 0.95, 0.98, and 0.86, respectively. Among them, the best results observed for the RF and XGBoost models, and the second-best for the BPNN, but the AUROCs of the machine learning models were all above 0.95. The prediction results of the 4 prediction models are analyzed for differences, and the results are shown in Table 2. The prediction accuracy of the three machine learning models on the test set is better than that of the traditional Logistic regression model, but the significant difference is not statistically significant (P>.05).

Figure 3. Receiver operating characteristic (ROC) curves of 4 prediction models: (a) random forest; (b) backpropagation neural network; (c) extreme gradient boosting; and (d) logistic regression.

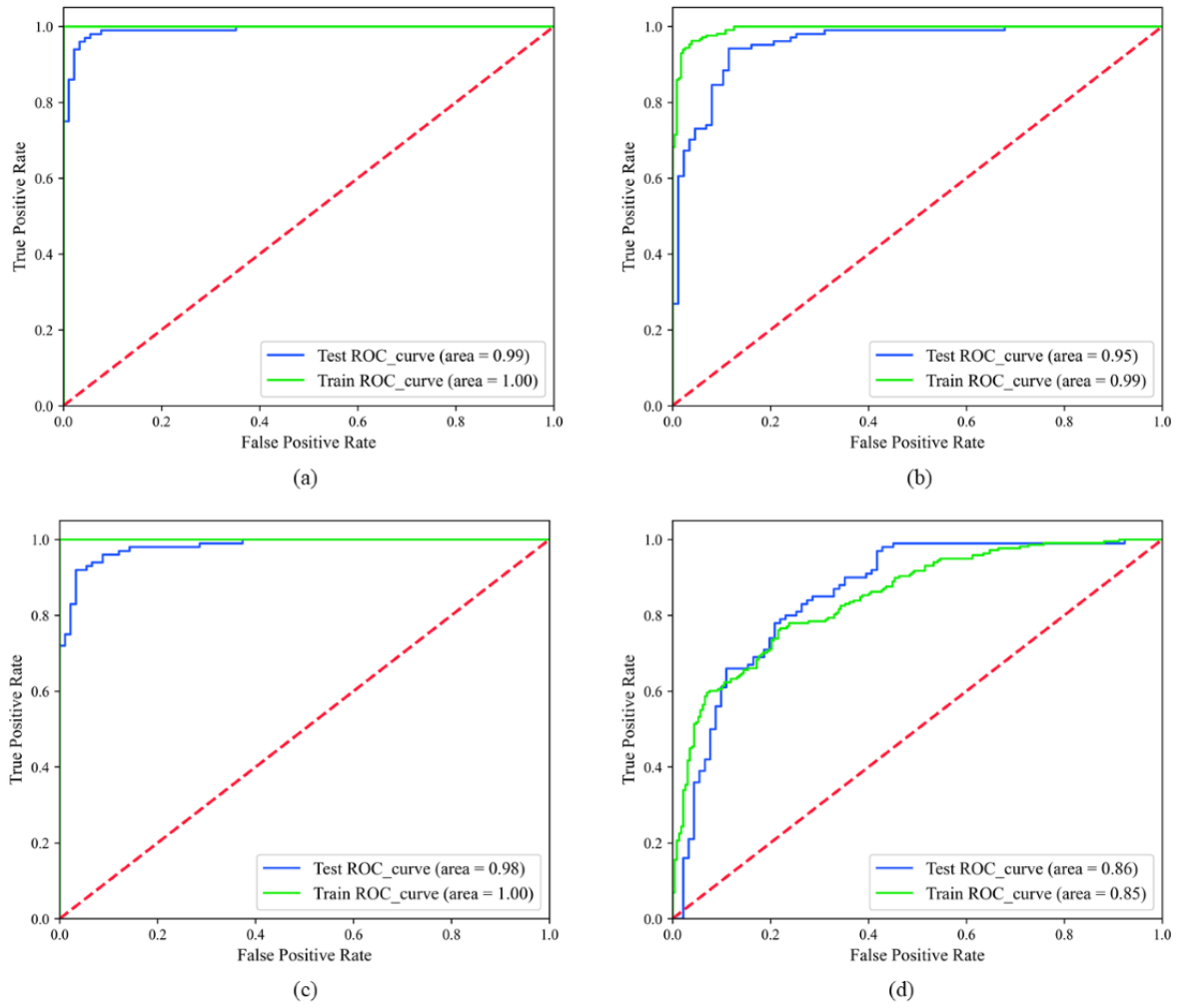


Table 2. Significance analysis of the prediction results of 4 models.

Prediction models	Outcome, n (%)		χ^2 (*df*)	*P* value
	In-hospital death	Survival		
RF[a]	103 (53.93)	88 (46.07)	2.240 (3)	.52
BPNN[b]	104 (54.45)	87 (45.55)	2.240 (3)	.52
XGBoost[c]	101 (52.88)	90 (47.12)	2.240 (3)	.52
Logistic regression	91 (47.64)	100 (52.36)	2.240 (3)	.52

[a]RF: random forest.

[b]BPNN: backpropagation neural network.

[c]XGBoost: extreme gradient boosting.

Confusion Matrix

The predictive performance of the 4 models was evaluated using accuracy, precision, sensitivity, specificity, and F_1-score. The RF model had the best overall prediction with accuracy, precision, sensitivity, specificity, and F_1-scores of 0.96, 0.97, 0.96, 0.97, and 0.92, respectively. The F_1-score of both the XGBoost and BPNN was 0.89, but the accuracy, precision, sensitivity, and specificity of XGBoost were higher than those of the BPNN. All 3 machine learning models outperformed the traditional logistic regression model (Figure 4) in terms of prediction performance (Table 3).

Figure 4. Confusion matrices for 4 prediction models; label 1 for the in-hospital death group and label 0 for the survival group: (a) random forest; (b) backpropagation neural network; (c) extreme gradient boosting; and (d) logistic regression.

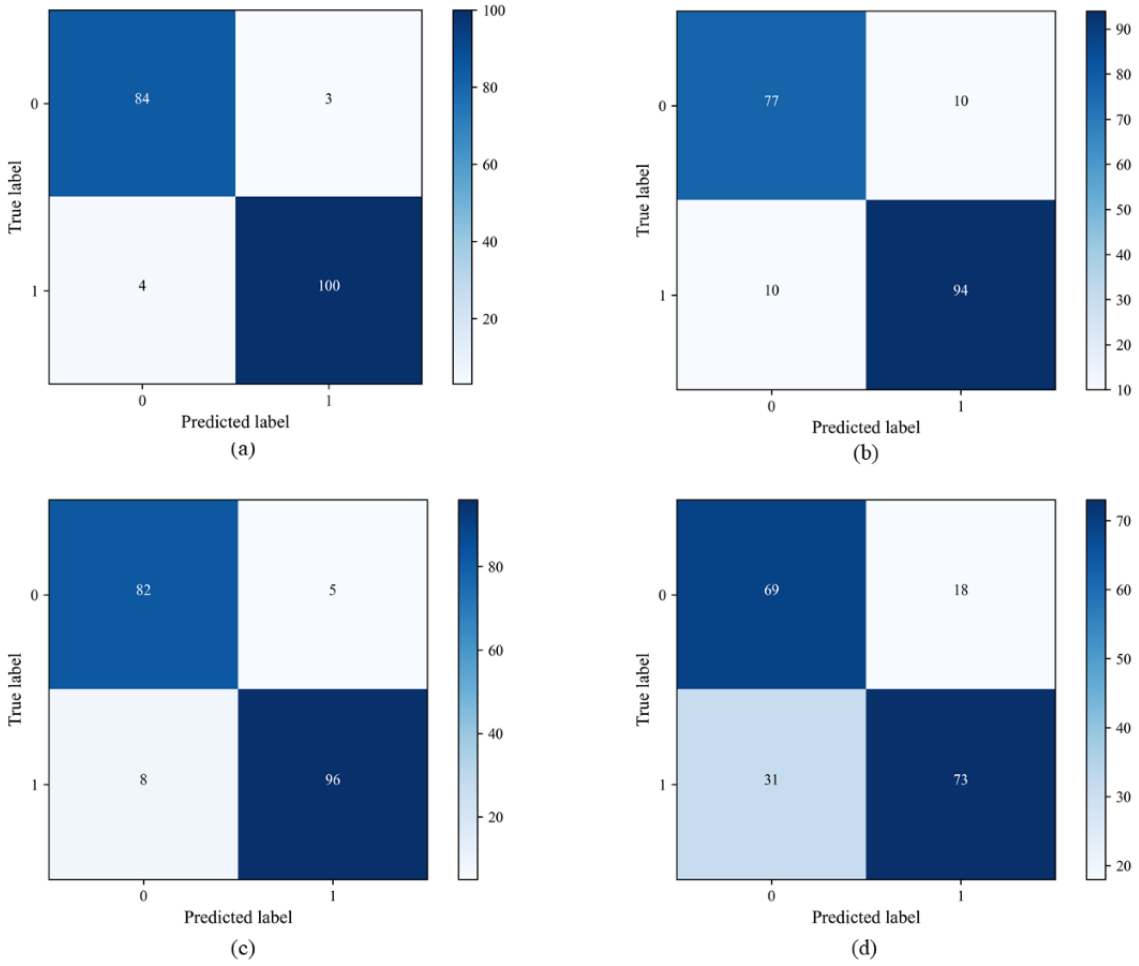


Table 3. The prediction performance evaluation of four models.

Prediction model	AUROC[a]	Accuracy	Precision	Sensitivity	Specificity	F_1-score
RF[b]	0.99	0.96	0.97	0.96	0.97	0.92
BPNN[c]	0.95	0.90	0.90	0.90	0.89	0.89
XGBoost[d]	0.98	0.93	0.95	0.92	0.94	0.89
Logistic regression	0.86	0.74	0.80	0.70	0.79	0.79

[a]AUROC: area under the receiving operating characteristic curve.

[b]RF: random forest.

[c]BPNN: backpropagation neural network.

[d]XGBoost: extreme gradient boosting.

Discussion

Principal Findings

In this study, 3 high-performing machine learning algorithms were selected to develop in-hospital mortality risk prediction models for patients with severe FNFs, including an RF model, a BPNN model, and an XGBoost model. The 3 machine learning models exhibited excellent performance on both the training and validation sets, with AUROC of the test set being 0.99, 0.95, and 0.98, respectively, and with better predictive performance compared to the traditional statistical logistic model. Meanwhile, the RF model was used in this study to rank the common predictors by calculating the importance of the feature variables. SAPS II, lactate, creatinine, gender, vitamin D, calcium, CK, CKMB, white blood cell, and age were further identified as significant predictors of death in patients with FNFs.

Comparison With Prior Work

The logistic model, a traditional statistical prediction model, has been more widely used in the prediction of morbidity and mortality in FNF [22]. However, logistic regression is more sensitive to multiple covariance data; it is difficult to deal with the problem of data imbalance; the accuracy of the model is low; and the ability to fit the true distribution of the data is poor. In recent years, machine learning has been continuously applied to the prediction of disease occurrence and adverse outcomes in medicine. For example, the risk of acute kidney injury in patients in ICU was predicted using logistic regression, RF, and LightGBM algorithms by Gao [23]. The 3 models predicted the risk of acute kidney injury after 24 hours with increasing sensitivity, and the model efficacy of the RF and LightGBM algorithms was significantly better than that of logistic regression. Huan et al [24] used machine learning to construct models to predict and analyze the risk factors of femoral head necrosis after internal fixation in patients with FNF, and the results proved that there was a good consistency between the predicted probability of machine learning and the actual risk of necrosis. In this study, the prediction effect of machine learning models was compared with that of the traditional logistic regression model, and it was confirmed that machine learning models had good performance in predicting in-hospital mortality of patients with severe FNF, which was consistent with the above conclusion.

Meanwhile, the RF model was used in this study to rank the common predictors by calculating the importance of the feature variables. SAPS II, lactate, creatinine, gender, vitamin D, calcium, CK, CKMB, white blood cell, and age were further identified as significant predictors of death in patients with FNFs. In a previous study, Seitz et al [25] found that defective bone mineralization and a decrease in 25-hydroxy vitamin D were associated with increased mortality in FNFs. 25-hydroxy vitamin D is the primary form of vitamin D present in the blood. Vitamin D and serum calcium were important, influential factors affecting in-hospital mortality in patients with FNFs in this study, which validated this finding, suggesting that balancing serum 25-hydroxy vitamin D levels through calcium supplementation and other measures in clinical treatment may reduce mortality in FNFs. In a prospective controlled study by Paccou et al [26], lactate dehydrogenase levels and creatinine levels were significant predictors of bone mineral density (BMD) loss; this is while BMD was associated with mortality, and faster BMD loss was associated with a higher risk of death [27], which is consistent with the results of this study. In addition, compared to previous studies regarding the prediction of mortality in FNFs [28-30], this study found that the SAPS II score was also significant for predicting mortality in patients. SAPS II consists of 12 physiological variables, age, type of hospitalization, and 3 types of chronic disease, and the measurement of SAPS II daily after admission to the ICU can predict the risk of death [31]. However, in existing prediction studies [32-34], the SAPS II score is commonly used in prognostic studies of patients with neurological diseases, abdominal infections, and respiratory distress, though there are fewer studies on the predictive ability of the SAPS II critical score in FNFs. The results of this study are important for further refining the prediction of morbidity and mortality in patients with FNFs.

Limitations

This study also has some limitations. First, this was a single-center study based on the MIMIC III database without external database validation, and the performance of the model needs to be further validated by prospective studies. Second, the interpretability of the machine learning model was poor, and although feature importance ranking was performed, the causal relationship between these features and in-hospital mortality in patients with FNFs could not be evaluated from a statistical perspective. Finally, some imaging metrics could not be included in the model due to limitations in the available data types in the MIMIC III database. Next, we will further integrate the existing model with the domestic database to validate the model performance, adjust the parameters to improve the model performance, and better adapt the model to the domestic database. Furthermore, we will extend the study timeline to establish a clinically applicable in-hospital mortality risk prediction model for patients with severe FNFs.

Conclusions

In summary, we used patients' clinical data to develop 3 machine learning models for predicting the risk of in-hospital death in patients with severe FNFs. The prediction performance of all 3 machine learning models was better than that of the traditional logistic model, and the RF model displayed the best prediction performance among the 3 models. In the future, after validating the domestic database and adjusting the model parameters, this model can be applied to clinical practice to better assist clinicians in decision-making, adjust treatment plans for patients with severe FNFs, better allocate medical supplies, and reduce the occurrence of adverse outcomes. Considering that MIMIC is a foreign database with fewer Asian patients, which is not universal for domestic FNF cases, more domestic patient data will be included in future work to adjust the model to make it more compatible with the characteristics of the domestic FNF population.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (grant number 81872718), Shanghai Municipal Health and Family Planning Commission (201840041), and Key undergraduate course project of Shanghai Education Commission (201965).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Paper code.
[PDF File (Adobe PDF File), 253 KB - bioinform_v3i1e38226_app1.pdf]

References

1. Zhang Y. Selection strategy and progress on the treatment of femoral neck fractures. Zhongguo Gu Shang 2015;28(9):781-783. [doi: 10.1093/med/9780199550647.003.012051]
2. Thorngren K, Hommel A, Norrman P, Thorngren J, Wingstrand H. Epidemiology of femoral neck fractures. Injury 2002 Dec;33:1-7. [doi: 10.1016/s0020-1383(02)00324-8]
3. Sun X, Zeng R, Hu Z. Femoral head necrosis after treatment of femoral neck fractures with compressive hollow screws. Chin J Orthop Trauma 2012;14(6):477-479.
4. Johnson AE, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3(1):160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
5. Zhang Y. Prediction of mortality in intensive care patients based on machine learning. University of Electronic Science and Technology of China 2018.
6. Anthony Celi L, Mark RG, Stone DJ, Montgomery RA. "Big Data" in the Intensive Care Unit. Closing the Data Loop. Am J Respir Crit Care Med 2013 Jun 01;187(11):1157-1160. [doi: 10.1164/rccm.201212-2311ed]
7. Frost SA, Nguyen ND, Black DA, Eisman JA, Nguyen TV. Risk factors for in-hospital post-hip fracture mortality. Bone 2011 Sep;49(3):553-558. [doi: 10.1016/j.bone.2011.06.002] [Medline: 21689802]
8. Sheikh HQ, Hossain FS, Aqil A, Akinbamijo B, Mushtaq V, Kapoor H. A Comprehensive Analysis of the Causes and Predictors of 30-Day Mortality Following Hip Fracture Surgery. Clin Orthop Surg 2017 Mar;9(1):10-18 [FREE Full text] [doi: 10.4055/cios.2017.9.1.10] [Medline: 28261422]
9. Dhingra M, Goyal T, Yadav A, Choudhury A. One-year mortality rates and factors affecting mortality after surgery for fracture neck of femur in the elderly. J Midlife Health 2021;12(4):276-280 [FREE Full text] [doi: 10.4103/jmh.jmh_208_20] [Medline: 35264833]
10. Tang F, Ishwaran H. Random Forest Missing Data Algorithms. Stat Anal Data Min 2017 Dec 13;10(6):363-377. [doi: 10.1002/sam.11348] [Medline: 29403567]
11. Yin W, Xu Z, Sheng J, Zhang C, Zhu Z. Logistic regression analysis of risk factors for femoral head osteonecrosis after healed intertrochanteric fractures. Hip Int 2016 May 16;26(3):215-219. [doi: 10.5301/hipint.5000346] [Medline: 27013487]
12. Pavlou M, Ambler G, Seaman S, Guttmann O, Elliott P, King M, et al. How to develop a more accurate risk prediction model when there are few events. BMJ 2015 Aug 11;351:h3868 [FREE Full text] [doi: 10.1136/bmj.h3868] [Medline: 26264962]
13. Bevevino A, Dickens J, Potter B, Dworak T, Gordon W, Forsberg J. A model to predict limb salvage in severe combat-related open calcaneus fractures. Clin Orthop Relat Res 2014 Oct;472(10):3002-3009 [FREE Full text] [doi: 10.1007/s11999-013-3382-z] [Medline: 24249536]
14. Breiman L. Random forests. Machine Learning 2001;45:5-32. [doi: 10.1023/A:1010933404324]
15. Khoshgoftaar T, Golawala M, Van HJ. An empirical study of learning from imbalanced data using random forest. 2007 Presented at: 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI); Jan 04, 2008; Patras, Greece. [doi: 10.1109/ictai.2007.46]
16. Wang L, Zeng Y, Zhang J, Huang W, Bao Y. The criticality of spare parts evaluating model using artificial neural network approach. 2006 Presented at: International Conference on Computational Science; May 28-31, 2006; Reading, UK p. 728-735. [doi: 10.1007/11758501_97]
17. Li H, Li H. Game design of self-automation based on artificial neural nets and genetic algorithms. 2009 Presented at: Second International Conference on Intelligent Computation Technology and Automation; October 10-11, 2009; Changsha, China. [doi: 10.1109/icicta.2009.86]
18. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. ACM Digital Library. 2016. URL: https://dl.acm.org/doi/pdf/10.1145/2939672.2939785 [accessed 2022-08-12]
19. Ogunleye A, Wang Q. XGBoost Model for Chronic Kidney Disease Diagnosis. IEEE/ACM Trans. Comput. Biol. and Bioinf 2020 Nov 1;17(6):2131-2140. [doi: 10.1109/tcbb.2019.2911071]
20. Wang L, Wang X, Chen A, Jin X, Che H. Prediction of Type 2 Diabetes Risk and Its Effect Evaluation Based on the XGBoost Model. Healthcare (Basel) 2020 Jul 31;8(3):247 [FREE Full text] [doi: 10.3390/healthcare8030247] [Medline: 32751894]
21. Torlay L, Perrone-Bertolotti M, Thomas E, Baciu M. Machine learning-XGBoost analysis of language networks to classify patients with epilepsy. Brain Inform 2017 Sep 22;4(3):159-169 [FREE Full text] [doi: 10.1007/s40708-017-0065-7] [Medline: 28434153]
22. Zheng JQ, Wang H, Gao YS, Ai ZS. Establishment and initial validation of the prediction model for postoperative complications of femoral neck fracture. Journal of TONGJI University (Medical Science) 2020;41(06):739-746. [doi: 10.16118/j.1008-0392.2020.06.010]

23. Gao WP, Lv HJ, Zhou L, Guo SW. Decision tree algorithm applied to MIMIC-lll database for the prediction of acute kidney injury in ICU patients. Beijing Biomedical Engineering 2021;40(06):609-617. [doi: 10.3969/j.issn.1002-3208.2021.06.010]
24. Wang H, Wu W, Han C, Zheng J, Cai X, Chang S, et al. Prediction Model of Osteonecrosis of the Femoral Head After Femoral Neck Fracture: Machine Learning-Based Development and Validation Study. JMIR Med Inform 2021 Nov 19;9(11):e30079 [FREE Full text] [doi: 10.2196/30079] [Medline: 34806984]
25. Seitz S, Koehne T, Ries C, De Novo Oliveira A, Barvencik F, Busse B, et al. Impaired bone mineralization accompanied by low vitamin D and secondary hyperparathyroidism in patients with femoral neck fracture. Osteoporos Int 2013 Feb 12;24(2):641-649. [doi: 10.1007/s00198-012-2011-0] [Medline: 22581296]
26. Paccou J, Merlusca L, Henry-Desailly I, Parcelier A, Gruson B, Royer B, et al. Alterations in bone mineral density and bone turnover markers in newly diagnosed adults with lymphoma receiving chemotherapy: a 1-year prospective pilot study. Ann Oncol 2014 Feb;25(2):481-486 [FREE Full text] [doi: 10.1093/annonc/mdt560] [Medline: 24401926]
27. Marques EA, Elbejjani M, Gudnason V, Sigurdsson G, Lang T, Sigurdsson S, et al. Proximal Femur Volumetric Bone Mineral Density and Mortality: 13 Years of Follow-Up of the AGES-Reykjavik Study. J Bone Miner Res 2017 Jun 20;32(6):1237-1242 [FREE Full text] [doi: 10.1002/jbmr.3104] [Medline: 28276125]
28. Bokshan SL, Marcaccio SE, Blood TD, Hayda RA. Factors influencing survival following hip fracture among octogenarians and nonagenarians in the United States. Injury 2018 Mar;49(3):685-690. [doi: 10.1016/j.injury.2018.02.004] [Medline: 29426609]
29. Fakler JK, Grafe A, Dinger J, Josten C, Aust G. Perioperative risk factors in patients with a femoral neck fracture - influence of 25-hydroxyvitamin D and C-reactive protein on postoperative medical complications and 1-year mortality. BMC Musculoskelet Disord 2016 Feb 01;17(1):51 [FREE Full text] [doi: 10.1186/s12891-016-0906-1] [Medline: 26833068]
30. Sebestyén A, Boncz I, Sándor J, Nyárády J. Effect of surgical delay on early mortality in patients with femoral neck fracture. Int Orthop 2008 Jun 24;32(3):375-379 [FREE Full text] [doi: 10.1007/s00264-007-0331-z] [Medline: 17323093]
31. Le Gall J. A New Simplified Acute Physiology Score (SAPS II) Based on a European/North American Multicenter Study. JAMA 1993 Dec 22;270(24):2957. [doi: 10.1001/jama.1993.03510240069035]
32. Ma LS, Su YY, Li X. Application of simplified acute physiological score II to predict the probability of death in patients with critical neurological diseases. Chinese Journal of Neurology 2010;11:774-777. [doi: 10.3760/cma.j.issn.1006-7876.2010.11.009]
33. Kuang G, Chen Y, Wei XS. The role of 24h LCR, SOFA score and SAPS II score in the prognosis evaluation of sepsis-induced by abdominal infection. J Hunan Normal Univ (Med Sci) 2020;17(01):26-29.
34. Liu H, Xiao J, Hu X, Wang I, Zhou F. The role of simplified acute physiological score-3 in selecting cortisol hormone therapy in patients with moderate to severe acute respiratory distress syndrome. Journal of Capital Medical University 2021;42(06):915-922. [doi: 10.3969/j.issn.1006-7795.2021.06.003]

Abbreviations

AUROC: area under the receiving operating characteristic curve
BMD: bone mineral density
BPNN: backpropagation neural network
CK: creatine kinase
CKMB: creatine kinase isoenzyme
FNF: femoral neck fracture
ICU: intensive care unit
MIMIC: Medical Information Mart for Intensive Care
RF: random forest
SAPS II: Simplified Acute Physiology Score II
SMOTE: synthetic minority oversampling technique
XGBoost: extreme gradient boosting

Edited by A Mavragani; submitted 24.03.22; peer-reviewed by O Fajarda Oliveira, DZ Pan; comments to author 29.06.22; revised version received 13.07.22; accepted 09.08.22; published 19.08.22.

Please cite as:
Xu L, Liu J, Han C, Ai Z
The Application of Machine Learning in Predicting Mortality Risk in Patients With Severe Femoral Neck Fractures: Prediction Model Development Study
JMIR Bioinform Biotech 2022;3(1):e38226
URL: https://bioinform.jmir.org/2022/1/e38226
doi:10.2196/38226
PMID:

Original Paper

Convolutional Neural Network–Based Automatic Classification of Colorectal and Prostate Tumor Biopsies Using Multispectral Imagery: System Development Study

Remy Peyret[1], PhD; Duaa alSaeed[2], PhD; Fouad Khelifi[1], PhD; Nadia Al-Ghreimil[2], PhD; Heyam Al-Baity[2], PhD; Ahmed Bouridane[1], PhD

[1]Northumbria University at Newcastle, Newcastle, United Kingdom

[2]College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia

Corresponding Author:
Duaa alSaeed, PhD
College of Computer and Information Sciences
King Saud University
King Abdullah Road
Riyadh, 11451
Saudi Arabia
Phone: 966 555442477
Email: dalsaeed@ksu.edu.sa

Abstract

Background: Colorectal and prostate cancers are the most common types of cancer in men worldwide. To diagnose colorectal and prostate cancer, a pathologist performs a histological analysis on needle biopsy samples. This manual process is time-consuming and error-prone, resulting in high intra- and interobserver variability, which affects diagnosis reliability.

Objective: This study aims to develop an automatic computerized system for diagnosing colorectal and prostate tumors by using images of biopsy samples to reduce time and diagnosis error rates associated with human analysis.

Methods: In this study, we proposed a convolutional neural network (CNN) model for classifying colorectal and prostate tumors from multispectral images of biopsy samples. The key idea was to remove the last block of the convolutional layers and halve the number of filters per layer.

Results: Our results showed excellent performance, with an average test accuracy of 99.8% and 99.5% for the prostate and colorectal data sets, respectively. The system showed excellent performance when compared with pretrained CNNs and other classification methods, as it avoids the preprocessing phase while using a single CNN model for the whole classification task. Overall, the proposed CNN architecture was globally the best-performing system for classifying colorectal and prostate tumor images.

Conclusions: The proposed CNN architecture was detailed and compared with previously trained network models used as feature extractors. These CNNs were also compared with other classification techniques. As opposed to pretrained CNNs and other classification approaches, the proposed CNN yielded excellent results. The computational complexity of the CNNs was also investigated, and it was shown that the proposed CNN is better at classifying images than pretrained networks because it does not require preprocessing. Thus, the overall analysis was that the proposed CNN architecture was globally the best-performing system for classifying colorectal and prostate tumor images.

(JMIR Bioinform Biotech 2022;3(1):e27394) doi:10.2196/27394

KEYWORDS

convolutional neural networks; classification; colorectal tumor; prostate tumor; machine learning; image processing

Introduction

Background

According to the World Health Organization 2014 report, 14 million new cases of cancer were diagnosed in 2012, and the disease caused 8 million people to die in the same period [1]. Colorectal cancer is the third most common cancer globally, whereas prostate cancer is the second most common cancer among men, accounting for 9.7% and 7.9% of all cancers in both sexes, respectively [1]. Both colorectal and prostate tissues are glandular and therefore have a similar histological appearance.

For prostate cancer diagnosis, the European Association of Urology guidelines [2] recommend the performance of a histological analysis on a sample taken from a needle biopsy by a pathologist who decides the grade and stage of cancer or the type of tumor based on their experience and expertise. However, this process is time consuming and it also results in a high intra- and interobserver variability [3,4], which affects diagnosis reliability. In December 1999, a study [5] of more than 6000 patients conducted by Johns Hopkins researchers found that up to 2 out of every 100 people who came to larger medical centers for treatment were given an incorrect diagnosis after histological analysis. These results suggest that second-opinion pathology examinations not only prevent errors but also save lives and money. Consequently, there is an increasing interest among pathology experts in the use of machine vision (or computational diagnosis tools) to reduce diagnosis error rates by lowering the fallible aspect of human image interpretation.

Computer-aided diagnosis can assist pathologists in reducing the human analysis time, improving efficiency, and acting as a second opinion [6-8]. Adding computer-based quantitative analysis to human qualitative interpretation could significantly reduce the intra- and interobserver variability revealed in [4]. The main objective of this study is to develop an automatic computerized system for the diagnosis of colorectal and prostate tumors using images of biopsy samples.

Numerous investigations concerning prostate or colorectal tumor classification have been carried out [9,10]. However, most use color spaces limited to gray-scale or red, green, blue (RGB) images. In the last decade, many studies have used multispectral images [11-18], which are acquired using a more precise sampling of the light spectrum. This approach aims to better capture the spectrum of the reflected light coming from the observed sample, offering more discriminative information. Lasch et al [19] suggested that multispectral imagery can improve histopathological analysis by capturing patterns that are invisible to the human vision system and standard RGB imaging. Multispectral imaging studies have shown promising results and often outperformed systems using traditional gray-scale or RGB images [9,10]. However, multispectral images contain a large amount of data, making them more difficult to process because of increased execution time and problems caused by the curse of dimensionality [13].

Since the emergence of graphic processing units (GPUs) with sufficient processing power to train Convolutional neural networks (CNNs) in 2011, these models have seen a growing interest in image classification. Several models have been developed and tested on the ImageNet data set. As an example, the AlexNet architecture was developed in 2012 [20] and won several international competitions, including the ImageNet competition. GoogLeNet [21], a 22 layers deep network, won the ImageNet competition of 2014. He et al [22] deepened the networks even more with ResNet and won the best paper in 2015 at the Conference on Computer Vision and Pattern Recognition. To reduce training times, they developed a framework in which layers are formulated as a residual function with reference to the layer input, as opposed to the unreferenced learning functions previously used. The residual network comprised 152 layers. In 2016, Google DeepMind used a mix of supervised deep learning and reinforcement learning (ie, deep reinforcement learning) to create a system capable of learning how to play the game of *Go* [23]. This program, called AlphaGo, achieved a 99.8% winning rate against other Go programs and defeated the human European Go champion by 5 games to 0. In 2017, they created AlphaGo Zero [24], which outperformed the original AlphaGo in terms of performance and learning time without using any human knowledge. CNNs seem particularly adapted to the problem of microscopic images of tumor classification. A previous study [25] applied CNNs to microscopic images of colorectal cancer and found a promising accuracy of 99.1%. However, in this study, images were preprocessed using an active contour model before being fed to the CNN model. This operation requires the intervention of a pathologist to select the region of interest from the segmented image. Otherwise, this step can be replaced by another supervised learning model, which requires more training and thus dramatically increases the processing time. This study proposes a model that does not require a preprocessing phase and uses a single CNN model for the entire classification task using multispectral images.

Deep learning is a branch of machine learning that attempts to mimic the thinking process. To process data, information is passed through a network consisting of different layers, where each layer serves as input to the following layer. The first layer of a network is referred to as the input layer, whereas the last layer is the output layer. All the layers in between are called hidden layers. Typically, a layer is a simple algorithm that consists of an activation function. This field of machine learning is now very active, and the research community is focused on solving practical applications using modern deep learning. This study aims to apply the deep learning framework to the problem at hand.

Objective

The primary objective of this study is to develop a computerized automatic system for the diagnosis of colorectal and prostate tumors using images of biopsy samples to reduce time and diagnosis error rates associated with human analysis. To achieve this, we propose a CNN model for the classification of colorectal and prostate tumors from multispectral images of biopsy samples. The key idea is based on removing the last block of

the convolutional layers and halving the number of filters per layer.

This paper is organized as follows: we first describe the principles of deep neural networks. The second section discusses the proposed method, whereas the data sets of multispectral tumor images are described in the third section. In the fourth section, the experiments carried out to validate the approach are detailed, and finally their results are presented and analyzed.

Feedforward Neural Networks

Overview

Feedforward neural networks, also called multilayer perceptrons (MLPs), are the basis of deep learning models. They aim to approximate the function f:~x!y, where ~*x* is an input feature vector and *y* is its corresponding class. The network builds a mapping ~y=$f_{(~x;)}$ by learning the parameters that provide the best approximation function to *f*. In this type of network, information moves from the input to the output through intermediate layers with no feedback connections. The number of layers is called the network depth. Each layer consists of a vector of functions or units that act in parallel, and the dimension of this vector is the width of the layer. Therefore, many hyperparameters need to be chosen when designing a neural network model, including its architecture, that is, the number of layers and units per layer.

A hidden layer computes an affine transformation of its input and then applies a nonlinear function *g*. This is defined by h=$g_{(W\sim x+b)}$, where *h* is the output of the hidden layer, *W* is the weight of the affine transformation, and *b* is the bias. *W* and *b* are the parameters learned when training the model.

The function chosen for each unit is called the activation function and is inspired by the behavior of biological neurons. The most widely used activation function is the rectified linear unit (ReLU), defined by $g_{(z)}$=$max_{(0, z)}$. Many other options are available, and the research on activation function is still a very active field. However, the ReLU has proven to perform well and is the default choice for activation functions.

Network training is performed using gradient descent. The main difference from other models is that the nonlinearity of neural networks causes the loss function to be nonconvex. Unlike convex optimization used with support vector machines or deep reinforcement learning, there is no guarantee of global convergence of a gradient descent applied to a nonconvex loss function. Consequently, the learning process is sensitive to the initial values of weights and biases. To apply gradient-based learning, a cost function must be chosen. The problem at hand in this study defines a conditional distribution $p(y|x; \theta)$ and the maximum likelihood principle is well adapted for it [26]. As a result, the cross-entropy between the training data and the model's prediction, which is equivalent to the negative log-likelihood, is used as the cost function. It enables the model to estimate the conditional probability of the classes if the input is known. The cost function model is as follows:

where  is the distribution of the training data and p_{model} is the model distribution and the set of parameters for which the cost function is calculated. Consequently, the specific form of the cost function changes depending on the form of the log p_{model}.

Back-Propagation

During training, the gradient of the cost function $\Delta_\theta J(\theta)$ is computed using a back-propagation algorithm [27-29] to allow information to flow backward through the network and compute the error made on each network weight. A gradient descent was then used to minimize the cost function. Learning was subsequently performed by updating the weights of the units. This procedure is described in the algorithm shown in Figure 1.

Training a neural network consists of applying a series of forwarding propagations—the network output is generated from the data through the network, and back-propagations compute the error at each unit. Each of these forward propagation and back-propagation combinations is called a pass. A pass of all the training examples is performed to compute the gradient used for the gradient-descent algorithm. A pass of every training example is called an epoch. At the end of each epoch, the network weights are updated using a learning rate hyperparameter, which is multiplied by the gradient calculated with back-propagation.

The learning rate is one of the most important hyperparameters for tuning in a neural network, as it controls the effective capacity of the network [26]. Therefore, it needs to be carefully optimized. If the learning rate is too large, the gradient descent can have the opposite of the desired effect, and training accuracy can decrease [30]. However, when it is too small, the training is slower, and sometimes the training accuracy can stay permanently small [30]. The number of epochs is also a hyperparameter that can be tuned ahead of the training.

Figure 1. Back-propagation algorithm.

Algorithm Back-propagation algorithm for a L-layer network with weights $\theta^{(l)}$ and a training set $\{(\mathbf{x_1}, \mathbf{y_1}), \ldots, (\mathbf{x_m}, \mathbf{y_m})\}$.

```
for l ← 1 to L do
    θ^(l) = small random value ;  // Initialise network weights for each layer
end
foreach epoch do
    for l ← 1 to L do
        Δ^(l) = 0 ;                // Initialise gradient matrices
    end
    // For each training example
    foreach (x_i, y_i) ∈ {(x_1, y_1), ..., (x_m, y_m)} do
        // Forward propagation
        w^(1) ← x_i;
        for l ← 2 to L do
            w^(l) ⇐ g(θ^(l−1) w^(l−1)) ;   // For each layer of the network
        end
        // Back-propagation
        δ^(L) ← w^(L) − y_i ;   // Compute the error at the output layer
        for l ← L − 1 to 2 do
            δ^(l) ⇐ ((θ^(l))^T δ^(l)) .* w^(l) .* (1 − w^(l)) ;   // Compute the error of each unit at the hidden layers
            Δ^(l) ← Δ^(l) + δ^(l) (w^(l))^T ;   // Update the matrix Δ for each layer
        end
    end
    // Gradient-descent: Update weights using learning rate η and gradient (1/m)Δ
    for l ← 1 to L do
        θ^(l) ← θ^(l) − η (1/m) Δ^(l)
    end
end
return θ^(1), ..., θ^(L);
```

Methods

Overview

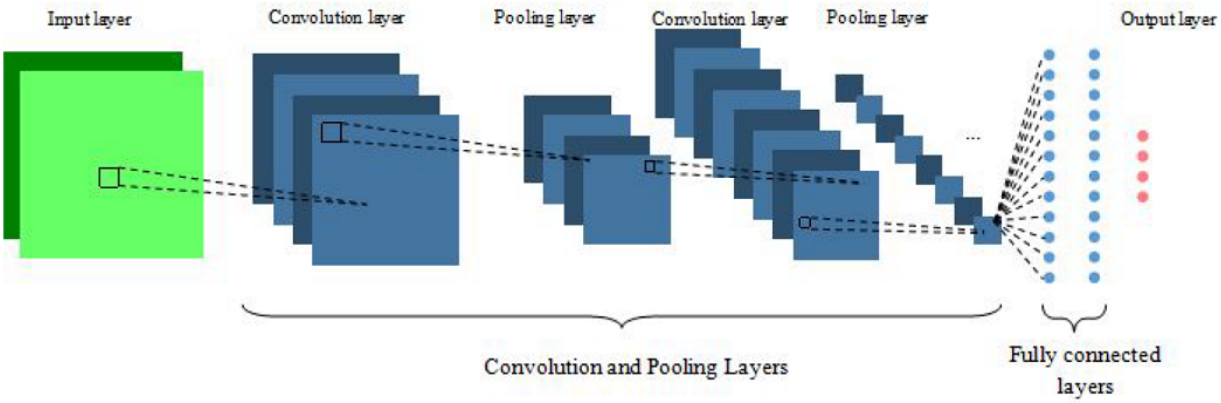
As previously mentioned, the research community is now focusing on solving practical applications using deep learning approaches. Our proposed solution to the problem of diagnosing colorectal and prostate cancer is to apply a deep learning framework.

CNNs [27,31] are a type of neural network that specialize in data with a grid-like topology. They are particularly adapted for image processing. Similar to conventional neural networks, they consist of units with weights and biases that are learned during training. However, with the assumption of the data topology, it is possible to add some properties to the architecture to reduce the number of parameters to learn and improve the network implementation efficiency. These key ideas are local

connections, shared weights, pooling, and the use of many layers [32].

The CNN units are arranged in three dimensions in each layer of the network: width, height, and depth of the activation volume. As depicted in Figure 2, a total of 3 different types of layers are usually stacked to form the full CNN architecture: convolutional layer, pooling layer, and fully connected layer. Fully connected layers are layers of a traditional MLP, as described in the section *Feedforward Neural Networks*.

Figure 2. Convolutional neural network architecture.



Convolutional Layer

The convolutional layer is the core layer of a CNN. The basic idea is that instead of connecting a unit to every unit of the previous layer, it is only connected to a local region of the previous layer. The spatial extent of this connection is called the receptive field of the unit or filter size. This is a hyperparameter of the model. The filter size along the depth axis is the same as that of the previous layer. This shows an asymmetry in the way spatial dimensions (width and height) and the depth dimension are treated, making the network particularly adapted for multispectral images. The connectivity of the convolutional layer is local along the width and height, but the layer is fully connected along with depth.

A convolutional layer's parameters can also be seen as a set of spatially small-sized learnable filters or kernels. During the forward pass, the filters are convolved across the width and height dimensions of the input volume. This action produces a 2D activation map outputting the responses of the filter at each position of the input layer [26,32]. The output volume of a convolutional layer depends on three hyperparameters: the number of filters, the stride, and zero padding.

The number of filters in the same receptive field determines the depth of the output volume. A different filter activates for every different pattern. A set of units with the same receptive field is called the breadth of the output layer.

The stride used when the filters are slid along the spatial dimensions of the previous layer affects the height and width of the output volume. The higher the stride, the smaller is the output volume.

The input volume can be padded with zeros around the border to keep the information at the border. Without zero padding, the information carried by the pixels at the border of the input image vanishes quickly after successive convolutional layers. This artificially increases the size of the input layer, thereby increasing the size of the output layer.

Furthermore, the parameter-sharing scheme is used to reduce the number of parameters to be learned. It is based on the assumption that a useful feature at one position of the input layer is also useful at a different position. This means that the units on the same output depth slice use the same weights and biases. This explains the fact that the forward propagation through a convolutional layer is equivalent to convoluting a filter or kernel with the input layer.

Pooling Layer

Typically, a pooling layer is inserted between the successive convolution layers. The pooling function replaces the output of a convolutional layer at a certain unit with the statistic of its neighboring units. The most popular pooling function used is the max-pooling method introduced by Zhou et al [33]. The pooling layer aims to make the system invariant to small input translations. This property gives more importance to whether a feature is present in the input rather than its exact position.

CNN Feature Extraction and Classification

The combination of convolutional and pooling layers aims to learn the best features that can be extracted from the data set. This contrasts with most current methods that use handcrafted feature extraction techniques, such as those presented in the previous sections. These approaches can yield very good results but are usually sensitive to the data set and perform poorly when applied to different data sets. The combination of convolutional and pooling layers of a CNN provides a more versatile method for extracting features from images. The fully connected layers of the CNN correspond to the classifier. It aims at learning to classify learned features. As a result, a CNN is a unified versatile scheme for feature extraction and classification. As medical image classification is often a very complex task, it requires carefully manufactured feature sets for each type of data or even each different data set; doing just that with a unified framework, CNNs seem particularly adapted to the field.

Data Set Description

The prostate gland and the colorectum have a similar tissue structure, with the tubular glandular mucosa—composed of epithelium and lamina propria—being their main functional tissue. This characteristic implies that these tissues are subject to development of the same types of tumors and cancers. Carcinomas are the most common type of malignant tumor and they are derived from epithelial cells [34]. Carcinomas are called adenocarcinomas when derived from glandular tissues, which is the case for both organs studied in this paper. All growths are not necessarily malignant, and benign polyps can occur [35]. They are usually noncancerous growths of the mucosa into the lumen and can be of different types.

Although most polyps are completely benign, such as hyperplastic polyps or hyperplasia, some types of polyps can transform into adenocarcinoma and can be considered as a precancerous stage. They are called adenomas and can be tubular or villous, depending on their growth patterns [36]. Hyperplastic polyps are characterized by an increase in the number of cells, resulting in an increased size of the tissue because of enhanced cell division. In contrast to an adenoma or a carcinoma, the division rate in a hyperplastic polyp returns to normal as soon as the stimulus is removed.

To best describe the different types of tumor recognized by pathologists, the following two data sets were used for the purpose of this study:

1. The prostate data set, which was used in previous works by Tahir and Bouridane [13] and Peyret et al [17], consists of 512 different multispectral prostate tumor tissue images of size 128×128. The images were taken at 16 spectral channels (500-650 nm) and 40× magnification power. The samples were evaluated by 2 highly experienced independent pathologists and labeled into four classes: 128 cases of stroma, which is normal muscular tissue, 128 cases of benign prostatic hyperplasia, a benign condition, 128 cases of prostatic intraepithelial neoplasia, a precancerous stage, and 128 cases of prostatic carcinoma, an abnormal tissue development corresponding to cancer.
2. The colorectal data set, which consists of multispectral colorectal histology data with a 40× magnification power, was developed by the University of Qatar in collaboration with Al-Ahli Hospital, Doha. It splits into 4 classes, each composed of 40 images. The images were acquired on a wider spectrum than the first data set, as it was spread on the visible and infrared ranges of the electromagnetic spectrum with an interval of 23 nm between each wavelength. That is to say, in the visible range, the wavelength interval is 23 nm starting from 465 to 695 nm, and in the infrared range, the wavelength interval is also 23 nm and ranges from 900 to 1590 nm. The special size was 128×60 pixels. The 4 classes were defined as carcinoma, containing images of cancerous colon biopsies; tubular adenoma, a precancerous stage; hyperplastic polyp, a benign polyp; and no remarkable pathology.

Experiments

Hardware and Software Specifications

To train deep CNNs, a GPU is required. The system used for this experiment was equipped with 1 NVIDIA K80 GPU and 4 central processing units. It had 61-GB RAM. Regarding software, Keras with a TensorFlow backend was used. Keras has the advantage of making available deep learning models alongside pretrained weights.

Selected Architecture

The proposed CNN architecture evaluated for the task at hand was based on Visual Geometry Group 16 (VGG16) [37]. To design the proposed architecture, the last block of the convolutional layers of VGG16 was removed, and the number of filters per layer was halved. The idea is to reduce the capacity of the network because the interclass similarity in the data sets used for the task was high compared with the data set on which VGG16 was tested.

As represented in Figures 3 and 4, the overall proposed network architecture consists of a total of 13 layers with weights—the first 10 being convolutional layers, and the remaining 3 fully connected layers. The output of the last fully connected layer was fed to a SoftMax classifier, which is a generalization of the logistic regression classifier to the multiclass problem and produces a distribution of the 4 class labels. The network uses cross-entropy as a loss function.

Similar to VGG16, we decided to use a small kernel with a size of 3 pixels for every convolutional layer. The strategy of stacking convolutional layers with a small filter size is preferred to using a single large receptive eld convolutional layer. For the same final receptive field, the former strategy includes nonlinearities (ReLU functions) at each layer, whereas the latter computes a simple linear function on the input, which makes the features less expressive. A stride of 1 was also adopted for the entire network to minimize information loss.

To achieve better control over the output size of each layer and maintain border information, a zero padding of 1 is added before each convolutional layer. The first 2 convolutional layers use 32 kernels followed by a 2 2 max-pooling layer. The max-pooling layer reduces the size of the output and thus the network capacity. The number of kernels is doubled in the next convolutional layer to compensate for this loss. Consequently, this sequence is followed by 2 convolutional layers with 64 filters, and then a new max-pooling layer is applied. This is followed by a series of 3 convolutional layers with 128 filters and a max-pooling layer. A final series of 3 convolutional layers with 256 filters and a max-pooling layer was applied. The neurons in the 3 fully connected layers with sizes of 1024, 1024, and 4, respectively, are connected to all neurons in the previous layer. The ReLU nonlinearity was applied to the output of every layer with weights.

Dropout is used after every max-pooling and fully connected layer to reduce overfitting. An early stopping strategy is also adopted to reduce the training time and regularization. Finally, data augmentation is carried out using the following transformations: each image is flipped along the 2 special axes,

and 30 rotations in both directions are applied. This results in the generation of 27 fake images for each real data image. To ensure that the generalization is not overestimated, data set augmentation is performed after splitting the data set into training and test sets.

Figure 3. Illustration of the architecture of the proposed convolutional neural network for prostate cancer images. ReLU: rectified linear unit.

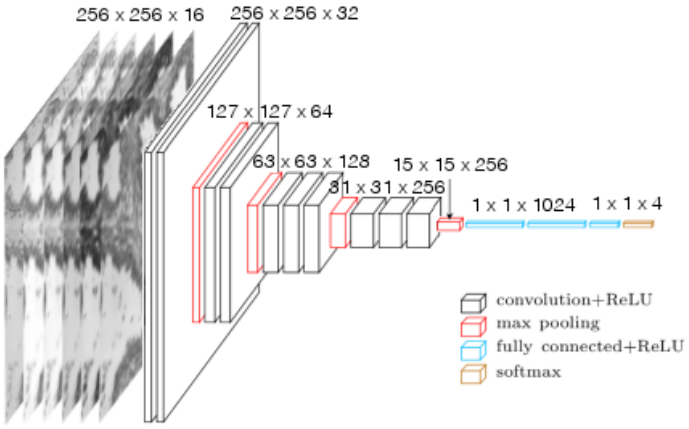
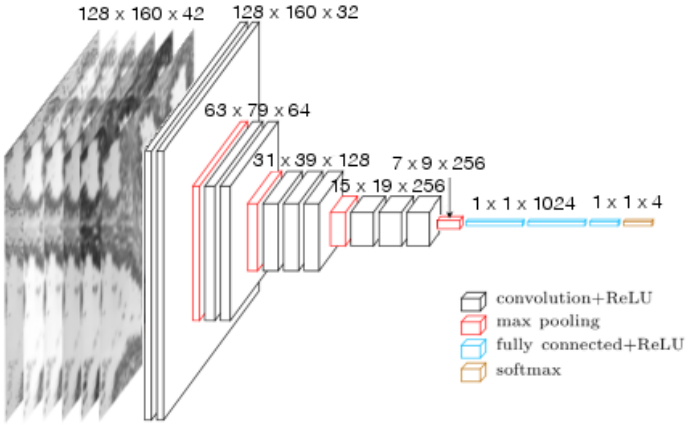


Figure 4. Illustration of the architecture of the proposed convolutional neural network for colorectal cancer images. ReLU: rectified linear unit.



Details of Learning

The weights of each layer are initialized using the Xavier initialization method [38], where the weights are drawn from a normal distribution centered on zero and with an SD of the following:

where N_{in} and N_{out} are the numbers of input and output units, respectively. The network was trained separately on the 2 data sets.

The learning rate used was the same for all layers. It is optimized using a grid-search scheme, the results of which are presented in Figures 5 and 6. The learning rate selected for training was 0.0001 for both data sets.

For each model training, a 10-fold cross-validation technique was adopted to obtain a good estimate of the systems' generalization accuracy. This provides a large training set for better learning.

Figures 7 and 8 illustrate the evolution of the loss function during training for the prostate and colorectal data sets, respectively. Figures 9 and 10 show the evolution of their accuracies. It can be observed from these figures that the validation accuracy is very close to the training accuracy, which proves that the model is not in the overfitting regime. The higher variation in validation accuracy and loss can be explained by the smaller set used for validation compared with that used for training.

Figure 5. Validation accuracy obtained with different learning rates for the network trained on prostate data.

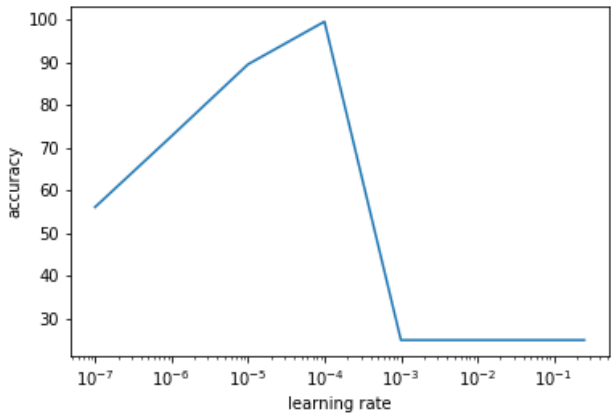


Figure 6. Validation accuracy obtained with different learning rates for the network trained on colorectal data.

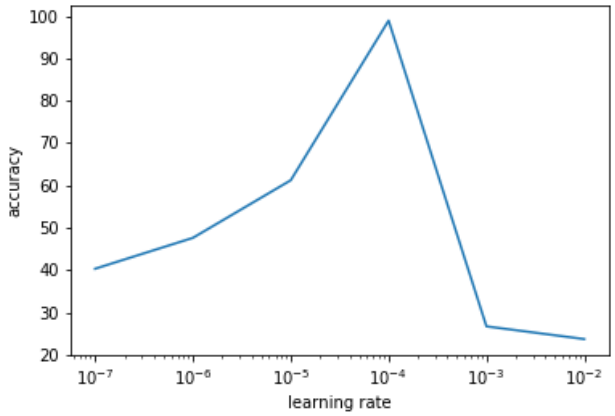


Figure 7. Loss function evolution during training for the prostate data set.

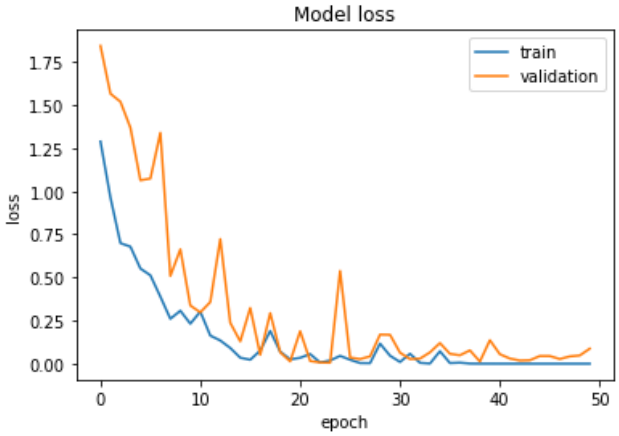


Figure 8. Loss function evolution during training for the colorectal data set.

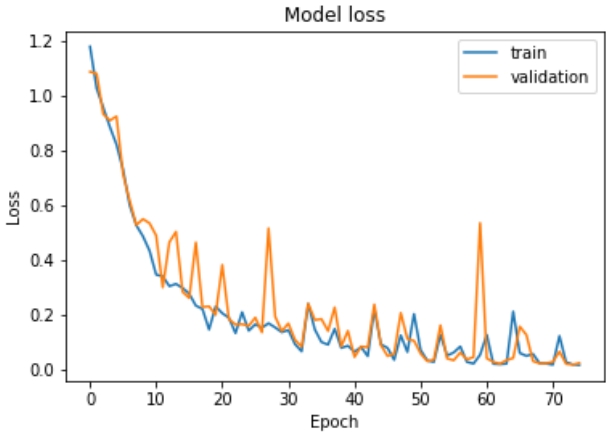


Figure 9. Accuracy evolution during training for the prostate data set.

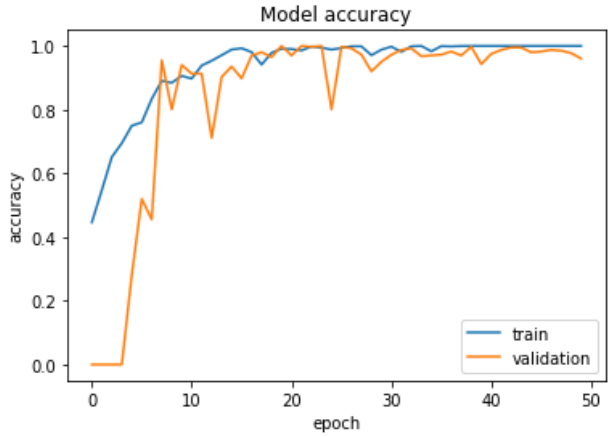
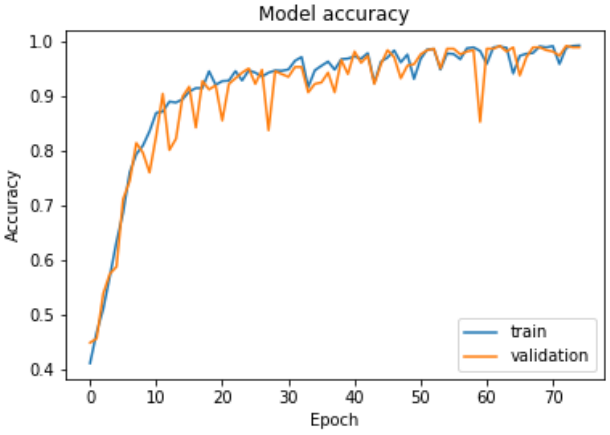


Figure 10. Accuracy evolution during training for the colorectal data set.



Transfer Learning

Transfer learning consists of using a network previously trained on another data set to use the knowledge acquired during this learning task for the new task at hand [39]. In most transfer learning for image classification tasks, the ImageNet data set [40], which contains 1.2 million images with 1000 categories, is used for pretraining the network. When only a small data set is available, this allows the CNN to be trained on a very large data set and therefore train a high-capacity network that captures details without overfitting. Very deep networks also require a lot of time and very powerful machines equipped with multiple GPUs. Using pretrained networks can be advantageous when appropriate resources are not provided. Several transfer-learning scenarios are practical.

In the first scenario, the pretrained CNN is used as a fixed feature extractor. The convolutional layers of the network are kept with the weights determined during training on the ImageNet data set, and the pretrained fully connected layers are replaced with fully connected layers initialized with random

weights. During training, only the newly added fully connected layers were marked as trainable. They used the features extracted by pretrained convolutional layers as inputs. These features are usually referred to as CNN codes [26,39].

Another strategy is to retrain the fully connected layers from scratch to fine-tune the weights of the pretrained convolutional layers by continuing back-propagation. Either all the convolutional layers can be retuned or only some of the higher-level layers to avoid overfitting. This derives from the observation that the lower-level layers usually learn more generic features, such as edge detectors, that can be used for many different learning tasks. In contrast, the high-level layers tend to learn features that are more specific to the characteristics of the classes of the original data set.

In this study, only the first scenario was investigated. The pretrained CNNs are very deep and require very high computational power to be retuned. Using them as feature extractors is, in fact, equivalent to training only a relatively shallow MLP. The proposed architecture was compared with popular CNN architectures: VGG16 [37], InceptionV3 [21], and ResNet50 [22]. These networks were initialized with the weights obtained when pretraining them on the ImageNet data set. However, InceptionV3 and ResNet50 are very deep networks (48 and 152 layers, respectively), and a minimum input image size is required. InceptionV3 requires a minimum width and height of 139 pixels and ResNet50 of 197 pixels. The images of the colorectal data set were smaller, and zero padding was added to reach the required dimensions. Moreover, the ImageNet images are RGB images and therefore have a depth of 3 channels. To meet the dimension requirements, a principal component analysis (PCA) was carried out to reduce the dimensionality of the multiscale images to 3 channels.

Results and Discussion

Principal Results and Findings

To visualize the effect of the kernels on images through the network, Figures 11 and 12 present examples of outputs of the first convolutional layer of the networks trained with the prostate and colorectal data sets, respectively. Figures 13 and 14 depict examples of outputs of the last convolutional layers of the same networks. It can be observed that after the first layer, the outputs are very similar to the input image, for instance, with transformations resembling edge detections. Once the image has its own through the network, different regions or features of the input image are represented in the outputs of the last convolutional layer. Thus, the different layers learn a succession of transformations, leading to an isolation of relevant regions or features of the input image. The fully connected layers of the network are then able to classify these particular features into the 4 classes.

Figure 11. Example of an output of the first convolutional layer for the network trained on the prostate data set.

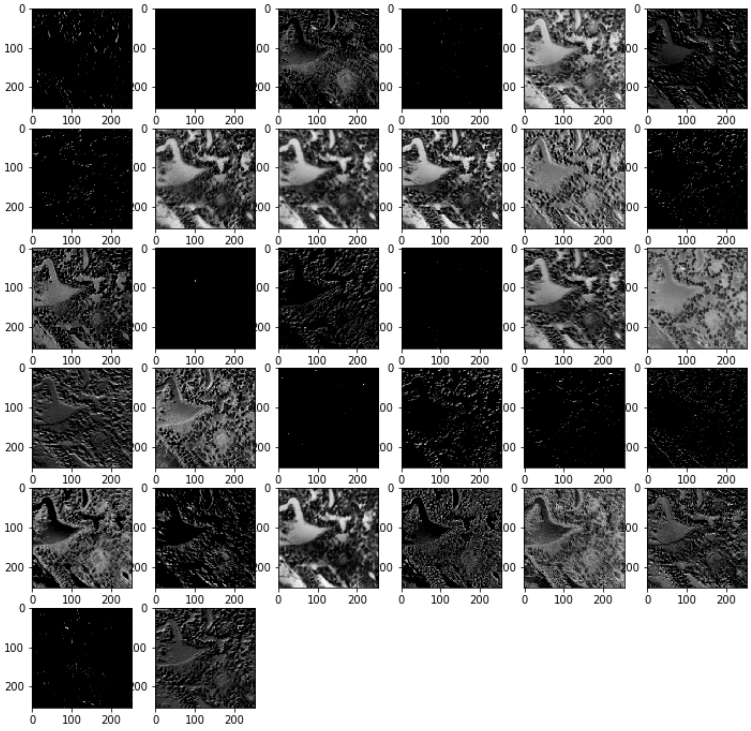


Figure 12. Example of an output of the first convolutional layer for the network trained on the colorectal data set.

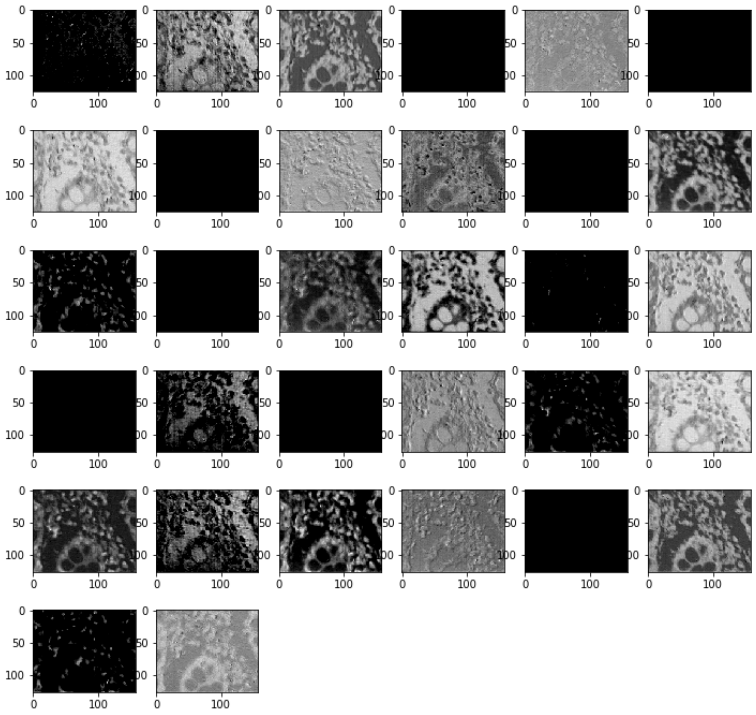


Figure 13. Example of an output of the last convolutional layer for the network trained on the prostate data set.

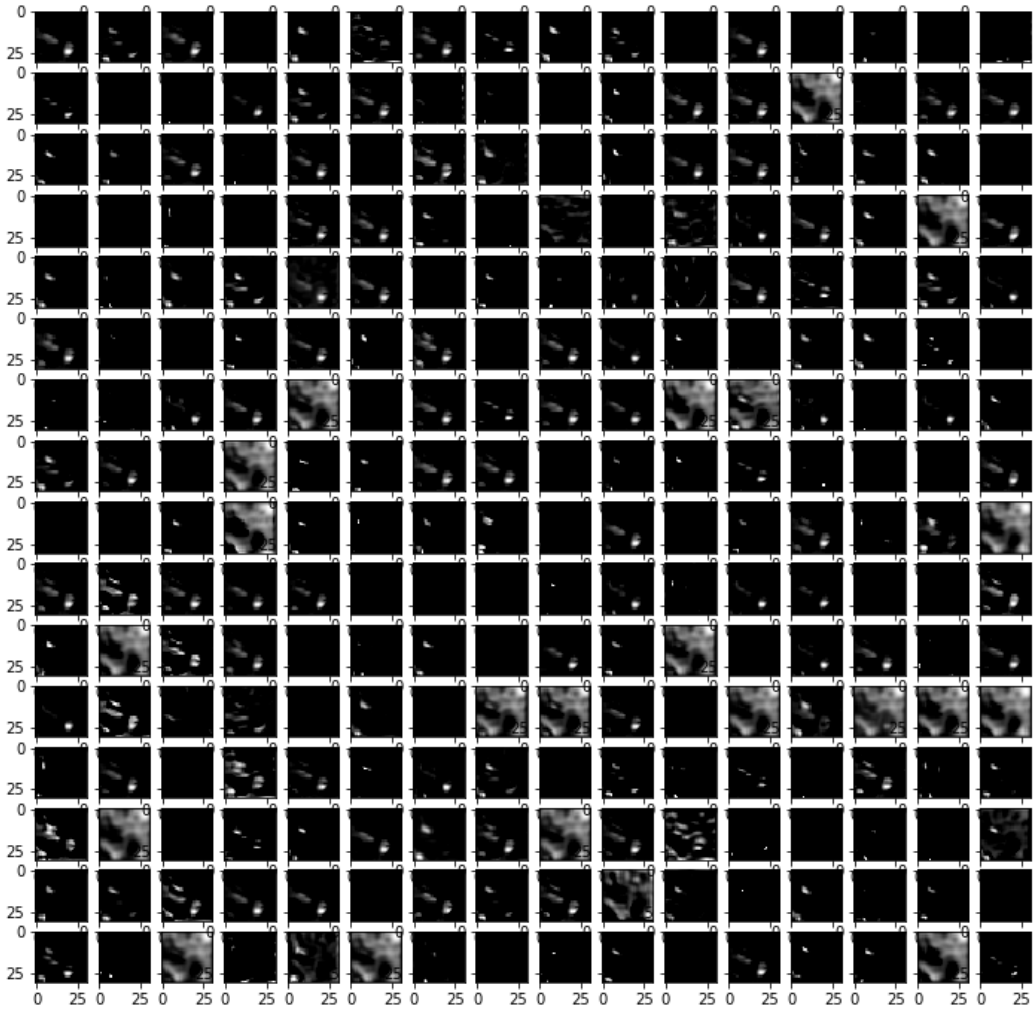


Figure 14. Example of an output of the last convolutional layer for the network trained on the colorectal data set.

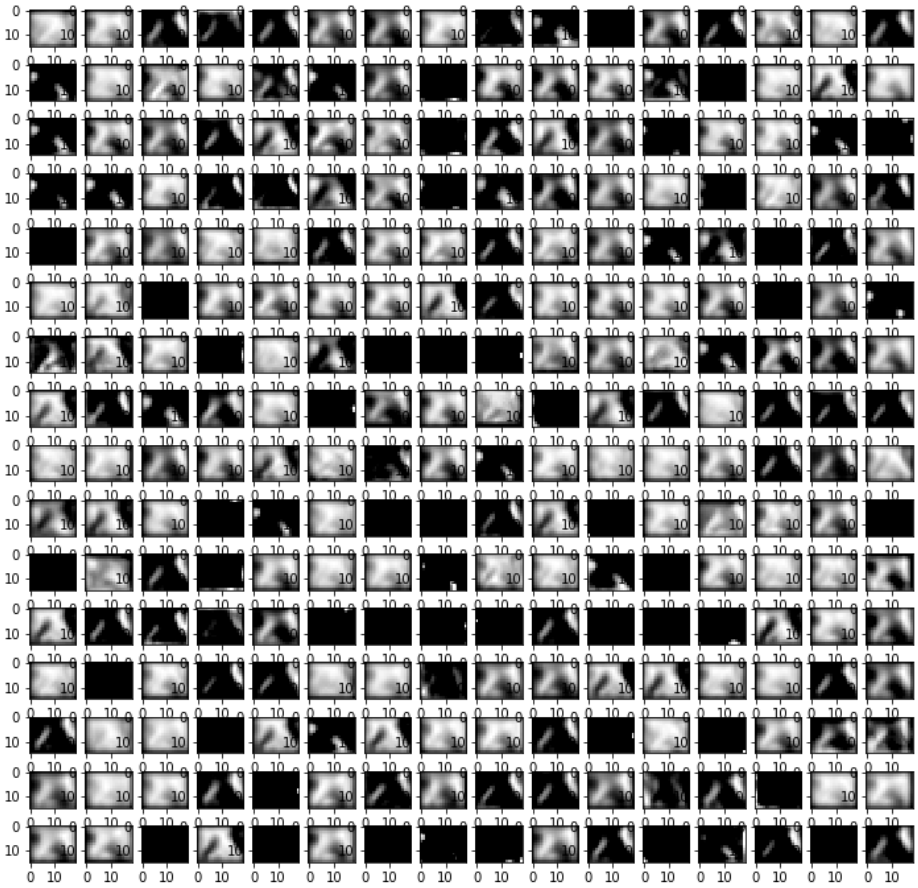


Table 1 displays the validation and test accuracies obtained using the prostate and colorectal data sets for different CNN models. This shows that the validation and test accuracies are very close, proving a good generalization of the systems and that overfitting was avoided.

The proposed CNN model achieved an average test accuracy of 99.8% and 99.5% for the prostate and colorectal data sets, respectively. Table 2 shows that the optimal CNN weights were obtained after 44 and 70 epochs, respectively. The VGG16 model initialized with Xavier weights trains very quickly for the prostate data set; the optimal validation accuracy was obtained after 19 epochs, as illustrated in Table 2. However, it is less efficient at learning for the colorectal data set and requires as many as 70 epochs to obtain the minimum validation loss. The results also show slight overfitting for the colorectal data set, as the validation accuracy is lower than the training accuracy. This is because of the high capacity of the network. When using this network with pretrained weights from ImageNet, the training loss reaches a minimum after only a few epochs, but the validation loss shows that the network overfits marginally for both data sets. The test accuracy was also lower than that of the proposed CNN by 99.5% and 98.1%, respectively. This is because the CNN codes learned with the ImageNet data set are not as adapted to the classification task at hand as those learned with the proposed CNN. The InceptionV3 model shows a higher overfitting and a lower generalization for both data sets with 99.0% and 94.5% accuracy for the prostate and colorectal data sets, respectively. This shows once again that the CNN codes learned on the ImageNet data set with this network are not adapted to the classification task at hand. Finally, the pretrained ResNet50 achieved optimal accuracy with the lowest number of epochs: 5 and 22 for the prostate and colorectal data sets, respectively. It also achieves 100% average accuracy for the prostate data set, outperforming the proposed CNN, and 99% for the colorectal data set, which is slightly lower than the proposed data set. This lower performance compared with the proposed CNN architecture for the colorectal data set might be owing to some loss of information when performing PCA on the 42 channels of the colorectal data set images. The prostate data set consisted of images with only 16 channels, and it is logical that the loss of information is not as important during this transformation.

Therefore, the proposed CNN architecture is more adapted to the task at hand than the other methods it was compared with. However, ResNet50 shows very good performance when used as a feature extractor and is trained with fewer epochs. In every case, it was noted that the colorectal data set is more prone to overfitting. This is probably owing to the size of the images, which are spatially smaller than those for the prostate data set. Therefore, a model with the correct capacity for the prostate data set might be overestimated for the colorectal data set.

Table 1. Validation and test accuracy comparison of different architectures.[a]

Method	Prostate data set (% accuracy), mean (SD)		Colorectal data set (% accuracy), mean (SD)	
	Validation	Test	Validation	Test
Proposed CNN[b]	100	99.8 (0.1)	100	99.5 (0.1)
VGG16[c] Xavier initial	100	99.6 (0.1)	99.0 (0.1)	99.2 (0.1)
VGG16 pretrained	100	99.5 (0.1)	97.5 (0.2)	98.1 (0.1)
InceptionV3 pretrain	98.8 (0.2)	99.0 (0.1)	92.3 (0.3)	94.5 (0.3)
ResNet50 pretrained	100	100	99.5 (0.1)	99.0 (0.2)

[a]SD values have been provided wherever applicable.

[b]CNN: convolutional neural network.

[c]VGG16: Visual Geometry Group 16.

Table 2. Number of epochs until early stopping.

Method	Prostate data set	Colorectal data set
Proposed CNN[a]	44	70
VGG16[b] Xavier initialization	19	70
VGG16 pretrained	10	38
InceptionV3 pretrained	48	53
ResNet50 pretrained	5	22

[a]CNN: convolutional neural network.

[b]VGG16: Visual Geometry Group 16.

Comparison Against Other Machine Learning Methods

Table 3 shows the test accuracy of the best-performing CNN architectures compared with other methods from Tahir et al [15], Bouatemane et al [16], Haj-Hassan et al [25], and Peyret et al [17] stacked multispectral multiscale local binary pattern (MMLBP) + gray-level co-occurrence matrix (GLCM), and concatenated local binary pattern [18]. Regarding the prostate data set, 5 systems have an accuracy above 99%: Bouatemane et al [16], Stacked MMLBP+GLCM, the proposed CNN, Haj-Hassan et al [25], and ResNet50 with pretrained weights. The highest classification accuracy was achieved using ResNet50 with 100% accuracy. The proposed CNN and the study by Bouatemane et al [16] achieved 99.8% accuracy; however, the SD was not given for the latter. Therefore, it is not possible to determine the precision of the accuracy estimation. The stacked MMLBP+GLCM system achieves 99.5% (SD 0.3 pp), which makes this performance similar to that of the proposed CNN. However, a higher SD shows lower precision in the accuracy estimation. Therefore, the proposed CNN was preferred. The study by Haj-Hassan et al [25] achieved a 99.17% accuracy with segmentation. Their system without this preprocessing phase achieved an accuracy of 79.23%.

This can be explained by the lower capacity of their model compared with ours. This has the advantage of requiring less processing power. However, this is counterbalanced by the fact that their system requires a preprocessing phase with the intervention of a pathologist, which dramatically increases the processing time of the system. Furthermore, they state that their CNN model requires 500 epochs to be trained, which is much higher than that of the proposed model. With respect to the colorectal data set. Peyret et al [17] stacked MMLBP+GLCM system and the proposed CNN both provided the same accuracy and SD. They outperform ResNet50 with pretrained weights by 0.5 pp.

Finally, when considering the results obtained with both data sets, the stacked MMLBP+GLCM system and the proposed CNN appear to provide the most stable results as well as the highest accuracy. However, on average, the SD of the accuracy achieved by the proposed CNN is lower than that obtained with the stacked MMLBP+GLCM system. The performance of the ResNet50 network seems to be more dependent on the data set used. Moreover, it would be interesting to compare the system proposed by Bouatemane et al [16] using the colorectal data set to verify whether it performs as well on different data sets. Considering the current information available on the system performance and with the data sets available, the proposed CNN is selected as the best-performing system in terms of accuracy for the classification task at hand.

Table 3. Accuracy comparison against other methods.[a]

Method	Prostate data set (% accuracy), mean (SD)	Colorectal data set (% accuracy), mean (SD)
Tahir et al [15]	• 98.9	• N/A[b]
Bouatemane et al [16]	• 99.83	• N/A
Concatenated LBP[c] [18]	• 92.4 (0.4) • 99.5 (0.3)	• 88.2 (0.5) • 99.5 (0.1)
Stacked MMLBP[d] + GLCM[e] [41]	• 99.5 (0.3)	• 99.5 (0.1)
Haj-Hassan et al [25]	• 99.17	• N/A
Proposed CNN	• 99.8 (0.1)	• 99.5 (0.1)
ResNet50 pretrained	• 100	• 99.0 (0.2)

[a]SD values have been provided wherever applicable.

[b]N/A: not applicable.

[c]LBP: local binary pattern.

[d]MMLBP: multispectral multiscale local binary pattern.

[e]GLCM: gray-level co-occurrence matrix.

Computational Complexity Analysis

In computer-aided diagnosis systems (CADSs), an unlabeled image is fed to a previously trained system. Consequently, the time used to process this image is decisive, as it is crucial that the CADS works on the web. However, the forward pass of an image through the CNN architectures studied in this study is computationally nonexpensive. Table 4 displays the classification times per image for all CNN architectures tested. This demonstrates that only a few milliseconds are required to classify one image once the CNN has been trained. However, it must be noted that the proposed CNN architecture is much quicker at classifying images than the others. This is because, for the architectures described in the literature and the pretrained networks, a PCA must be carried out to reduce to 3 the number of channels of the image to be classified. This preprocessing stage lengthens the total classification time.

As mentioned, training is performed only once when a CADS is created. Consequently, training time is not a critical measure of the problem at hand. However, the computational complexity of deep learning systems can rapidly increase significantly. Such architectures require high-performing hardware, including GPUs. Some extremely deep architectures can also entail several weeks of training time [26]. Such long training times considerably slowed down the CADS development process. To verify that the proposed system can be trained within a reasonable duration, a comparison of the training times for each architecture was carried out (Table 5). The computational times depending on the hardware and software used, it is not possible to compare the CNN architectures with other classification systems proposed in other published works. However, this is one of the first attempts to use deep learning for this application. Therefore, this section aims to establish the ability of deep learning systems to be trained in a short period using the data sets used.

Unsurprisingly, Table 5 demonstrates that pretrained networks have a much shorter training time per epoch owing to the reduced number of layers to be trained; ResNet50 and InceptionV3 can be trained in a few minutes. When considering this measure of performance, the best architecture was ResNet50. However, the total training time for every CNN model is <2 hours, making it a reasonable time for developing a CADS.

Table 4. Average convolutional neural network (CNN) classification computation times for 1 image.

Method	Prostate data set (ms)	Colorectal data set (ms)
Proposed CNN	14	7
VGG16[a] Xavier initial	75	42
VGG16 pretrained	75	42
InceptionV3 pretrained	63	42
ResNet50 pretrained	65	47

[a]VGG16: Visual Geometry Group 16.

Table 5. Average convolutional neural network (CNN) training computation times for the complete data set.

Method	Prostate data set (seconds)		Colorectal data set (seconds)	
	Time per epoch	Total training	Time per epoch	Total training
Proposed CNN	90	3780	45	2925
VGG16[a] Xavier initial	245	4655	97	6790
VGG16 pretrained	83	3154	35	1400
InceptionV3 pretrained	39	1755	15	705
ResNet50 pretrained	41	205	32	704

[a]VGG16: Visual Geometry Group 16.

Conclusions

In this paper, the proposed CNN architecture was detailed and compared with previously trained network models used as feature extractors. These CNNs were also compared with other classification methods from other published studies. The proposed CNN demonstrated excellent performance compared with pretrained CNNs and other classification methods. The computational complexity of the CNNs was also analyzed, and it was demonstrated that the proposed CNN is faster at classifying images than pretrained networks because it avoids a preprocessing phase. The conclusion of this overall analysis is that the proposed CNN architecture was globally the best-performing system for classifying colorectal and prostate tumor images.

Acknowledgments

This research project was supported by a grant from the Research Supporting Program (Project Number: RSP2022R281), King Saud University, Riyadh, Saudi Arabia.

Conflicts of Interest

None declared.

References

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 2015 Mar 01;136(5):E359-E386 [FREE Full text] [doi: 10.1002/ijc.29210] [Medline: 25220842]
2. Heidenreich A, Bellmunt J, Bolla M, Joniau S, Mason M, Matveev V, European Association of Urology. EAU guidelines on prostate cancer. Part 1: screening, diagnosis, and treatment of clinically localised disease. Eur Urol 2011 Jan;59(1):61-71. [doi: 10.1016/j.eururo.2010.10.039] [Medline: 21056534]
3. Humphrey P. Prostate Pathology. Chicago: American Society for Clinical Pathology; 2003.
4. Thomas GD, Dixon MF, Smeeton NC, Williams NS. Observer variation in the histological grading of rectal carcinoma. J Clin Pathol 1983 Apr;36(4):385-391 [FREE Full text] [doi: 10.1136/jcp.36.4.385] [Medline: 6833507]
5. Kronz JD, Westra WH, Epstein JI. Mandatory second opinion surgical pathology at a large referral hospital. Cancer 1999 Dec 01;86(11):2426-2435. [Medline: 10590387]
6. Tobore I, Li J, Yuhang L, Al-Handarish Y, Kandwal A, Nie Z, et al. Deep learning intervention for health care challenges: some biomedical domain considerations. JMIR Mhealth Uhealth 2019 Aug 02;7(8):e11966 [FREE Full text] [doi: 10.2196/11966] [Medline: 31376272]
7. Owais M, Arsalan M, Mahmood T, Kang JK, Park KR. Automated diagnosis of various gastrointestinal lesions using a deep learning-based classification and retrieval framework with a large endoscopic database: model development and validation. J Med Internet Res 2020 Nov 26;22(11):e18563 [FREE Full text] [doi: 10.2196/18563] [Medline: 33242010]
8. Zhao Z, Wu C, Zhang S, He F, Liu F, Wang B, et al. A novel convolutional neural network for the diagnosis and classification of rosacea: usability study. JMIR Med Inform 2021 Mar 15;9(3):e23415 [FREE Full text] [doi: 10.2196/23415] [Medline: 33720027]
9. Mosquera-Lopez C, Agaian S, Velez-Hoyos A, Thompson I. Computer-aided prostate cancer diagnosis from digitized histopathology: a review on texture-based systems. IEEE Rev Biomed Eng 2015;8:98-113. [doi: 10.1109/RBME.2014.2340401] [Medline: 25055385]
10. Kunhoth S, Al Maadeed S. Multispectral biopsy image based colorectal tumor grader. In: Valdés Hernández M, González-Castro V, editors. Medical Image Understanding and Analysis. Cham: Springer; 2017:330-341.
11. Roula M, Diamond J, Bouridane A, Miller P, Amira A. A multispectral computer vision system for automatic grading of prostatic neoplasia. In: Proceedings IEEE International Symposium on Biomedical Imaging. 2002 Presented at: Proceedings

IEEE International Symposium on Biomedical Imaging; Jul 7-10, 2002; Washington, DC, USA. [doi: 10.1109/ISBI.2002.1029226]
12. Roula MA. Machine vision and texture analysis for the automated identification of tissue pattern in prostatic neoplasia. PhD Thesis. Belfast, Northern Ireland: Queen's University of Belfast; 2004.
13. Tahir M, Bouridane A. Novel round-robin tabu search algorithm for prostate cancer classification and diagnosis using multispectral imagery. IEEE Trans Inf Technol Biomed 2006 Oct;10(4):782-793. [doi: 10.1109/titb.2006.879596] [Medline: 17044412]
14. Tahir M, Bouridane A, Kurugollu F. Simultaneous feature selection and feature weighting using Hybrid Tabu Search/K-nearest neighbor classifier. Pattern Recog Letters 2007 Mar;28(4):438-446 [FREE Full text] [doi: 10.1016/j.patrec.2006.08.016]
15. Tahir M, Bouridane A, Roula M. Prostate cancer classification using multispectral imagery and metaheuristics. In: Computational Intelligence in Medical Imaging. London, United Kingdom: Chapman and Hall; 2009.
16. Bouatmane S, Roula M, Bouridane A, Al-Maadeed S. Round-Robin sequential forward selection algorithm for prostate cancer classification and diagnosis using multispectral imagery. Mach Vision Apps 2010 Sep 16;22(5):865-878. [doi: 10.1007/s00138-010-0292-x]
17. Peyret R, Khelifi F, Bouridane A, Al-Maadeed S. Automatic diagnosis of prostate cancer using multispectral based linear binary pattern bagged codebooks. In: Proceedings of the 2nd International Conference on Bio-engineering for Smart Technologies (BioSMART). 2017 Presented at: 2nd International Conference on Bio-engineering for Smart Technologies (BioSMART); Aug 30 -Sep 1, 2017; Paris, France. [doi: 10.1109/biosmart.2017.8095322]
18. Peyret R, Bouridane A, Al-Maadeed S, Kunhoth S, Khelifi F. Texture analysis for colorectal tumour biopsies using multispectral imagery. In: Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2015 Presented at: 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); Aug 25-29, 2015; Milan, Italy. [doi: 10.1109/embc.2015.7320057]
19. Lasch P, Chiriboga L, Yee H, Diem M. Infrared spectroscopy of human cells and tissue: detection of disease. Technol Cancer Res Treat 2002 Feb;1(1):1-7 [FREE Full text] [doi: 10.1177/153303460200100101] [Medline: 12614171]
20. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017 May 24;60(6):84-90 [FREE Full text] [doi: 10.1145/3065386]
21. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015 Presented at: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Jun 7-12, 2015; Boston, MA URL: http://arxiv.org/abs/1409.4842 [doi: 10.1109/cvpr.2015.7298594]
22. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. arXiv. Preprint posted online December 10, 2015 [FREE Full text]
23. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. Nature 2016 Jan 28;529(7587):484-489. [doi: 10.1038/nature16961] [Medline: 26819042]
24. Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, et al. Mastering the game of Go without human knowledge. Nature 2017 Oct 18;550(7676):354-359. [doi: 10.1038/nature24270] [Medline: 29052630]
25. Haj-Hassan H, Chaddad A, Harkouss Y, Desrosiers C, Toews M, Tanougast C. Classifications of multispectral colorectal cancer tissues using convolution neural network. J Pathol Inform 2017;8:1 [FREE Full text] [doi: 10.4103/jpi.jpi_47_16] [Medline: 28400990]
26. Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge, Massachusetts, United States: MIT Press; 2016.
27. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE 1998 Nov;86(11):2278-2324. [doi: 10.1109/5.726791]
28. Rumelhart D, Durbin R, Golden R, Chauvin Y. Backpropagation: the basic theory. In: Backpropagation Theory, Architectures, and Applications. Mahwah: Lawrence Erlbaum Associates; 1995.
29. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 1986 Oct;323(6088):533-536. [doi: 10.1038/323533a0]
30. LeCun Y, Bottou L, Orr G, Müller K. Efficient BackProp. In: Neural Networks: Tricks of the Trade. Berlin, Heidelberg: Springer; 2012.
31. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, et al. Backpropagation applied to handwritten zip code recognition. Neural Comput 1989 Dec;1(4):541-551. [doi: 10.1162/neco.1989.1.4.541]
32. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015 May 28;521(7553):436-444. [doi: 10.1038/nature14539] [Medline: 26017442]
33. Zhou Y, Chellappa R, Vaid A, Jenkins B. Image restoration using a neural network. IEEE Trans Acoust Speech Signal Processing 1988 Jul;36(7):1141-1151. [doi: 10.1109/29.1641]
34. Lackie J. A Dictionary of Biomedicine. Oxford, UK: Oxford University Press; 2010.
35. Jass J, Sobin L. Histological classification of intestinal tumours. In: Histological Typing of Intestinal Tumours. Berlin, Heidelberg: Springer; 1989.

36. Understanding your pathology report: colon polyps (sessile or traditional serrated adenomas). American Cancer Society. URL: https://www.cancer.org/treatment/understanding-your-diagnosis/tests/understanding-your-pathology-report/colon-pathology/colon-polyps-sessile-or-traditional-serrated-adenomas.html [accessed 2022-01-12]
37. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv. Preprint posted online September 4, 2014 [FREE Full text]
38. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. 2010 Presented at: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics; May 13-15, 2010; Sardinia, Italy.
39. Transfer learning. CiteSeerX. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.146.1515 [accessed 2022-01-12]
40. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis 2015 Apr 11;115(3):211-252. [doi: 10.1007/s11263-015-0816-y]
41. Peyret R, Bouridane A, Khelifi F, Tahir M, Al-Maadeed S. Automatic classification of colorectal and prostatic histologic tumor images using multiscale multispectral local binary pattern texture features and stacked generalization. Neurocomputing 2018 Jan;275(C):83-93 [FREE Full text] [doi: 10.1016/j.neucom.2017.05.010]

Abbreviations

CADS: computer-aided diagnosis system
CNN: convolutional neural network
GLCM: gray-level co-occurrence matrix
GPU: graphic processing unit
MLP: multilayer perceptron
MMLBP: multispectral multiscale local binary pattern
PCA: principal component analysis
ReLU: rectified linear unit
RGB: red, green, blue
VGG16: Visual Geometry Group 16

Edited by A Mavragani; submitted 23.01.21; peer-reviewed by M Wu, SM Mir Hosseini; comments to author 25.06.21; revised version received 08.09.21; accepted 11.12.21; published 09.02.22.

Please cite as:
Peyret R, alSaeed D, Khelifi F, Al-Ghreimil N, Al-Baity H, Bouridane A
Convolutional Neural Network–Based Automatic Classification of Colorectal and Prostate Tumor Biopsies Using Multispectral Imagery: System Development Study
JMIR Bioinform Biotech 2022;3(1):e27394
URL: https://bioinform.jmir.org/2022/1/e27394
doi:10.2196/27394
PMID:

Original Paper

Seasonality of Hashimoto Thyroiditis: Infodemiology Study of Google Trends Data

Robert Marcec[1], MD; Josip Stjepanovic[1], MD; Robert Likic[1], MD, PhD

Department of Internal Medicine, University of Zagreb School of Medicine and Clinical Hospital Centre Zagreb, Zagreb, Croatia

Corresponding Author:
Robert Likic, MD, PhD
Department of Internal Medicine
University of Zagreb School of Medicine and Clinical Hospital Centre Zagreb
Kispaticeva 12
Zagreb, 10000
Croatia
Phone: 385 12388288
Fax: 385 12388144
Email: robert.likic@mef.hr

Abstract

Background: Hashimoto thyroiditis (HT) is an autoimmune thyroid disease and the leading cause of hypothyroidism in areas with sufficient iodine intake. The quality-of-life impact and financial burden of hypothyroidism and HT highlight the need for additional research investigating the disease etiology with the aim of revealing potential modifiable risk factors.

Objective: Implementation of measures against such risk factors, once identified, has the potential to lessen the financial burden while also improving the quality of life of many individuals. Therefore, we aimed to examine the potential seasonality of HT in Europe using the Google Trends data to explore whether there is a seasonal characteristic of Google searches regarding HT, examine the potential impact of the countries' geographic location on the potential seasonality, and identify potential modifiable risk factors for HT, thereby inspiring future research on the topic.

Methods: Monthly Google Trends data on the search topic "Hashimoto thyroiditis" were retrieved in a 17-year time frame from January 2004 to December 2020 for 36 European countries. A cosinor model analysis was conducted to evaluate potential seasonality. Simple linear regression was used to estimate the potential effect of latitude and longitude on seasonal amplitude and phase of the model outputs.

Results: Of 36 included European countries, significant seasonality was observed in 30 (83%) countries. Most phase peaks occurred in spring (14/30, 46.7%) and winter (8/30, 26.7%). A statistically significant effect was observed regarding the effect of geographical latitude on cosinor model amplitude ($y = –3.23 + 0.13 x$; R^2=0.29; P=.002). Seasonal increases in HT search volume may therefore be a consequence of an increased incidence or higher disease activity. It is particularly interesting that in most countries, a seasonal peak occurred in spring and winter months; when viewed in the context of the statistically significant impact of geographical latitude on seasonality amplitude, this may indicate the potential role of vitamin D levels in the seasonality of HT.

Conclusions: Significant seasonality of HT Google Trends search volume was observed in our study, with seasonal peaks in most countries occurring in spring and winter and with a significant impact of latitude on seasonality amplitude. Further studies on the topic of seasonality in HT and factors impacting it are required.

(JMIR Bioinform Biotech 2022;3(1):e38976) doi:10.2196/38976

KEYWORDS

Hashimoto disease; Hashimoto thyroiditis; infodemiology; search engine; Google Trends; seasonality; cosinor analysis; Google; thyroid

Introduction

Hypothyroidism is a growing global public health problem affecting approximately 5% of the general global population [1]. The leading cause of hypothyroidism, in areas with sufficient iodine intake, is Hashimoto thyroiditis (HT)—an autoimmune thyroid disease, with an incidence of 0.3-0.5 of 1000 population per year [2]. HT can present with a long list of both local and systematic symptoms such as dyspnea, dysphagia, dysphonia, constipation, coronary artery disease, bradycardia, anemia, memory loss, depression, bile colic, and dry and thickened skin [3]. The etiology of HT is still insufficiently clarified, but it is known to be associated with genetic factors, environmental triggers, and epigenetic influences [4]. Although some predisposing factors, such as stress, age, and gender, are recognized in the pathogenesis [5], much about HT still remains unknown and warrants further investigation. Namely, hypothyroidism represents a significant global, financial, and quality-of-life burden [1]. For example, hypothyroidism-related medical costs in the United States are estimated to range from US $460 to US $2555 per patient per year, patients with hypothyroid reporting significantly higher work absenteeism and significant long- and short-term disability costs, resulting in further direct and indirect costs [6]. The high quality-of-life and financial burden of hypothyroidism and HT emphasize the need of additional research investigating the disease etiology to reveal potential modifiable risk factors. Implementing measures against such risk factors has the potential to lessen the financial burden while also improving the quality of life of many individuals. Various environmental and seasonal factors, such as insolation and UV exposure, seasonal incidence of infectious diseases, or seasonal changes in human behavior, may potentially play a role in the disease's etiology. Vitamin D is a particularly interesting seasonal factor implicated in the pathophysiology of numerous diseases. Studies about vitamin D alteration by climate variation [7,8] along with studies about vitamin D deficiency in the pathogenesis of many autoimmune diseases, such as multiple sclerosis [7], diabetes, and also cancers, [9] have been conducted. A recently published systematic review on the topic of the association between vitamin D deficiency and autoimmune thyroid disorders found out that most of the studies included in the review supported the association between low vitamin D levels and the occurrence of autoimmune thyroid diseases, in particular HT and Graves disease, but the authors highlighted the need for further randomized long-term follow-up studies to confirm the potential causal link and potential role of vitamin D supplementation [10]. Unfortunately, such studies are costly and difficult to conduct; therefore, until such trials are performed, other sources of information may serve to narrow the evidence gap. One such area of science, which may help to provide further evidence regarding potential risk factors for HT, is infodemiology (short for information epidemiology); infodemiology studies health-related user data available on the internet with goals of improving public health, reducing the impact of web-based misinformation, and narrowing the existing knowledge gaps [11]. In recent years, the increase in internet penetration and usage as well as the number of web-based social platforms have provided a rich source of user-related health information. Therefore, multiple internet services and platforms have been used in infodemiology studies in recent years with various relevant topics being explored. Multiple studies used the data retrieved from Twitter to explore the sentiment and analyze conversation themes regarding COVID-19 vaccines, with goals of providing information relevant to fighting vaccine hesitancy [12,13], while other studies explored the potential role of Instagram in raising awareness of skin cancer [14] or analyzed the platform's contents regarding vaping [15]. Apart from social media platforms, internet search engines represent a particularly interesting source of health-related data. Changes in internet search trends have been found to potentially reflect the general public's interest regarding medical-related topics [16], and disease-specific internet searches tend to increase in parallel with the increase in the specific burden of disease [17,18]. Worldwide, Google is the most widely used search engine, and it also provides a freely accessible tool set called Google Trends (GT) through which one can easily conduct an analysis of search trends. With GT data being accessible in real time, the issue of lingering survey methods becomes obsolete, as the data are available practically instantly. Another major advantage of GT is the fact that it enables obtaining information that would be difficult, costly, or even impossible to obtain with conventional methods, such as the estimated population sizes and geospatial distributions of marginalized populations (eg, LGBTQ+), which is important for public health planning [19]. In addition, GT makes delicate topics and disease research, such as HIV, suicide rates, and mental illnesses, more easily doable as web searches are executed anonymously [20]. Thus, GT has so far been used in a number of studies investigating a broad range of medical topics, and based on the GT data, seasonal patterns have been proposed for various diseases ranging from psoriasis [21], gout [22], bruxism [23], and cellulitis [24] to major mental disorders [25].

Therefore, the aim of this study was to explore the potential seasonal pattern of Google searches regarding HT in European countries in order to guide future real-world studies on this topic.

Methods

Research Questions

Our study examined the seasonality of HT Google-related searches in Europe using GT data, with the goal to explore whether there was a seasonal characteristics of Google searches regarding HT, examine the potential impact of the countries' geographic location on the potential seasonality, and identify possible modifiable risk factors for HT, thereby inspiring future research on the topic.

Data Retrieval

We conducted our GT data retrieval throughout May 2021, and using all the search categories, we queried GT using the search topic "Hashimoto thyroiditis." The data were retrieved for 38 European countries, covering a 17-year period of time, from January 2004 to December 2020. A total of 36 European countries were included in the analysis, while Kosovo and Moldova were excluded due to an extremely small search volume. As we investigated the potential impact of seasonality

on Google searches, the widest available search time frame of 17 years was used. A specific search topic and no search category restrictions were used to potentially capture the interest of the population of multiple European countries, as wide as possible. The search topic approach was used as it included alternative spellings and translations in other languages; this approach, when comparing European countries using different languages in their Google searches, provided a simple and uniform way to extract the data for each country. The methods were reported following suggestions from a systematic review article on the topic of the use of GT in health care research [18].

Data Analysis

GT search data were expressed as relative search volume (RSV) normalized to range from 0% to 100% for the set search time frame and geographical location. Seasonality was assessed using a cosinor regression model from the R programming language "seasons" package. The model fits a sinusoid to the monthly input data, and its outputs include the sinusoid's amplitude, the phase corresponding to the sinusoidal peak, and 2 *P* values, the smaller of which is being reported in Table 1. A *P* value of less than .025 was considered statistically significant. More information on the cosinor model can be found in studies by Cornelissen [26], Mei et al [27], and Wu et al [21]. Months with 0 RSV have been encoded as NA, allowed by the model.

Seasons have been defined as spring (March, April, and May), summer (June, July, and August), autumn (September, October, and November), and winter (December, January, and February).

To evaluate the potential influence of the countries' latitudes and longitudes, the weighted population center coordinates for each country were retrieved from the Baylor University population resource [28] except for Serbia and Montenegro, where the coordinates of the capital cities were used instead, as the weighted population center coordinates for the 2 countries were not reported in the population resource.

Simple linear regression was conducted to evaluate the potential effect of latitude and longitude on seasonal amplitude and phase of the cosinor model output for each country.

All statistical analyses and data visualizations were done in R programming language (version 4.0.5; R Core Team).

Table 1. Cosinor model analysis results regarding country seasonality, amplitude, phase, and cosinor *P* values. Phase corresponds to month. Number of observations column correspond to the number of months for each country with a Google Trends relative search volume. Latitude and longitude values are used in the simple linear regression.

Country	Seasonality	Amplitude	Phase	Phase season	*P* value	Number of observations	Latitude	Longitude
Albania	Yes	10.72	7.4	Summer	<.001	50	41.174529494701	19.929275580053
Austria	Yes	2.24	7	Summer	<.001	191	47.765386201318	14.645625300333
Belarus	Yes	1.67	9.3	Autumn	<.001	115	53.531624124024	27.847175354981
Belgium	Yes	2.17	3.3	Spring	.002	175	50.844005826061	4.4332869095216
Bosnia and Herzegovina	No	—[a]	—	No seasonality	.07	118	44.160791721547	17.753208075376
Bulgaria	Yes	2.21	7.3	Summer	<.001	180	42.754116369708	25.083976957381
Croatia	Yes	2.06	1	Winter	.001	152	45.317637428417	16.262950671815
Czech Republic	Yes	2.1	7.5	Summer	<.001	113	49.821456149539	15.617527756779
Denmark	Yes	2.01	2.9	Winter	.007	144	55.853326754724	10.856715208377
Estonia	Yes	5.63	8.7	Summer	<.001	48	58.957945858648	25.572740786761
Finland	Yes	6.19	2.3	Winter	<.001	126	61.755732589277	24.98467066628
France	Yes	2.58	4.7	Spring	<.001	202	47.143228746162	2.6764463428893
Germany	Yes	3.62	5.2	Spring	<.001	204	50.855573924694	9.6963409646128
Greece	Yes	1.49	5.9	Spring	.02	173	38.686808689502	23.323965300494
Hungary	Yes	1.75	4.5	Spring	<.001	146	47.288770753717	19.388772968949
Iceland	Yes	7.21	5.1	Spring	<.001	54	64.372216876845	–21.045029641756
Ireland	No	—	—	No seasonality	.57	121	53.111585555903	–7.4282382442794
Italy	Yes	3.43	4.3	Spring	<.001	203	42.870086858764	12.12890612484
Latvia	No	—	—	No seasonality	.24	96	56.831191706188	24.496056054831
Lithuania	Yes	3.63	8.4	Summer	<.001	63	55.223194780885	23.887086150639
Luxembourg	Yes	5.58	3.6	Spring	<.001	101	49.643734947502	6.0837996175026
Macedonia	Yes	4.66	2.4	Winter	<.001	79	41.742844591767	21.554126089671
Montenegro	Yes	3.96	3.3	Spring	<.001	73	42.442574	19.268646
Netherlands	Yes	2.61	2.3	Winter	.001	193	52.072871145825	5.2875541627667
Norway	Yes	3.77	5.9	Spring	<.001	123	61.128336570352	9.9468336009803
Poland	Yes	2.56	2.5	Winter	.005	196	51.707976823759	19.308388806995
Portugal	No	—	—	No seasonality	.11	157	39.74693753116	–9.1672490596965
Republic of Serbia	Yes	1.46	3.5	Spring	.02	150	44.787197	20.457273
Romania	Yes	1.93	10.4	Autumn	.001	161	45.692835166704	25.283411622442
Slovakia	Yes	3.2	12.8	Winter	<.001	98	48.662536038829	19.164300350224
Slovenia	Yes	2.5	3.9	Spring	<.001	111	46.169295822703	14.89236963709
Spain	Yes	2.01	5.2	Spring	.008	191	39.720397339383	–3.2923251997811
Sweden	No	—	—	No seasonality	.04	165	58.913317696441	15.528746561364
Switzerland	Yes	2.68	4.2	Spring	<.001	185	47.025712614417	7.9586515381984
Ukraine	No	—	—	No seasonality	.38	153	48.808076188342	31.766935926448
United Kingdom	Yes	2.03	2.4	Winter	<.001	196	52.745166767654	–1.6847761296012

[a]Not applicable.

Ethical Considerations

The ethical committee of the University of Zagreb School of Medicine exempted this study from review.

Results

Cosinor model results on the seasonality of the search term "Hashimoto thyroiditis" in 36 European countries can be seen in Multimedia Appendix 1 and Table 1. Boxplot graphs in Multimedia Appendix 1 represent the monthly GT RSV for each country along with the sinusoid resulting from the cosinor model.

Of the 36 included European countries, significant seasonality was observed in 30 (83%) countries, with a mean amplitude of 3.3 (SD 2.0; median=2.6) and a mean phase value of 5.24 (SD 2.8; median=4.6).

Distribution of the phase months can be seen in Figure 1; most of the phase peaks occurred during spring (14/30, 46.7%), winter (8/30, 26.7%), and summer (6/30, 20%), while the least phase peaks were in autumn (2/30, 6.7%). Geographical distribution of phase seasons is shown in Figure 2.

Simple linear regression results of the effect of latitude and longitude on seasonal amplitude and phase of the cosinor model are demonstrated in Figure 3. A statistically significant effect was observed regarding the effect of latitude on seasonality amplitude ($y = -3.23 + 0.13x$; R^2=0.29; P=.002). No statistically significant effects were observed regarding the effect of latitude on phase month (P=.22), longitude on amplitude (P=.94), or longitude on phase month (P=.07). The amplitude value of Albania was excluded from the linear regression models as it was identified as an extreme outlier, most likely due to low quantity of the monthly RSV data.

Figure 1. Density plot showing the distribution of phase months.

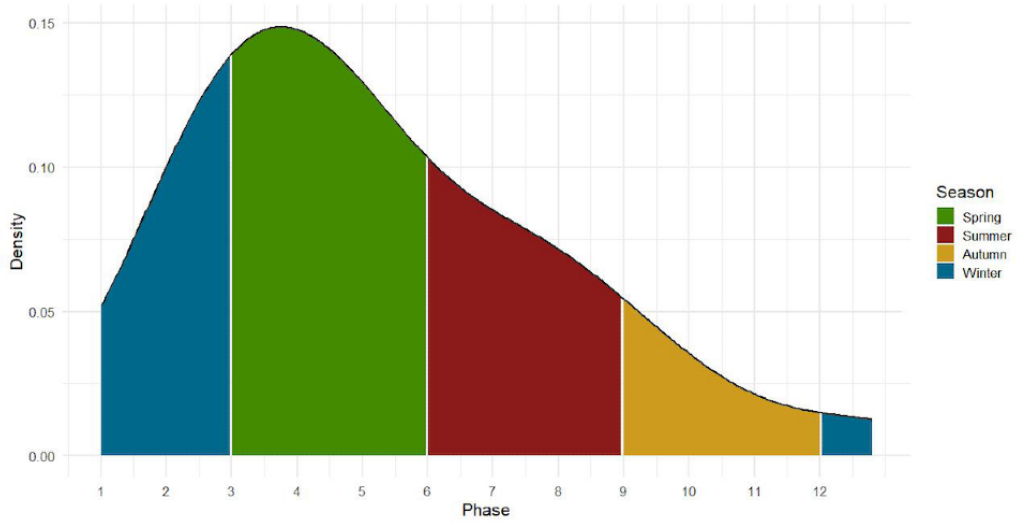


Figure 2. Map of Europe colored by season of phase. Points represent weighted population centers for each country, except Serbia and Montenegro, where coordinates of capital cities were used.

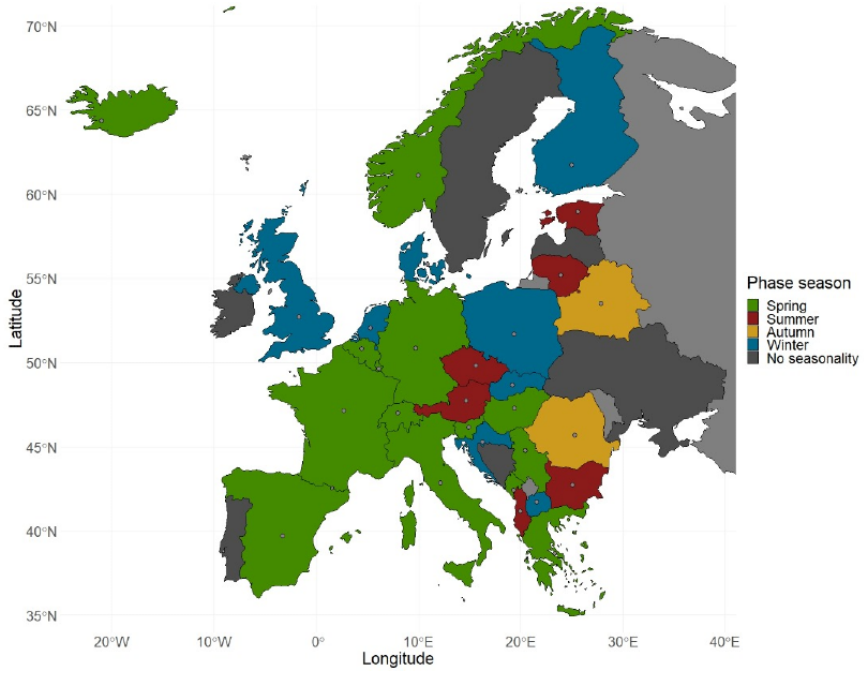
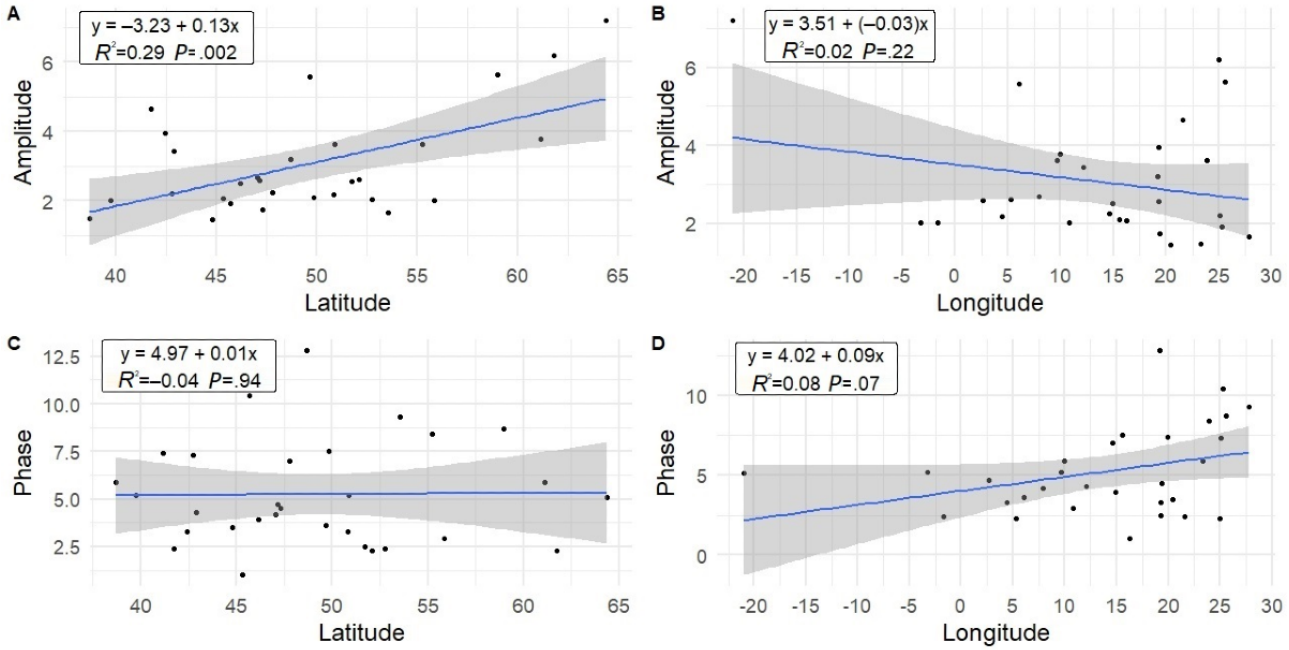


Figure 3. Graphs showing simple linear regression results of the impact of latitude and longitude on amplitude and phase.



Discussion

Principal Findings

The results of our study demonstrate statistically significant seasonality in HT-related RSV across Europe, with most seasonal peaks occurring in spring and winter months (22/30, 73.3%). Additionally, we have also observed a statistically significant impact of geographical latitude on seasonal amplitude.

Comparison With Prior Work

A study exploring internet searches patterns regarding hypothyroidism found similar results, with more hypothyroidism-related internet searches occurring during spring globally [29]. Seasonal increases in RSV for HT may be a consequence of increased incidence or higher disease activity. Namely, the volume of disease-related internet searches is known to correlate with patients' desire to gather more information before an appointment and to supplement information provided by the physician [30]. It is particularly interesting that in most countries, a seasonal peak occurred in spring and winter months; when viewed in the context of the statistically significant impact of geographical latitude on seasonality amplitude, this may indicate that vitamin D levels could play a role in the seasonality of HT. Significant discrepancies in vitamin D levels between northern and southern European countries have been observed, with people in northern European countries having lower levels of vitamin D, which can be associated with less sun exposure, geographical latitude, and solar zenith angle [31]. People living in higher-latitude countries receive lower yearly amounts of sunlight, which leads to a predisposition for developing vitamin D deficiency. Studies have shown that shorter days and insufficient sunlight exposure at latitudes above 40 degrees North lead to poor vitamin D synthesis in the skin [7]. Furthermore, seasonal changes in serum vitamin D levels have been implicated in the seasonality and outcomes of infectious disease [32]. Vitamin D levels are also known to have seasonal fluctuations with the lowest serum levels occurring in the late winter and early spring months [33], which seems to correlate well with the seasonal increase in the RSVs of HT in most countries. Research by Kim [34] and Jamka et al [35] highlighted the importance of vitamin D in the pathogenesis of HT. Kim's [34] cross-sectional study showed significantly higher prevalence of vitamin D insufficiency in patients with autoimmune thyroid disease, while Jamka et al [35] demonstrated a presumed vitamin D effect on reduction in the levels of thyroid peroxidase antibodies, which has an important role in HT pathogenesis [36]. Multiple studies have demonstrated beneficial effects of vitamin D supplementation on autoimmune diseases, but most notable is the recently published study in the *BMJ*, which found that vitamin D supplementation (2000 IU/day) could reduce autoimmune disease rate by 22% [37]. Future research should focus on the association between vitamin D serum seasonal changes as a potential trigger of HT and vitamin D supplementation during winter and early spring months, which may prove to be an easily implemented public health measure to decrease the burden of HT.

Limitations

It is important to consider some of the possible limitations of this study. First, selection bias might be present, as this study only used data pertaining to the population with internet access and those who used Google instead of other search engines. Second, medical-related searches may be performed by anyone interested in a particular medical topic and not only patients. Third, deeper analysis of individual users could not be performed due to the limitations of the available data set. Finally, the impact of potential confounding factors, such as academic cycling (ie, higher search volumes in spring during college exams), could not be excluded.

Conclusions

Significant seasonality of GT search volume for HT was observed in our study, with seasonal peaks in most European countries occurring during spring and winter. A significant impact of latitude on seasonality amplitude was also demonstrated. Additional studies on the topic of seasonality in HT and factors impacting it are required. If vitamin D deficiency is unequivocally proven as a contributing factor in the development of HT, vitamin D supplementation during winter and early spring months might be an easily implemented public health measure aimed at decreasing the burden of this disease.

Authors' Contributions

RM conceived the idea, retrieved the data, and conducted the analysis. JS performed the literature search and wrote the first draft. RM and RL revised the first draft. RL provided input and advice in all steps of the study's design and conduct.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Boxplot graphs showing monthly Google Trends data with cosinor model output. Y-axis represents relative search volume.
[DOCX File , 901 KB - bioinform_v3i1e38976_app1.docx]

References

1. Chiovato L, Magri F, Carlé A. Hypothyroidism in context: where we've been and where we're going. Adv Ther 2019 Sep;36(Suppl 2):47-58 [FREE Full text] [doi: 10.1007/s12325-019-01080-8] [Medline: 31485975]
2. Hiromatsu Y, Satoh H, Amino N. Hashimoto's thyroiditis: history and future outlook. Hormones 2013 Jan 1;12(1):12-18. [doi: 10.1007/bf03401282]
3. Caturegli P, De Remigis A, Rose N. Hashimoto thyroiditis: clinical and diagnostic criteria. Autoimmun Rev 2014 Apr;13(4-5):391-397. [doi: 10.1016/j.autrev.2014.01.007] [Medline: 24434360]
4. Ralli M, Angeletti D, Fiore M, D'Aguanno V, Lambiase A, Artico M, et al. Hashimoto's thyroiditis: An update on pathogenic mechanisms, diagnostic protocols, therapeutic strategies, and potential malignant transformation. Autoimmun Rev 2020 Oct;19(10):102649. [doi: 10.1016/j.autrev.2020.102649] [Medline: 32805423]
5. Ajjan RA, Weetman AP. The pathogenesis of Hashimoto's thyroiditis: further developments in our understanding. Horm Metab Res 2015 Sep 16;47(10):702-710. [doi: 10.1055/s-0035-1548832] [Medline: 26361257]
6. Hepp Z, Lage MJ, Espaillat R, Gossain VV. The direct and indirect economic burden of hypothyroidism in the United States: a retrospective claims database study. J Med Econ 2021 Mar 30;24(1):440-446 [FREE Full text] [doi: 10.1080/13696998.2021.1900202] [Medline: 33685322]
7. Ghareghani M, Reiter RJ, Zibara K, Farhadi N. Latitude, vitamin D, melatonin, and gut microbiota act in concert to initiate multiple sclerosis: a new mechanistic pathway. Front Immunol 2018 Oct 30;9:2484 [FREE Full text] [doi: 10.3389/fimmu.2018.02484] [Medline: 30459766]
8. Kashi Z, Saeedian FS, Akha O, Gorgi MAH, Emadi SF, Zakeri H. Vitamin D deficiency prevalence in summer compared to winter in a city with high humidity and a sultry climate. Endokrynol Pol 2011;62(3):249-251 [FREE Full text] [Medline: 21717408]
9. Sizar O, Khare S, Goyal A, Bansal P, Givler A. Vitamin D Deficiency. StatPearls, Treasure Island (FL): StatPearls Publishing; 2021.
10. Khozam S, Sumaili A, Alflan M, Shawabkeh R. Association between vitamin D deficiency and autoimmune thyroid disorder: a systematic review. Cureus 2022 Jun;14(6):e25869 [FREE Full text] [doi: 10.7759/cureus.25869] [Medline: 35836431]
11. Eysenbach G. Infodemiology: the epidemiology of (mis)information. Am J Med 2002 Dec;113(9):763-765. [doi: 10.1016/S0002-9343(02)01473-0]
12. Boucher J, Cornelson K, Benham JL, Fullerton MM, Tang T, Constantinescu C, et al. Analyzing social media to explore the attitudes and behaviors following the announcement of successful COVID-19 vaccine trials: infodemiology study. JMIR Infodemiology 2021;1(1):e28800 [FREE Full text] [doi: 10.2196/28800] [Medline: 34447924]
13. Marcec R, Likic R. Using Twitter for sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and Moderna COVID-19 vaccines. Postgrad Med J 2022 Jul 09;98(1161):544-550 [FREE Full text] [doi: 10.1136/postgradmedj-2021-140685] [Medline: 34373343]
14. Gomaa B, Houghton R, Crocker N, Walsh-Buhi E. Skin cancer narratives on Instagram: content analysis. JMIR Infodemiology 2022 Jun 2;2(1):e34940 [FREE Full text] [doi: 10.2196/34940]
15. Gao Y, Xie Z, Sun L, Xu C, Li D. Electronic cigarette-related contents on Instagram: observational study and exploratory analysis. JMIR Public Health Surveill 2020 Nov 05;6(4):e21963 [FREE Full text] [doi: 10.2196/21963] [Medline: 33151157]
16. Bundorf MK, Wagner TH, Singer SJ, Baker LC. Who searches the internet for health information? Health Serv Res 2006 Jun;41(3 Pt 1):819-836 [FREE Full text] [doi: 10.1111/j.1475-6773.2006.00510.x] [Medline: 16704514]

17. Cervellin G, Comelli I, Lippi G. Is Google Trends a reliable tool for digital epidemiology? Insights from different clinical settings. J Epidemiol Glob Health 2017 Dec;7(3):185-189 [FREE Full text] [doi: 10.1016/j.jegh.2017.06.001] [Medline: 28756828]
18. Nuti SV, Wayda B, Ranasinghe I, Wang S, Dreyer RP, Chen SI, et al. The use of google trends in health care research: a systematic review. PLoS One 2014 Oct;9(10):e109583 [FREE Full text] [doi: 10.1371/journal.pone.0109583] [Medline: 25337815]
19. Card KG, Lachowsky NJ, Hogg RS. Using Google Trends to inform the population size estimation and spatial distribution of gay, bisexual, and other men who have sex with men: proof-of-concept study. JMIR Public Health Surveill 2021 Nov 29;7(11):e27385 [FREE Full text] [doi: 10.2196/27385] [Medline: 34618679]
20. Mavragani A, Ochoa G. Google Trends in infodemiology and infoveillance: methodology framework. JMIR Public Health Surveill 2019 May 29;5(2):e13439 [FREE Full text] [doi: 10.2196/13439] [Medline: 31144671]
21. Wu Q, Xu Z, Dan Y, Zhao C, Mao Y, Liu L, et al. Seasonality and global public interest in psoriasis: an infodemiology study. Postgrad Med J 2020 Mar 11;96(1133):139-143. [doi: 10.1136/postgradmedj-2019-136766] [Medline: 31511319]
22. Kardeş S. Seasonal variation in the internet searches for gout: an ecological study. Clin Rheumatol 2019 Mar;38(3):769-775. [doi: 10.1007/s10067-018-4345-2] [Medline: 30374747]
23. Kardeş S, Kardeş E. Seasonality of bruxism: evidence from Google Trends. Sleep Breath 2019 Jun 21;23(2):695-701. [doi: 10.1007/s11325-019-01787-6] [Medline: 30790220]
24. Zhang X, Dang S, Ji F, Shi J, Li Y, Li M, et al. Seasonality of cellulitis: evidence from Google Trends. IDR 2018 May;Volume 11:689-693. [doi: 10.2147/idr.s163290]
25. Ayers JW, Althouse BM, Allem J, Rosenquist JN, Ford DE. Seasonality in seeking mental health information on Google. Am J Prev Med 2013 May;44(5):520-525. [doi: 10.1016/j.amepre.2013.01.012] [Medline: 23597817]
26. Cornelissen G. Cosinor-based rhythmometry. Theor Biol Med Model 2014 Apr 11;11:16 [FREE Full text] [doi: 10.1186/1742-4682-11-16] [Medline: 24725531]
27. Mei Y, Mao Y, Cao F, Wang T, Li Z. Using internet search data to explore the global public concerns in ankylosing spondylitis. Postgrad Med J 2021 Feb 24;97(1144):93-96. [doi: 10.1136/postgradmedj-2019-137407] [Medline: 32094142]
28. European states population-weighted centers. Cs.baylor.edu. URL: https://cs.baylor.edu/~hamerly/software/europe_population_weighted_centers.html [accessed 2022-08-27]
29. Ilias I, Alexiou M, Meristoudis G. Is there seasonality in hypothyroidism? A Google Trends pilot study. Cureus 2019 Jan 25;11(1):e3965 [FREE Full text] [doi: 10.7759/cureus.3965] [Medline: 30956917]
30. Orgaz-Molina J, Cotugno M, Girón-Prieto M, Arrabal-Polo M, Ruiz-Carrascosa J, Buendía-Eisman A, et al. A study of internet searches for medical information in dermatology patients: The patient–physician relationship. Actas Dermo-Sifiliográficas (English Edition) 2015 Jul;106(6):493-499. [doi: 10.1016/j.adengl.2015.01.019]
31. O'Neill CM, Kazantzidis A, Ryan M, Barber N, Sempos C, Durazo-Arvizu R, et al. Seasonal changes in vitamin D-effective UVB availability in Europe and associations with population serum 25-hydroxy vitamin D. Nutrients 2016 Aug 30;8(9):533 [FREE Full text] [doi: 10.3390/nu8090533] [Medline: 27589793]
32. Abhimanyu A, Coussens AK. The role of UV radiation and vitamin D in the seasonality and outcomes of infectious disease. Photochem Photobiol Sci 2017 Mar 16;16(3):314-338. [doi: 10.1039/c6pp00355a] [Medline: 28078341]
33. Klingberg E, Oleröd G, Konar J, Petzold M, Hammarsten O. Seasonal variations in serum 25-hydroxy vitamin D levels in a Swedish cohort. Endocrine 2015 Aug 14;49(3):800-808 [FREE Full text] [doi: 10.1007/s12020-015-0548-3] [Medline: 25681052]
34. Kim D. Low vitamin D status is associated with hypothyroid Hashimoto's thyroiditis. Hormones (Athens) 2016 Jul 9;15(3):385-393 [FREE Full text] [doi: 10.14310/horm.2002.1681] [Medline: 27394703]
35. Jamka M, Ruchała M, Walkowiak J. [Vitamin D and Hashimoto's disease]. Pol Merkur Lekarski 2019 Sep 25;47(279):111-113. [Medline: 31557141]
36. Ihnatowicz P, Drywień M, Wątor P, Wojsiat J. The importance of nutritional factors and dietary management of Hashimoto's thyroiditis. Ann Agric Environ Med 2020 Jun 19;27(2):184-193 [FREE Full text] [doi: 10.26444/aaem/112331] [Medline: 32588591]
37. Hahn J, Cook NR, Alexander EK, Friedman S, Walter J, Bubes V, et al. Vitamin D and marine omega 3 fatty acid supplementation and incident autoimmune disease: VITAL randomized controlled trial. BMJ 2022 Jan 26;376:e066452 [FREE Full text] [doi: 10.1136/bmj-2021-066452] [Medline: 35082139]

Abbreviations

GT: Google Trends
HT: Hashimoto thyroiditis
RSV: relative search volume

Edited by A Mavragani; submitted 24.04.22; peer-reviewed by M K., JK Kumar; comments to author 04.07.22; revised version received 24.07.22; accepted 15.08.22; published 01.09.22.

Please cite as:
Marcec R, Stjepanovic J, Likic R
Seasonality of Hashimoto Thyroiditis: Infodemiology Study of Google Trends Data
JMIR Bioinform Biotech 2022;3(1):e38976
URL: https://bioinform.jmir.org/2022/1/e38976
doi:10.2196/38976
PMID:

Original Paper

Multiple-Inputs Convolutional Neural Network for COVID-19 Classification and Critical Region Screening From Chest X-ray Radiographs: Model Development and Performance Evaluation

Zhongqiang Li[1], PhD; Zheng Li[1]; Luke Yao[1]; Qing Chen[2], PhD; Jian Zhang[2], PhD; Xin Li[3], PhD; Ji-Ming Feng[4], PhD; Yanping Li[5], PhD; Jian Xu[1], PhD

[1]Division of Electrical and Computer Engineering, College of Engineering, Louisiana State University, Baton Rouge, LA, United States

[2]Division of Computer Science and Engineering, College of Engineering, Louisiana State University, Baton Rouge, LA, United States

[3]Department of Visualization, Texas A & M University, College Station, TX, United States

[4]Department of Comparative Biomedical Science, School of Veterinary Medicine, Louisiana State University, Baton Rouge, LA, United States

[5]School of Environment and Sustainability, University of Saskatchewan, Saskatoon, SK, Canada

Corresponding Author:
Jian Xu, PhD
Division of Electrical and Computer Engineering
College of Engineering
Louisiana State University
Patrick F Taylor Hall
3304 S Quad Dr
Baton Rouge, LA, 70803
United States
Phone: 1 (225) 578 4483
Email: jianxu1@lsu.edu

Abstract

Background: The COVID-19 pandemic is becoming one of the largest, unprecedented health crises, and chest X-ray radiography (CXR) plays a vital role in diagnosing COVID-19. However, extracting and finding useful image features from CXRs demand a heavy workload for radiologists.

Objective: The aim of this study was to design a novel multiple-inputs (MI) convolutional neural network (CNN) for the classification of COVID-19 and extraction of critical regions from CXRs. We also investigated the effect of the number of inputs on the performance of our new MI-CNN model.

Methods: A total of 6205 CXR images (including 3021 COVID-19 CXRs and 3184 normal CXRs) were used to test our MI-CNN models. CXRs could be evenly segmented into different numbers (2, 4, and 16) of individual regions. Each region could individually serve as one of the MI-CNN inputs. The CNN features of these MI-CNN inputs would then be fused for COVID-19 classification. More importantly, the contributions of each CXR region could be evaluated through assessing the number of images that were accurately classified by their corresponding regions in the testing data sets.

Results: In both the whole-image and left- and right-lung region of interest (LR-ROI) data sets, MI-CNNs demonstrated good efficiency for COVID-19 classification. In particular, MI-CNNs with more inputs (2-, 4-, and 16-input MI-CNNs) had better efficiency in recognizing COVID-19 CXRs than the 1-input CNN. Compared to the whole-image data sets, the efficiency of LR-ROI data sets showed approximately 4% lower accuracy, sensitivity, specificity, and precision (over 91%). In considering the contributions of each region, one of the possible reasons for this reduced performance was that nonlung regions (eg, region 16) provided false-positive contributions to COVID-19 classification. The MI-CNN with the LR-ROI data set could provide a more accurate evaluation of the contribution of each region and COVID-19 classification. Additionally, the right-lung regions had higher contributions to the classification of COVID-19 CXRs, whereas the left-lung regions had higher contributions to identifying normal CXRs.

Conclusions: Overall, MI-CNNs could achieve higher accuracy with an increasing number of inputs (eg, 16-input MI-CNN). This approach could assist radiologists in identifying COVID-19 CXRs and in screening the critical regions related to COVID-19 classifications.

(JMIR Bioinform Biotech 2022;3(1):e36660) doi:10.2196/36660

KEYWORDS

COVID-19; chest X-ray radiography; multiple-inputs convolutional neural network; screening critical COVID regions

Introduction

Background

In early 2020, COVID-19 was officially announced as a pandemic by the World Health Organization (WHO), which rapidly spread to become one of the largest unprecedented health crises worldwide [1]. To date, the COVID-19 pandemic has heavily impacted the global economy and threatened many people's lives [1]. According to the latest WHO reports, by July 2021, 190,671,330 people have been confirmed to have COVID-19, contributing to 4,098,758 deaths. In the United States, there have been 33,741,532 confirmed cases with 603,880 deaths [2]. At the time of writing, the United States, India, and Brazil have the highest numbers of confirmed cases globally, followed by France, Russia, Turkey, and the United Kingdom [2].

Luckily, several types of COVID-19 vaccines have been rapidly and accurately developed, such as Pfizer-BioNTech, Moderna, Johnson & Johnson's Janssen, and others, which are reaching an increasing number of populations worldwide [3]. For instance, by July 2021, 3,436,534,998 vaccine doses had been administered worldwide and 341,759,270 vaccine doses had been administered in the United States [2]. Unfortunately, the crisis of COVID-19 remains severe, primarily since the Delta variant of SARS-CoV-2 was first identified in December 2020 in India, followed by the second large wave in the country. This new variant quickly spread to more than 92 countries to become the dominant viral COVID-19 strain in the world [4]. Moreover, a more recent variant of SARS-CoV-2, Omicron, was reported in December 2021 and spread globally thereafter [5-7].

Currently, polymerase chain reaction (PCR), especially real-time reverse-transcription-PCR (RT-PCR), is considered the gold standard for diagnosing COVID-19. However, this method has many problems such as being time-consuming or requiring specialized personnel and laboratories [8,9]. In addition, medical imaging such as chest X-ray radiography (CXR), chest computed tomography (CT), and magnetic resonance imaging also serves as an important alternative method for COVID-19 diagnosis [1,8,10]. CXR is the imaging technique that was first used to diagnose COVID-19 and continues to play an important role in clinical diagnosis [11-16].

A chest CT scan could be more sensitive than CXR for the diagnosis of COVID-19; however, some significant issues hinder its use, such as high costs, time-intensive processes to scan a single patient, high levels of ionizing radiation, and limited access in some hospitals or health centers [8-16]. Therefore, CXR remains an affordable imaging technique that is widely used to diagnose COVID-19 with a much lower radiation dose [17]. In addition, in clinical practice, the RT-PCR test is often combined with a CXR examination to reduce the false negatives, and to assess the extent and severity of the disease [8,9].

Prior Work

In some conditions, extracting and finding useful image features from CXRs impose a heavy workload for radiologists [9,15,18]. In recent years, deep learning has become one of the most popular research topics in image classification, identification, and segmentation [8,19-24]. Compared to conventional approaches of image analysis, deep-learning methods usually have better efficiency in extracting image features since they do not require human supervision to determine the critical image features. The convolutional neural network (CNN) is one of the most representative examples of deep learning for learning and recognizing specific image features [8,19,20,25]. Therefore, integrating an efficient CNN architecture into diagnostic systems would help to reduce the workload of radiologists, while increasing the reliability of the result and enabling quantitative analysis [20]. To date, several CNN models have been reported to differentiate COVID-19 cases from other (non-COVID-19) cases with CXR, including GoogleNet, ResNet50, VGG19, MobileNetV2, and Inception. Most of these models could achieve very high accuracy (up to 99%) in the classification of COVID-19 [11,13-16,18,26-31]. Thus, deep learning with CXRs could be a valuable method to identify COVID-19.

In most of these previous studies, the CNN models were trained with whole-image CXRs as a single input for the classification [9,11,13-16,18,26-31]. Other studies also attempted to develop new CNN models that accept multiple inputs; such multiple-input CNNs (MI-CNN) could effectively improve the classification accuracy and demonstrated better performance than single-input CNNs [32-34]. Because an MI-CNN could provide different features, fusing these network features together could improve the accuracy of the entire system [34]. To date, MI-CNNs have been applied in the fields of facial expression and gender recognition [33,34] or flower grading [35].

However, most of the MI-CNN models developed in previous studies used whole images as at least one of the CNN inputs, and the prefeatured images were used as the other inputs [33,34]. To our best knowledge, few studies have reported using MI-CNNs to detect and analyze COVID-19 CXRs. In addition, some of the obtained features can allow the network to determine the correct result, while other features can also cause serious misjudgment [34]. Thus, evaluation of feature contribution and removal of negative CNN features are critical steps toward increasing the reliability of disease diagnosis. However, few studies have explored the feasibility of using an MI-CNN to extract important image regions and exclude the contribution of the irrelevant features for the classification of COVID-19.

In this study, we developed a novel COVID-19 classification strategy with MI-CNN models. CXR images could be evenly segmented into different regions, and each MI-CNN input could process only one part of the COVID-19 CXRs. Furthermore, MI-CNNs could screen the critical regions for the classification

of COVID-19 CXRs and exclude irrelevant image regions that falsely contribute to the COVID-19 classification.

Methods

COVID-19 and Normal CXR Image Data Sets

In this study, 6205 CXR images (including 3021 COVID-19 CXRs and 3184 normal CXRs) were obtained from previous reports [14,36,37]. All of the CXR images were resized to 320×320 pixels to obtain 16 image segmentations. All CXR images were used in the original PNG format without any modification.

Ethics Approval

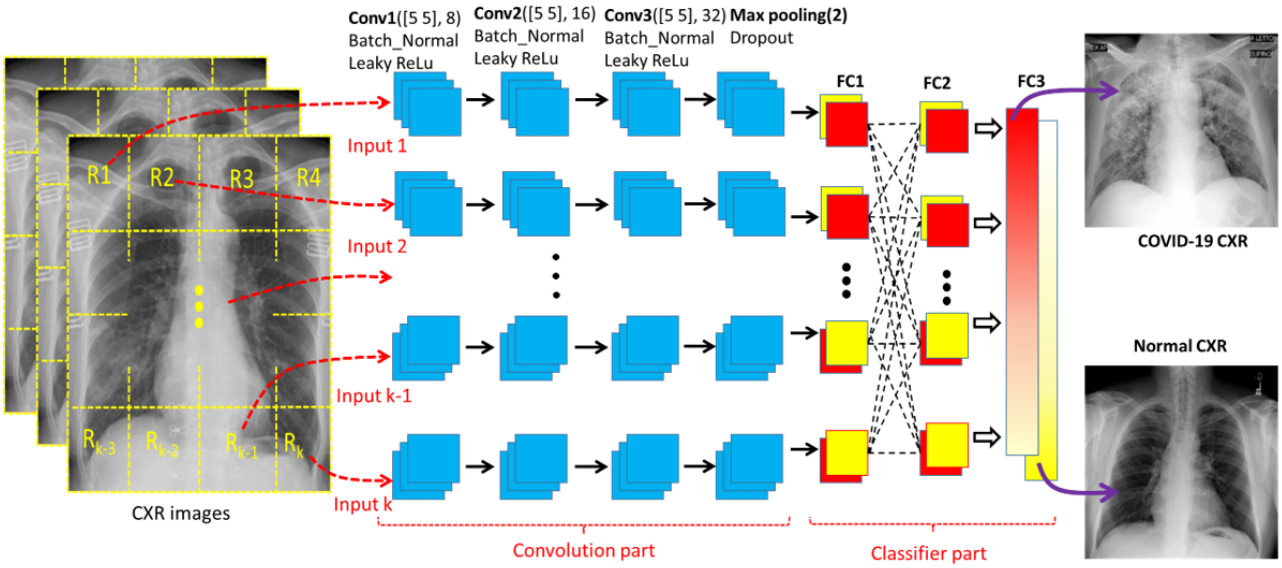
The ethics review was waived due to the use of secondary publicly available data, along with a lack of manipulation or intervention of human subjects, as determined by the Louisiana State University Institutional Review Board.

Design of the MI-CNN Architecture

Figure 1 provides a schematic diagram of the MI-CNN architecture, which is composed of the convolution part and classifier part. Each CXR was evenly segmented into 2, 4, or 16 different regions, and each type of image segmentation was loaded into the corresponding MI-CNN model (2-input, 4-input, and 16-input MI-CNNs). For the single-input CNN, the whole CXR image was directly loaded into the model. For the convolution part, each MI-CNN input had three convolutional sections. Each convolutional section included one 2D convolutional layer, one batch normalization (Batch_Normal), and one leaky rectified linear unit (ReLu) layer. All three 2D convolutional layers were set to a (5, 5) filter size. The filter number was set to 8 for the first 2D convolutional layer (Conv1), 16 for the second 2D convolutional layer (Conv2), and 32 for the third 2D convolutional layer (Conv3). A one max-pooling layer was used after the three convolutional sections.

There were three fully connected (FC) layers for the classifier part: the first FC (FC1) was set to receive the outputs from each MI-CNN input, FC2 was used to fully connect all of the FC1 outputs from all MI-CNN inputs, and FC3 was used to determine the CXR category (COVID-19 or normal). Accuracy, sensitivity, specificity, and precision were calculated for model performance evaluation.

Figure 1. Schematic diagram of the multiple-input convolutional neural network (MI-CNN) architecture. There are two parts in the MI-CNN: the convolution part and classifier part. The convolution part consists of up to 16 MI-CNN inputs (depending on the type of MI-CNN), and each MI-CNN input has three convolutional sections and one max-pooling layer. The classifier part is composed of three fully connected (FC) layers. CXR: chest X-ray radiograph; ReLu: rectified linear unit.

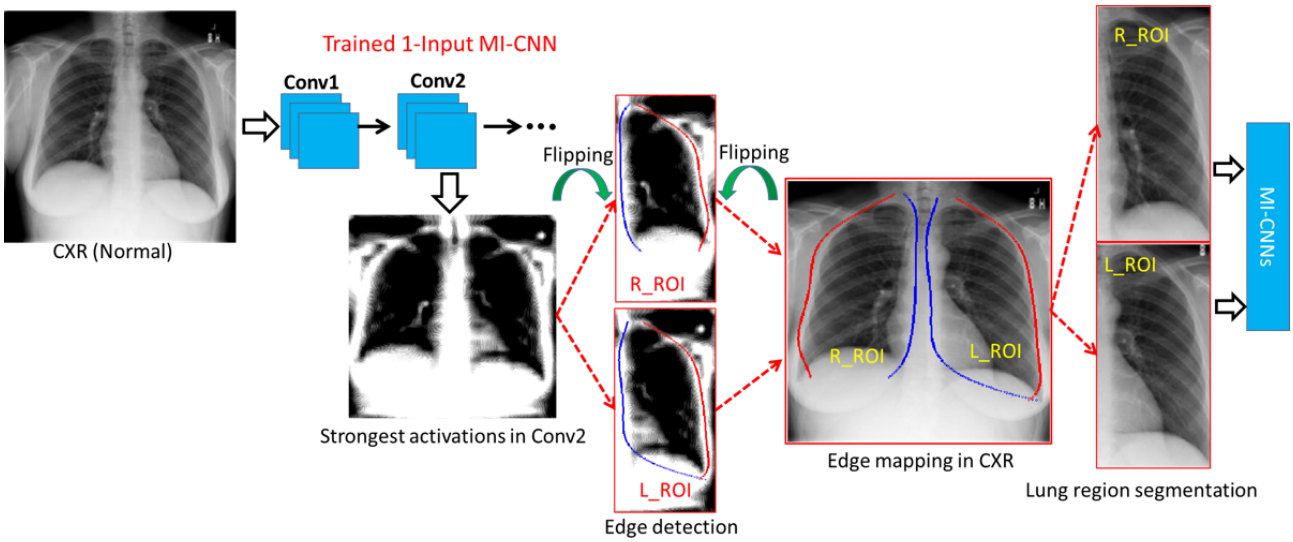


Segmentation of Lung Regions Through the Most Strongly Activated Convolutional Layer

The crucial steps in the automatic analysis of CXRs are accurate lung boundaries detection and their classification as normal or abnormal [17]. Segmentation of lung boundaries in medical imaging allows for disease identification, including for the detection of COVID-19 [17]. In the single-input CNN, the strongest activations of the second convolutional layer could yield the best profile of the lung regions, as shown in Figure 2. To extract both the left-lung region of interest (ROI) and the right-lung ROI, the CXR was first divided into two even parts and then the left part of the CXRs was flipped over horizontally.

Since the lung regions are relatively darker than the surrounding anatomical structures (the white regions), the edges of the left and right lung regions could be determined by the starting (blue lines) and end (red lines) points in each column. The coordinates of the lung edges were then projected onto the original CXRs. However, for some CXRs from patients at severe COVID-19 stages, it can be challenging to identify the lung regions from surrounding regions. Therefore, a rectangular region with minimum and maximum coordinates of the lung edges was used to crop the whole lung regions.

Figure 2. Schematic diagram of the regions of interest for the right and left lung (R_ROI and L_ROI, respectively) based on the strongest activation of max-pooling layers in the 1-Input convolutional neural network (CNN) model. Conv: convolutional layer; CXR: chest X-ray radiography; MI-CNN: multiple-input convolutional neural network.



Screening Critical Regions for COVID-19 Classification

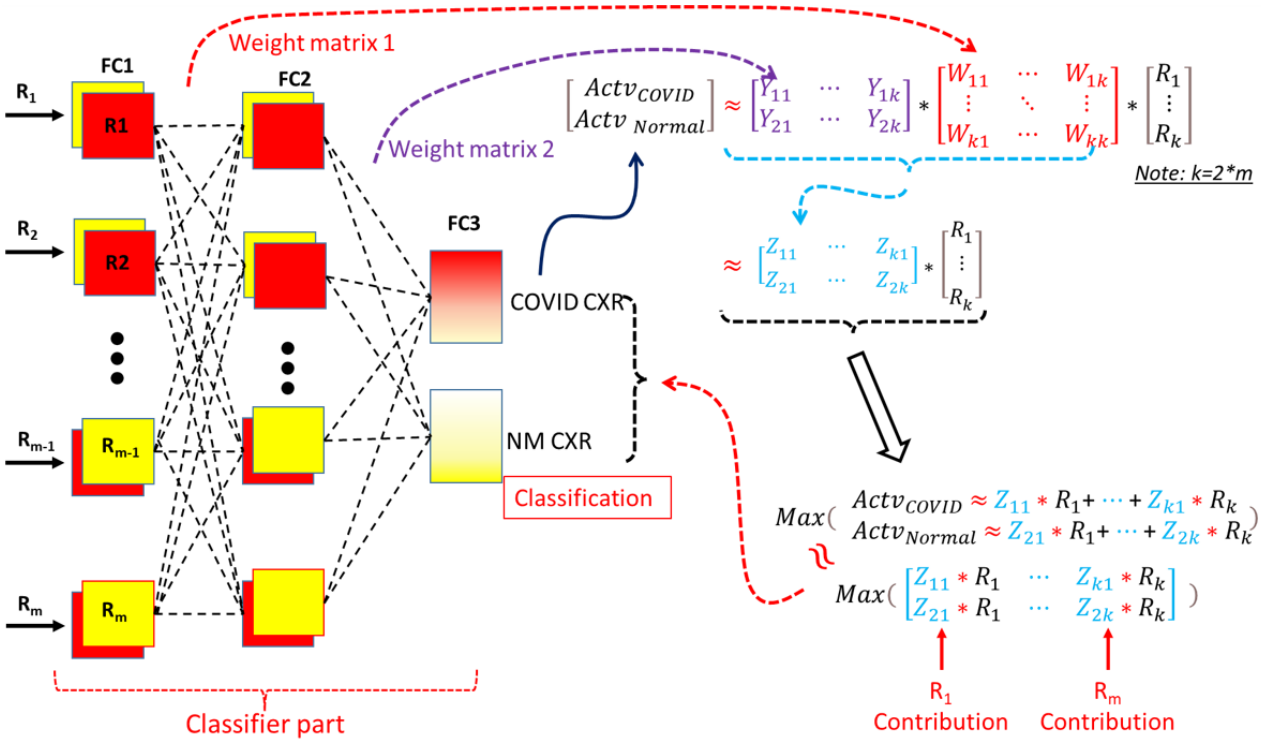
Although deep-learning approaches have been widely applied in analyses of medical images, few studies have reported the use of MI-CNN models to analyze the contributions of the critical ROIs and exclude the false contributions of regions that are irrelevant for the classification of diseases. To find the critical regions and exclude the irrelevant regions for the classification of COVID-19 CXRs, we explored the relationship between the outputs (*R* matrix) of each convolutional branch (also serving as the inputs of the classifier part) and the final activations (FC3 layer) of the classifier part, as shown in Figure 3.

The approximate relationship between FC1 and FC2 activations could be found through the weight matrix of the FC2 layer (*W* matrix in Figure 3). The weight matrix of the FC3 layer (*Y* matrix in Figure 3) could provide an approximate relationship between FC2 and FC3 activations. Therefore, the relationship between MI-CNN inputs (*R* matrix) and the final classification (FC3 activations) could be approximately evaluated through the matrix product (*Z* matrix) of the *W* and *Y* weight matrices. Furthermore, the final classification of COVID-19 or normal CXR was approximately determined by the maximum values of FC3 activations. The classification could then be approximately determined by the maximum elements of the element-wise multiplication between the *Z* matrix and *R* matrix (Figure 3).

Regarding the region contributions, the correctly classified images in the testing data sets were grouped with the labels of the corresponding regions (R1 to R16) that gave the maximum elements of element-wise multiplication (*Z* and *R* matrices). The region contributions could then be evaluated according to the percentages of the correctly classified images of each MI-CNN input.

All analyses were conducted with MATLAB R2020b (MathWorks Inc) on an HP Z2620 Workstation computer with a NVIDIA Tesla K80 GPU Accelerator.

Figure 3. Schematic diagram of the classifier in the multiple-input convolutional neural network (MI-CNN) and screening of critical regions that classify COVID-19 and normal (NM) chest X-ray radiographs (CXRs). FC: fully connected layer; R: matrix between MI-CNN inputs; W: weight matrix of the second fully connected layer (FC2); Y: weight matrix of the last fully connected layer (FC3); Z: product of the Y and W matrices.



Results

Evaluation of MI-CNNs With Different Inputs for Discrimination of COVID-19 and Normal CXRs

To evaluate the performance of the single-input CNN and MI-CNNs with 2, 4, and 16 inputs, 90% of the CXR data sets were used for training and the rest (10%) were used for testing. Five-fold cross-validation was used for all MI-CNN models. Figure 4 shows the training accuracy and loss curves of the 1-, 2-, 4-, and 16-input MI-CNNs, and each training had 50 epochs with a 0.01 learning rate. Throughout the 50-epochs training, the three MI-CNNs had higher training accuracy than the 1-input CNN. In addition, the 1-input CNN had much higher loss at the beginning of the training (with 20 epochs) than the MI-CNNs, but showed similar loss to that of the MI-CNNs until 35 epochs.

After 50 epochs, there was an approximate 0.02 training loss for the MI-CNNs and 0.05 loss for the 1-input CNN. The MI-CNNs also had approximately 3% higher accuracy than the 1-input CNN (~99% vs 96%). In addition, at the beginning of the training curves, MI-CNNs showed higher training accuracy, which was 50.94% for the 1-input CNN, 62.53% for the 2-inputs MI-CNN, 66.66% for the 4-inputs MI-CNN, and 72.60% for the 16-inputs MI-CNN.

Regarding the testing evaluations (Figure 5), all MI-CNNs exhibited good performance in the classification of COVID-19 and normal CXRs, which could achieve over 93% accuracy, sensitivity, specificity, and precision. Similar to the training accuracy and loss, MI-CNNs with more than 2 inputs also exhibited better classification efficiency than the 1-input CNN. For instance, the MI-CNNs usually had over 95% accuracy, sensitivity, specificity, and precision, whereas these metrics only reached around 93% for the 1-input CNN under the same conditions.

However, for MI-CNNs, the testing performance increased with more inputs, in which the 16-inputs MI-CNN showed the best classification of COVID-19 CXRs and exhibited the highest accuracy (up to a mean of 97.10%, SD 1.08%) and sensitivity (up to a mean of 97.77%, SD 1.71%); however, there was only a minimal difference in the specificity and precision between the 2-input and 16-input MI-CNNs (approximately 1%-2% smaller than those of the 4-inputs MI-CNN).

As shown in the receiver operating characteristic (ROC) curves in Figure 6, all MI-CNNs with different inputs had good efficiency in the classification of COVID-19 and normal CXR, with the 4-inputs MI-CNN showing the best performance (Figure 6a). Similar to the testing data sets, the MI-CNNs had an area under the ROC curve (AUC) value of 0.98 for all inputs. However, the MI-CNNs had much higher AUC values (over 0.99) than that of the 1-input CNN (mean 0.982, SD 0.005), and the 4-inputs and 16-input MI-CNNs had the largest AUC values overall (0.995) (Figure 6b).

Figure 4. (a) Accuracy-epoch and (b) loss-epoch curves of the 1-input convolutional neural network (CNN) and 2-, 4-, and 16-input CNNs. Each curve represents the average of 5 five-fold cross-validation; the learning rate was 0.01 in all cases.

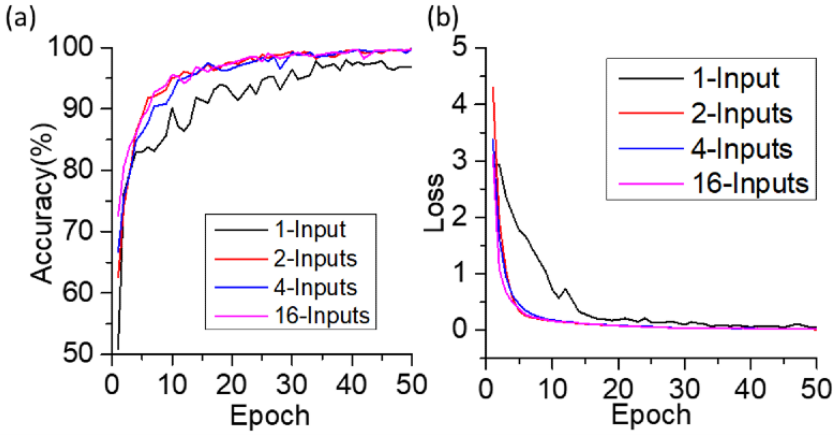


Figure 5. (a) Accuracy, (b) sensitivity, (c) specificity, and (d) precision of the of 1-input convolutional neural network (CNN), and the 2-, 4-, and 16-input CNNs.

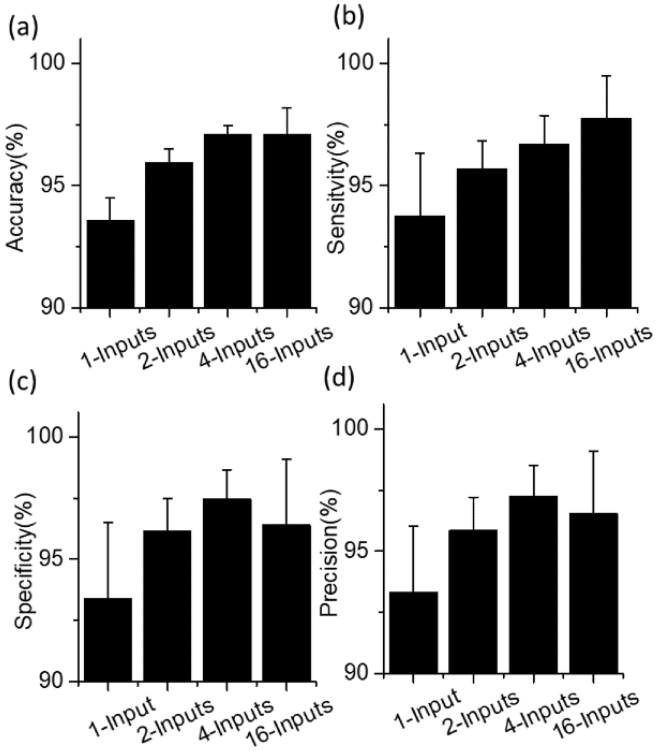
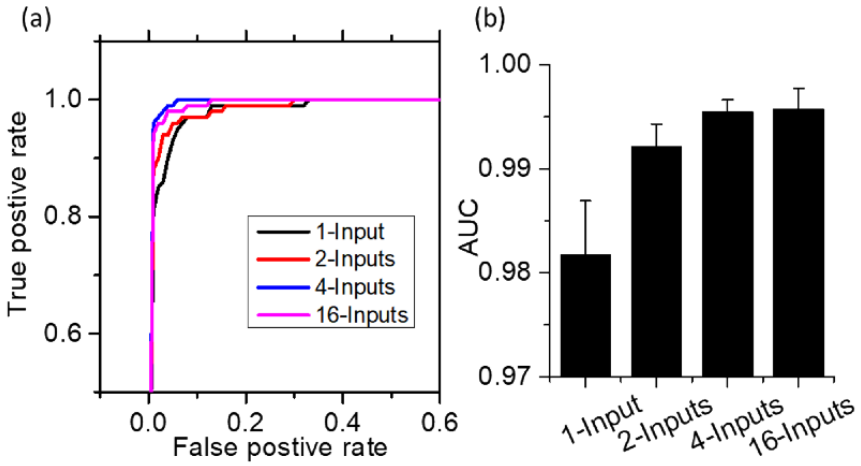


Figure 6. Receiver operating characteristic curves (a) and area under the curve (AUC) values (b) of the 1-input convolutional neural network (CNN), and 2-, 4-, and 16-input CNNs.



Classification of COVID-19 and Normal CXRs With Left- and Right-Lung ROI Data Sets

Figure 7 indicates the training accuracy epoch and loss epoch of the ROI data sets with the 1-input CNN and 2-inputs MI-CNN. Left- and right-lung ROI curves were obtained from the corresponding data sets of left or right lung regions, which were trained with the 1-input CNN. The LR-ROI curves were obtained from the left and right data sets, serving as the two inputs of the 2-inputs MI-CNN. All of these models were run with 100 epochs and a 0.001 learning rate, and were repeated five times.

From Figure 7, it can be seen that the 1-input CNN (both left- and right-lung ROIs) exhibited a higher accuracy curve and lower loss curve than the 2-inputs MI-CNN with the LR-ROI data set; however, all three methods showed similar accuracy (~97.0%) and loss (~0.05) at the end of the training. The LR-ROI data set resulted in approximately 15% higher accuracy than that of the left and right ROI data sets at the beginning of the training.

For the LR-ROI testing data sets (Figure 8), MI-CNNs showed good efficiency for the classification of COVID-19 CXRs, although the accuracy, sensitivity, specificity, and precision were slightly lower than those of the whole-image data sets (but still above 90%). There was almost no difference in the accuracy between the 1-input CNN and 2-inputs MI-CNN in the accuracy (~92%) for all three data sets (Figure 8a). LR-ROI showed the largest sensitivity (up to a mean of 94.55%, SD 2.90%), followed by the right ROI (93.30%, SD 3.27%) and then the left ROI with a slightly lower value (91.27%, SD 2.64%). By contrast, the left-lung ROI data set had larger specificity (mean 93.09%, SD 0.23%) and precision (mean 92.71%, SD 0.52%) than those of the right-lung ROI and LR-ROI (both approximately 90%).

Therefore, the three ROIs showed similar efficiency (similar accuracy) in classifying COVID-19 CXRs. Compared to the left-lung ROI method, higher sensitivity with the LR-ROI and right-lung ROI data sets indicated that the two CNN models (especially the 2-inputs MI-CNN) had better capability to identify COVID-19 CXRs from the normal CXRs correctly. Left-lung ROI had a lower probability of falsely recognizing normal CXRs as COVID-19 CXRs. Overall, LR-ROI and right-lung ROI had better efficiency in detecting COVID-19 CXRs, while the left-lung ROI method was better for identifying normal CXRs.

Based on the ROC curve, the 2-inputs MI-CNN with the LR-ROI data sets also showed relatively better performance than the 1-input CNN (left- or right-lung ROI) in classifying COVID-19 and normal CXRs, given its larger AUC value (mean 0.980, SD 0.005), as shown in Figure 9. In the 1-input CNN, the right-lung ROI data set showed better efficiency (mean AUC 0.975, SD 0.008) in identifying COVID-19 CXRs than the left-lung ROI (mean AUC 0.972, SD 0.008), as shown in Figure 9b.

In addition, LR-ROI data sets were also evaluated using the 4-input and 16-input MI-CNNs (see Figures S1 and S2 in Multimedia Appendix 1). The results showed almost no difference among different inputs of MI-CNNs, although the 2-inputs model had much higher sensitivity (mean 94.55%, SD 2.90%) than that of the 4-input (mean 92.72%, SD 4.37%) and 16-input (mean 93.42%, SD 2.25%) MI-CNNs.

Figure 7. (a) Accuracy-epoch and (b) loss-epoch curves of the 1-input convolutional neural network (CNN) and 2-inputs CNN with the left-lung region of interest (L-ROI), right-lung region of interest (R-ROI), and left and right lung region of interest (LR-ROI) data sets. Each curve represents the average of five replicates; the learning rate was 0.001.

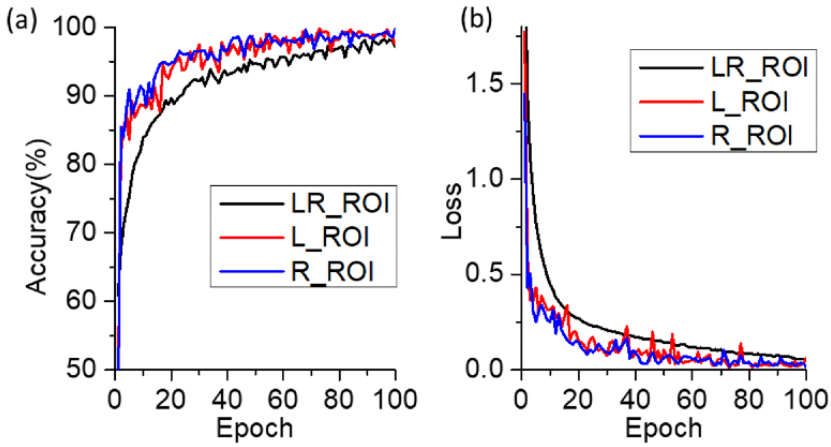


Figure 8. (a) Accuracy, (b) sensitivity, (c) specificity, and (d) precision of the 1-input convolutional neural network (CNN) with the left-lung region of interest (L-ROI) and right-lung region of interest (R-ROI), and the 2-inputs CNNs with the combined left- and right-lung region of interest (LR-ROI) data set.

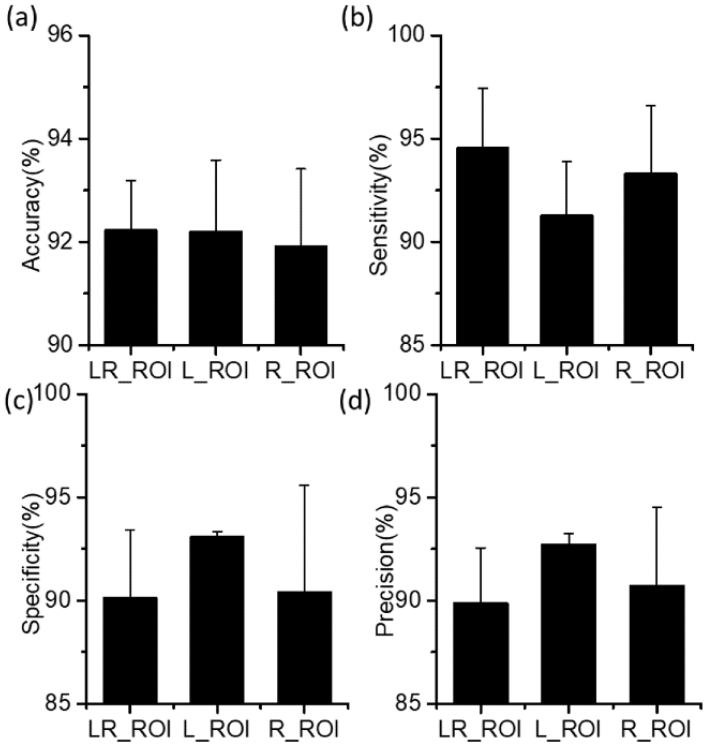
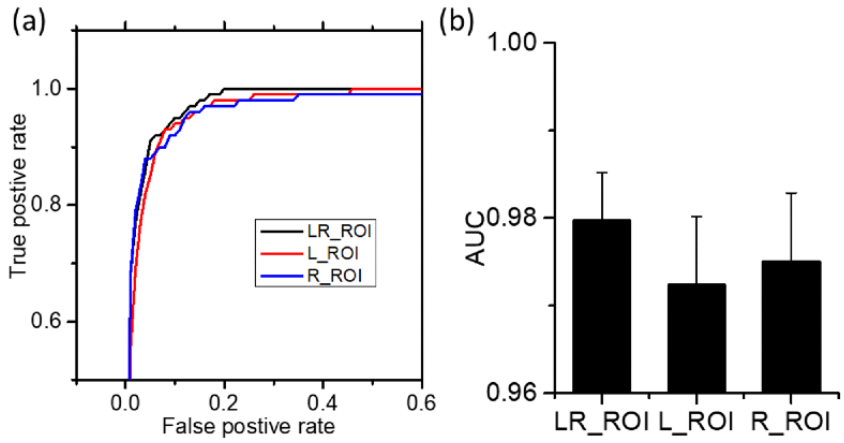


Figure 9. Receiver operating characteristic curves (a) and area under the curve (AUC) values (b) of the 1-input convolutional neural network (CNN) with the left-lung region of interest (L-ROI) and right-lung region of interest (R-ROI), and the 2-inputs CNN with the left- and right-lung region of interest (LR-ROI).



Screening of Critical Regions for COVID-19 Identification Using MI-CNNs

Regarding the region contributions, 90% of the whole-image and LR-ROI data sets were trained under the same training conditions as shown in Figure 4 and Figure 7, but repeated 10 times. The percentages of the correctly classified images from each image region were then calculated following the procedures outlined in Figure 3.

For the 2-inputs MI-CNN, both the whole-image and LR-ROI data sets showed that more COVID-19 CXRs were classified by the R1 regions (right-lung ROI), while more normal CXRs were recognized by the R2 regions (left-lung ROI) (Figures 10a and 11a). However, over 86.5% of COVID-19 CXRs were identified through the R1 regions in LR-ROI data sets, while only 13.5% were identified through the R2 regions. Moreover, normal CXRs showed slight changes in R1 (~30%) and R2 (~70%) between the whole-image and LR-ROI data sets.

Compared to the 2-inputs MI-CNN, all four regions contributed to the COVID-19 classification in the 4-inputs MI-CNN. In the whole-image data sets, R2 had the largest contributions to both COVID-19 and normal CXRs, whereas R3 had the lowest contributions. However, R4 regions showed the greatest difference in the classification of COVID-19 and normal images, which was approximately 35% in normal CRXs but only 10% in COVID-19 CRXs (Figure 10b). In the LR-ROI data sets, COVID-19 had the largest image percentage in the R3 regions (up to 60%) and the lowest image percentage (only 10%) in R2 regions (Figure 11b). In normal CXRs, the largest image percentage was found in R2 (up to 50%, the lowest region for COVID-19), followed by R3 (up to 30%) (Figure 11b).

In the 16-inputs MI-CNN, the critical regions became more obvious because smaller regions were used as MI-CNN inputs. From the whole-image data sets in Figure 10c, R1 regions had the largest contributions in COVID-19 CXRs, accounting for approximately 22% of the total accurately classified COVID-19 images. R6 had the second-largest contribution (accounting for approximately 15% of the correctly classified images). Compared to COVID-19 CXRs, the greatest difference in normal CXRs was found in the R9 regions (up to 20% vs ~3% in COVID-19 images), and R4 regions had the largest contributions (up to 27%) (Figure 10c).

In LR-ROI data sets, the critical regions and irrelevant regions become more clear. In COVID-19 CRXs, significant regions could be found in the R1, R2, R5, and R9 regions, especially R5 accounting for approximately 35%. These regions had almost no contribution in normal CXRs, whereas the greatest critical regions in normal CXRs were R10, R12, and R14, which were much higher (with each contribution reaching up to 20%) than other regions. These regions had almost no contributions to COVID-19 classification.

From the 16-inputs MI-CNN, nonlung regions were found to play critical roles in classifying COVID-19 (eg, R16 contributed up to 15% of testing images). When combining the left- and right-lung regions in 4- and 16-input MI-CNNs, the left lung had a greater contribution to the classification of both COVID-19 and normal CXRs (Figure 12a), which was not consistent with the results for the 2-inputs MI-CNN (Figure 12a). However, if removing the nonlung region R16, the right lung had a greater contribution (approximately 60% of classified testing images) in the classification of COVID-19 CXRs. In comparison, a greater contribution (also approximately 60% of classified testing images) was found from the left lung to classify normal CXRs (Figure 12b). Thus, it seems that COVID-19 CXRs could be more efficiently classified from the right-lung data sets, while normal CXRs were much more easily found through the left-lung data sets. From the color mapping to CXRs shown in Figure S4 of Multimedia Appendix 1, most of the critical regions for the classification of COVID-19 CXRs were distributed in the right-lung regions, while those for normal CXRs were in the left-lung regions.

Figure 10. Screening and evaluating the critical regions (R1-R16) of the whole-image data sets through the initial inputs and final activations of the classifier part of multiple-input convolutional neural networks (MI-CNNs): (a) 2-input MI-CNN; (b) 4-input MI-CNN; (c) 16-input MI-CNN.

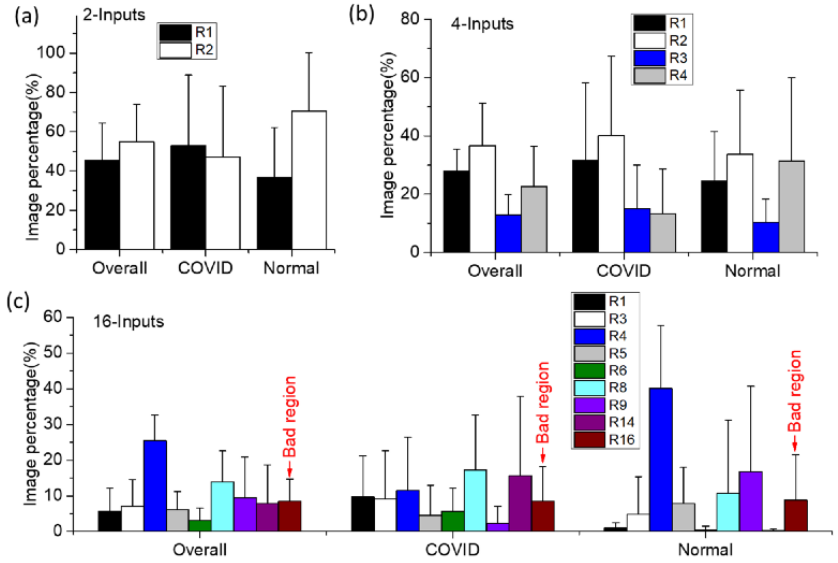


Figure 11. Screening and evaluation of the critical regions of the lung imaging data sets (R1-R16) through the initial inputs and final activations of the classifier part of multiple-input convolutional neural networks (MI-CNNs) with more than two inputs: (a) 2-inputs MI-CNN; (b) 4-inputs MI-CNN; (c) 16-inputs MI-CNN. L-ROI: left-lung region of interest; LR-ROI: left- and right-lung region of interest; R-ROI: right-lung region of interest.

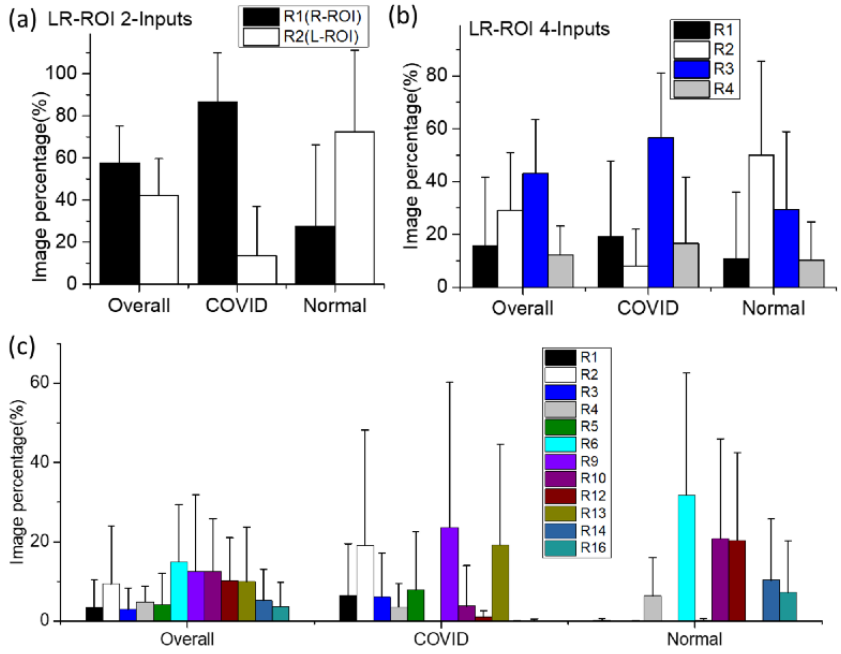
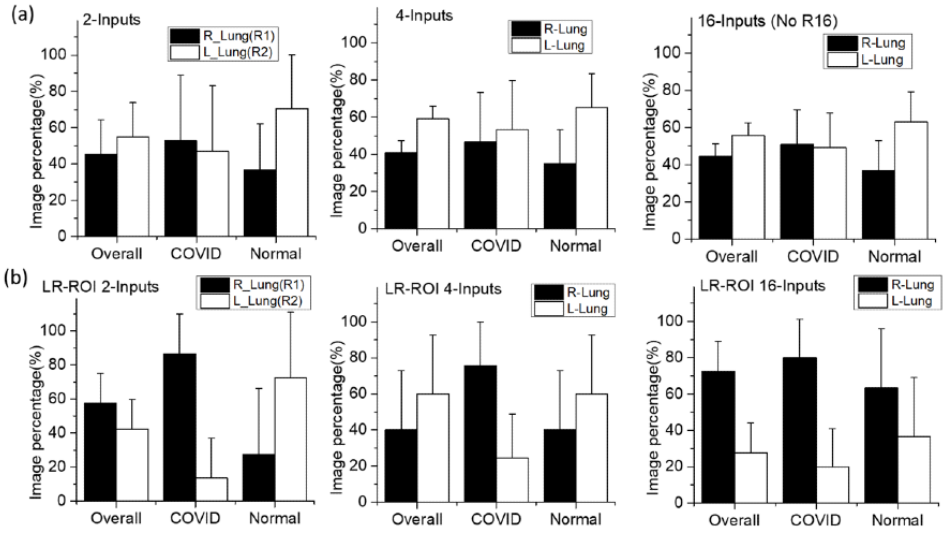


Figure 12. Contributions of the left-lung regions (L–Lung) and right-lung regions (R–Lung) to classifying COVID-19 chest X-ray radiographs (CXR) using multiple-input convolutional neural networks. (a) Whole-image CXR data sets; (b) left- and right-lung region of interest (LR-ROI) data sets.

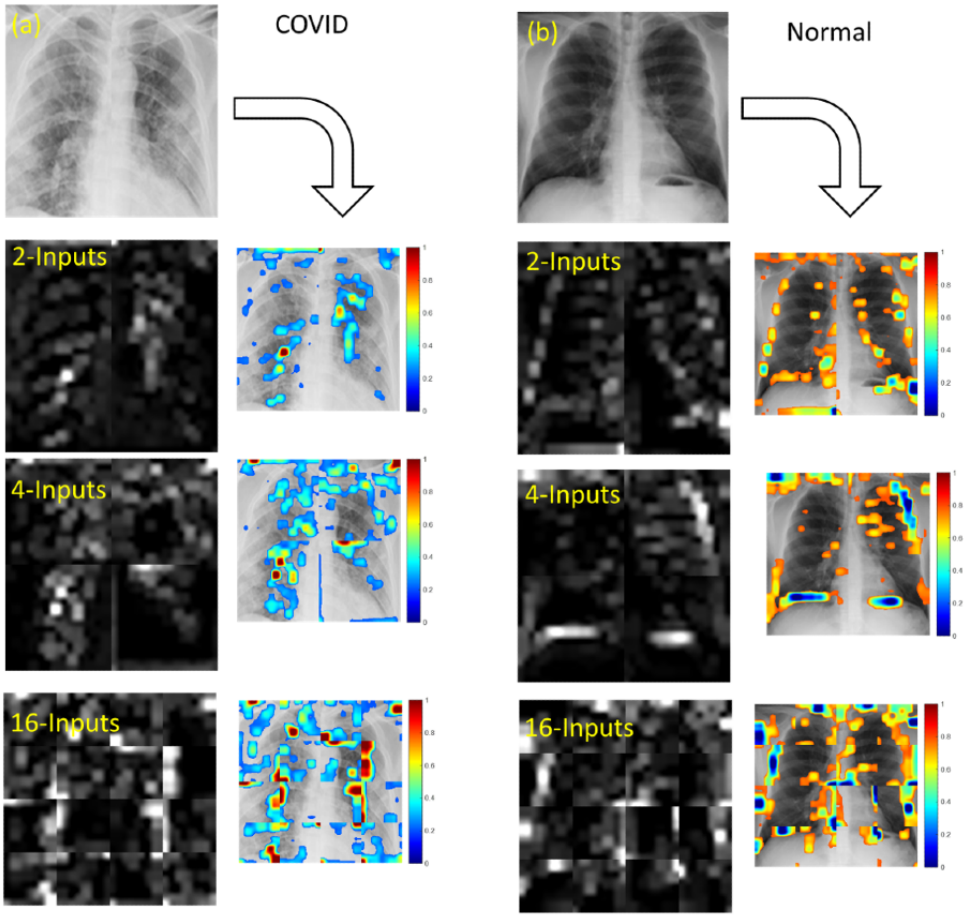


Visualization of the CNN Features From the Max-Pooling Layers in MI-CNNs With Different Inputs

Figure 13 provides a visual representation of the CNN features of MI-CNNs for the LR-ROI data sets. In COVID-19 CXRs, most of the strong-intensity pixels (red regions) in the visualized CNN features were distributed in the lung regions. Compared to the left lung, more color mapping was located in the right-lung regions, especially for the 4-input MI-CNN (Figure 13a). In normal CXRs (blue regions), most of the CNN features extracted by MI-CNNs with LR-ROI data sets could be found at the lung edges (Figure 13b).

Compared to the LR-ROI data sets, MI-CNNs had a much lower efficiency for extracting CNN features in the whole-image data sets. Most of the critical features (the strong-intensity pixels) were found in the nonlung regions (Figure S6a in Multimedia Appendix 1). In the normal CXRs, several of the CNN features (the blue regions in the color map) were distributed at the lung edges; however, some features were found in the nonlung regions (Figure S6b in Multimedia Appendix 1).

Figure 13. Visualization and mapping of convolutional neural network (CNN) features extracted from the strongest activations of the max-pooling layers using multiple-input CNNs with different numbers of inputs and the left- and right-lung region of interest data sets. (a) CNN features of COVID-19 chest X-ray radiographs (CXRs); (b) CNN features of normal CXRs. In color mapping, the red regions indicate COVID-19 features and the blue regions indicate normal features.



Discussion

Principal Findings

Currently, the COVID-19 pandemic is still deeply impacting the world and threatening many people's lives [1]. CXR remains one of the most commonly used imaging modalities for diagnosing COVID-19 owing to its advantages of low radiation, lack of side effects, economic feasibility, and moderate sensitivity [11-16,38]. In this study, we developed a novel MI-CNN method to classify COVID-19 and normal CXR images. In the whole-image CXR training, the results showed that MI-CNNs exhibited good efficiency for COVID-19 classification with high training accuracy and testing performance (over 95% accuracy, sensitivity, specificity, and precision). Compared to the 1-input CNN, the MI-CNNs had higher efficiency in recognizing COVID-19 CXRs, especially for the 4- and 16-input MI-CNNs, with testing accuracy over 97%. MI-CNNs also showed higher accuracy than the 1-input CNN from the beginning of training progress. Therefore, splitting the CXRs into different regions could improve the efficiency in classifying COVID-19 and normal CXRs. Regarding the learning rates, using a smaller learning rate (0.001) could indeed help to increase the accuracy, sensitivity, specificity, and precision from those of 1-input CNN models, by approximately 2%-3%. However, the learning rate had almost no effect for the MI-CNN models (2-input and 4-input MI-CNNs, as shown in Figure S7 of Multimedia Appendix 1).

In this study, CXRs were evenly segmented into different regions, and each image segmentation could individually serve as one of the MI-CNN inputs. Through assessment of the image percentage of the accurately classified CXRs in the testing data sets, the contributions of each region could be evaluated; in particular, more detailed contributions of each region could be screened with more MI-CNN inputs (eg, 16-inputs MI-CNN). According to our results, some CNN features could allow the network to determine the correct classification, whereas some image features may cause serious misjudgment [34]. Although the whole-image data sets could obtain higher accuracy for COVID-19 identification than LR-ROI data sets, some of the contributions were from the nonlung regions. For instance, in the 16-inputs MI-CNN, R1 regions had accurate contributions for COVID-19 classification, but the nonlung region R16 also had remarkable contributions for COVID-19 classification. Therefore, if using medical images as single inputs, not all regions will provide the correct contributions for classifying COVID-19 CXRs. Some of the nonlung regions may give noticeable contributions falsely.

Moreover, extraction of the lung regions (LR-ROI data sets) could greatly help to extract the critical regions for COVID-19 classification. Compared to the whole-image data sets, the

critical regions could be found at R1, R2, R3, R5, R9, and R13 in the LR-ROI data sets. These regions significantly contributed to accurately classifying COVID-19 CXRs (eg, R2, R9, and R13 contributed up to approximately 65% of accurately classified COVID-19 CXRs), but had no significant contribution to the normal CXRs. In comparison, R6, R10, R12, R14, and R16 were found in normal CXRs. Among them, R6, R10, and R12 contributed to over 80% of the accurately classified normal CXRs, whereas these regions had almost no contribution to COVID-19 classification.

In addition, right-lung regions had a higher contribution in the classification of COVID-19 than left-lung regions. By using the 16-inputs MI-CNN, more critical regions were screened in the right lung. Moreover, the sensitivity of the right-lung data sets with the 1-input CNN (approximately 94%, as shown in Figure 8b) was higher than that of the left-lung data sets (approximately 91%), which also indicates that right-lung regions tend to be more efficient in the classification of COVID-19 CXRs. By contrast, excluding the LR-ROI with the 16-input MI-CNN (Figure 12b), most of the critical regions related to normal CXRs could be found in the left-lung regions, which was further demonstrated through the higher precision (94%) in the left-lung ROI (Figure 8c). Based on their distributions, more critical regions related to COVID-19 were also found in the right-lung regions. In contrast, more regions related to normal CXRs were found in the left lung, especially for the 16-inputs MI-CNN with LR-ROI data sets (Figure S4b in Multimedia Appendix 1).

Finally, from the visualized CNN features, MI-CNNs still had better feature extractions than the 1-input CNN. For the CXRs in cases of severe COVID-19 in the whole-image data sets, CNN features extracted by the 1-input CNN were mainly distributed in the lung regions (Figure S5 in Multimedia Appendix 1). However, for the CRXs in cases of mild COVID-19, most of the CNN features were found at the lung edges (not within lung regions). Therefore, using the 1-input CNN to extract CNN features would be highly impacted by the quality of the whole-image CXRs. However, LR-ROI data sets could provide higher accuracy for COVID-19 classification than the whole-image data sets. Most of the critical features related to COVID-19 classification were distributed within the lung regions. This point is consistent with the evaluations of the critical regions, in which LR-ROI data sets could give more accurate COVID-19 classifications than the whole-image data sets. MI-CNNs tended to identify the normal CXRs from the edges of the lung regions in the LR-ROI data sets. Most of the strong-intensity regions in the visualized CNN features were distributed around the lung edges.

Comparison With Prior Work

Although previous studies reported that deep-learning methods could achieve very high accuracy in classifying COVID-19 CXRs or CT scans, most of these analyses were based on whole images as the single input, and the specific regions that contribute to the successful classification of COVID-19 have barely been explored [8,9,13,15,16,18,20,26-29]. Compared to the single-input CNN model, MI-CNNs will provide more CNN features for the classification, which could improve the accuracy of the entire system [33,34]. However, most previous studies focused on whole images with different formats as the CNN inputs [33,34], and few studies attempted to segment a medical image into different regions as the CNN inputs and evaluate their contributions to the final classification of disease images. Therefore, our MI-CNNs could independently extract features from different regions; more importantly, the contribution of each region could be evaluated through the MI-CNN models.

Moreover, compared to the traditional CNN models GoogLeNet and ResNet_50, our proposed 1-input model could achieve similar performance with the same data set (whole CXRs) and learning rate (0.001). Our model also required much less time (only approximately 14 minutes for each training), whereas GoogLeNet and ResNet_50 needed respectively more than 130 minutes and 300 minutes, representing an increase of 10 times and 20 times than required with our models.

Limitations

Although the MI-CNNs could achieve good efficiency in the classification of COVID-19 CXRs and screen the critical regions and features related to COVID-19, there are still several major limitations of this study. First, the size of the COVID-19 and normal data set is still small, and more CXRs are required to further test the reliability of our MI-CNNs. Second, the feature visualization exhibits relatively low efficiency. Other algorithms such as GRAD can be used to better map the critical features to the original CXR. Third, all of the MI-CNN models used the same structures of convolutional layers for all CXR regions. More complicated structures could be further explored in the future. For example, different regions could use different convolutional designs, such as the lung boundary with fewer convolutional layers and lung regions with more convolutional layers. Finally, the severity of COVID-19 cannot currently be evaluated with MI-CNN models, especially from the critical features. Finally, more parameters (eg, image resolution) could be used to better evaluate the accuracy and performance of MI-CNNs.

Conclusions

In summary, each MI-CNN input could individually process only one part of CXRs, which contributes to the highly efficient classification of the COVID-19 CXRs. In the whole-image data sets, MI-CNNs could achieve better classification efficiency (over 95% accuracy, sensitivity, specificity, and precision) than the 1-input CNN. In addition, the performance of MI-CNNs increased with the number of inputs, especially for the 4- and 16-input MI-CNNs with over 97% accuracy. In the LR-ROI data sets, the MI-CNNs showed an approximate 4% decrease in the classification of COVID-19 CXRs compared to the whole-image data sets. Some nonlung regions (eg, R16) had positive contributions to COVID-19 classification (also shown in the visualized CNN features), which fraudulently increased the higher performance in the whole-image data sets. Therefore, compared to the whole-image data sets, LR-ROI data sets could provide a more accurate evaluation for the contribution of each region, as well as the extraction of CNN features.

From the analysis of the contributions of critical regions in the testing data sets, the right lung had a greater contribution to the

classification of COVID-19 CXRs. However, the left-lung regions had a greater contribution to classifying normal CXRs. From LR-ROI data sets, MI-CNNs were sensitive to the lung edges and found more important features distributed around the lung edges in normal CXRs. For COVID-19 CXRs, visualized CNN features were primarily distributed within the lung regions (especially in the 16-inputs MI-CNN).

In conclusion, MI-CNNs have excellent efficiency in classifying COVID-19 CXRs. More MI-CNN inputs usually result in better classification efficiency. Our method could assist radiologists in automatically screening the regions playing critical roles in the classification of COVID-19 from CXRs.

Acknowledgments

This research was supported by Louisiana State University (LSU) Faculty Research Grants (Grant No. 009875); LSU Leveraging Innovation for Technology Transfer (LIFT2) Grant (LSU-2021-LIFT-009, LSU-2020-LIFT-008); Health Sciences Center New Orleans, Louisiana State University Grant (HSCNO-2019-LIFT-004); Louisiana Board of Regents Grant (LEQSF (2018-21)-RD-A-09); National Institutes of Health (1U01AA029348-01); and the National Science Foundation CAREER award (2046929).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supporting information; Figures S1-S8.
[DOCX File , 3275 KB - bioinform_v3i1e36660_app1.docx]

References

1. Aslan MF, Unlersen MF, Sabanci K, Durdu A. CNN-based transfer learning-BiLSTM network: a novel approach for COVID-19 infection detection. Appl Soft Comput 2021 Jan;98:106912 [FREE Full text] [doi: 10.1016/j.asoc.2020.106912] [Medline: 33230395]
2. WHO Coronavirus (COVID-19) Dashboard. World Health Organization. 2021. URL: https://covid19.who.int/ [accessed 2022-07-20]
3. COVID-19 advice for the public: Getting vaccinated. World Health Organization. 2021. URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/covid-19-vaccines/advice [accessed 2021-07-20]
4. Kupferschmidt K, Wadman M. Delta variant triggers new phase in the pandemic. Science 2021 Jun 25;372(6549):1375-1376. [doi: 10.1126/science.372.6549.1375]
5. Mahase E. Covid-19: Omicron and the need for boosters. BMJ 2021 Dec 14;375:n3079. [doi: 10.1136/bmj.n3079] [Medline: 34906956]
6. Karim SSA, Karim QA. Omicron SARS-CoV-2 variant: a new chapter in the COVID-19 pandemic. Lancet 2021 Dec 11;398(10317):2126-2128 [FREE Full text] [doi: 10.1016/S0140-6736(21)02758-6] [Medline: 34871545]
7. Torjesen I. Covid restrictions tighten as omicron cases double every two to three days. BMJ 2021 Dec 09;375:n3051. [doi: 10.1136/bmj.n3051] [Medline: 34887256]
8. Arias-Londono JD, Gomez-Garcia JA, Moro-Velazquez L, Godino-Llorente JI. Artificial intelligence applied to chest X-ray images for the automatic detection of COVID-19. A thoughtful evaluation approach. IEEE Access 2020;8:226811-226827 [FREE Full text] [doi: 10.1109/ACCESS.2020.3044858] [Medline: 34786299]
9. Islam MZ, Islam MM, Asraf A. A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images. Inform Med Unlocked 2020;20:100412 [FREE Full text] [doi: 10.1016/j.imu.2020.100412] [Medline: 32835084]
10. Kremer S, Lersy F, de Sèze J, Ferré JC, Maamar A, Carsin-Nicol B, et al. Brain MRI findings in severe COVID-19: a retrospective observational study. Radiology 2020 Nov;297(2):E242-E251 [FREE Full text] [doi: 10.1148/radiol.2020202222] [Medline: 32544034]
11. Abbas A, Abdelsamea MM, Gaber MM. Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network. Appl Intell 2021 Sep 05;51(2):854-864 [FREE Full text] [doi: 10.1007/s10489-020-01829-7] [Medline: 34764548]
12. Bui MM, Smith P, Agresta SV, Cheong D, Letson GD. Practical issues of intraoperative frozen section diagnosis of bone and soft tissue lesions. Cancer Control 2008 Jan 01;15(1):7-12 [FREE Full text] [doi: 10.1177/107327480801500102] [Medline: 18094656]
13. Misra S, Jeon S, Lee S, Managuli R, Jang I, Kim C. Multi-channel transfer learning of chest X-ray images for screening of COVID-19. Electronics 2020 Aug 27;9(9):1388. [doi: 10.3390/electronics9091388]

14. Rahman T, Khandakar A, Qiblawey Y, Tahir A, Kiranyaz S, Abul Kashem SB, et al. Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. Comput Biol Med 2021 May;132:104319 [FREE Full text] [doi: 10.1016/j.compbiomed.2021.104319] [Medline: 33799220]
15. Ismael AM, Şengür A. Deep learning approaches for COVID-19 detection based on chest X-ray images. Expert Syst Appl 2021 Feb;164:114054 [FREE Full text] [doi: 10.1016/j.eswa.2020.114054] [Medline: 33013005]
16. Haque K, Abdelgawad A. A deep learning approach to detect COVID-19 patients from chest X-ray images. AI 2020 Sep 22;1(3):418-435 [FREE Full text] [doi: 10.3390/ai1030027]
17. Candemir S, Antani S. A review on lung boundary detection in chest X-rays. Int J Comput Assist Radiol Surg 2019 Apr;14(4):563-576 [FREE Full text] [doi: 10.1007/s11548-019-01917-1] [Medline: 30730032]
18. Alam N, Ahsan M, Based MA, Haider J, Kowalski M. COVID-19 detection from chest X-ray images using feature fusion and deep learning. Sensors 2021 Feb 20;21(4):1480 [FREE Full text] [doi: 10.3390/s21041480] [Medline: 33672585]
19. Zhao X, Wu Y, Song G, Li Z, Zhang Y, Fan Y. A deep learning model integrating FCNNs and CRFs for brain tumor segmentation. Med Image Anal 2018 Jan;43:98-111 [FREE Full text] [doi: 10.1016/j.media.2017.10.002] [Medline: 29040911]
20. Narin A, Kaya C, Pamuk Z. Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks. Pattern Anal Appl 2021 May 09;24(3):1207-1220 [FREE Full text] [doi: 10.1007/s10044-021-00984-y] [Medline: 33994847]
21. Huang K, Lin C, Lee Y, Wu Z. A deep learning and image recognition system for image recognition. Data Sci Pattern Recognition 2019;3(2):1-11.
22. Ghosh S, Das N, Das I, Maulik U. Understanding deep learning techniques for image segmentation. ACM Comput Surv 2020 Jul 31;52(4):1-35. [doi: 10.1145/3329784]
23. Li Z, Li Z, Chen Q, Zhang J, Dunham ME, McWhorter AJ, et al. Machine-learning-assisted spontaneous Raman spectroscopy classification and feature extraction for the diagnosis of human laryngeal cancer. Comput Biol Med 2022 Jul;146:105617. [doi: 10.1016/j.compbiomed.2022.105617] [Medline: 35605486]
24. Zhou P, Liu Z, Wu H, Wang Y, Lei Y, Abbaszadeh S. Automatically detecting bregma and lambda points in rodent skull anatomy images. PLoS One 2020 Dec;15(12):e0244378 [FREE Full text] [doi: 10.1371/journal.pone.0244378] [Medline: 33373400]
25. Li Z, Li Z, Chen Q, Ramos A, Zhang J, Boudreaux JP, et al. Detection of pancreatic cancer by convolutional-neural-network-assisted spontaneous Raman spectroscopy with critical feature visualization. Neural Netw 2021 Dec;144(4):455-464. [doi: 10.1016/j.neunet.2021.09.006] [Medline: 34583101]
26. Loey M, Smarandache F, Khalifa NEM. Within the lack of chest COVID-19 X-ray dataset: a novel detection model based on GAN and deep transfer learning. Symmetry 2020 Apr 20;12(4):651. [doi: 10.3390/sym12040651] [Medline: 33177550]
27. Wang L, Lin ZQ, Wong A. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. Sci Rep 2020 Nov 11;10(1):19549. [doi: 10.1038/s41598-020-76550-z] [Medline: 33177550]
28. Apostolopoulos ID, Mpesiana TA. Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks. Phys Eng Sci Med 2020 Jun;43(2):635-640 [FREE Full text] [doi: 10.1007/s13246-020-00865-4] [Medline: 32524445]
29. de Sousa PM, Carneiro PC, Oliveira MM, Pereira GM, da Costa Junior CA, de Moura LV, et al. COVID-19 classification in X-ray chest images using a new convolutional neural network: CNN-COVID. Res Biomed Eng 2021 Jan 04;38(1):87-97 [FREE Full text] [doi: 10.1007/s42600-020-00120-5] [Medline: 32895587]
30. Abraham B, Nair MS. Computer-aided detection of COVID-19 from X-ray images using multi-CNN and Bayesnet classifier. Biocybern Biomed Eng 2020 Jun;40(4):1436-1445 [FREE Full text] [doi: 10.1016/j.bbe.2020.08.005] [Medline: 32895587]
31. Toğaçar M, Ergen B, Cömert Z. COVID-19 detection using deep learning models to exploit Social Mimic Optimization and structured chest X-ray images using fuzzy color and stacking approaches. Comput Biol Med 2020 Jun;121:103805 [FREE Full text] [doi: 10.1016/j.compbiomed.2020.103805] [Medline: 32568679]
32. Elmoufidi A, Skouta A, Jai-Andaloussi S, Ouchetto O. CNN with multiple inputs for automatic glaucoma assessment using fundus images. Int J Image Grap 2022 Jan 10:1-8 [FREE Full text] [doi: 10.1142/S0219467823500122]
33. Trivedi P, Mhasakar P, Prakash S, Mitra SK. Multichannel CNN for facial expression recognition. 2017 Presented at: PReMI: International Conference on Pattern Recognition and Machine Intelligence; December 5-8, 2017; Kolkata, India.
34. Lin C, Lin C, Jeng S. Using feature fusion and parameter optimization of dual-input convolutional neural network for face gender recognition. Appl Sci 2020 May 01;10(9):3166-3168. [doi: 10.3390/app10093166]
35. Sun Y, Zhu L, Wang G, Zhao F. Multi-input convolutional neural network for flower grading. J Electr Comput Eng 2017;2017:1-8. [doi: 10.1155/2017/9240407]
36. Chowdhury MEH, Rahman T, Khandakar A, Mazhar R, Kadir MA, Mahbub ZB, et al. Can AI help in screening viral and COVID-19 pneumonia? IEEE Access 2020;8:132665-132676 [FREE Full text] [doi: 10.1109/ACCESS.2020.3010287]
37. COVID-19 Radiography Database. Kaggle. URL: https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database [accessed 2021-07-20]
38. Zotin A, Hamad Y, Simonov K, Kurako M. Lung boundary detection for chest X-ray images classification based on GLCM and probabilistic neural networks. Proc Comput Sci 2019;159:1439-1448. [doi: 10.1016/j.procs.2019.09.314]

Abbreviations

AUC: area under the receiver operating characteristic curve
CNN: convolutional neural network
CT: computed tomography
CXR: chest X-ray radiography
FC: Fully connected layer
LR-ROI: left- and right-lung region of interest
MI-CNN: multiple-inputs convolutional neural network
PCR: polymerase chain reaction
ReLu: rectified linear unit
ROC: receiver operating characteristic
ROI: region of interest
RT-PCR: reverse transcription real-time polymerase chain reaction
WHO: World Health Organization

Edited by A Mavragani; submitted 19.01.22; peer-reviewed by JA Benítez-Andrades, D Oladele, Z Alhassan, J Sang; comments to author 10.08.22; revised version received 23.08.22; accepted 29.08.22; published 04.10.22.

Please cite as:
Li Z, Li Z, Yao L, Chen Q, Zhang J, Li X, Feng JM, Li Y, Xu J
Multiple-Inputs Convolutional Neural Network for COVID-19 Classification and Critical Region Screening From Chest X-ray Radiographs: Model Development and Performance Evaluation
JMIR Bioinform Biotech 2022;3(1):e36660
URL: https://bioinform.jmir.org/2022/1/e36660
doi:10.2196/36660
PMID:36277075

Original Paper

Identification of Potential Vaccine Candidates Against SARS-CoV-2 to Fight COVID-19: Reverse Vaccinology Approach

Ekta Gupta[1], MSc; Rupesh Kumar Mishra[2], PhD; Ravi Ranjan Kumar Niraj[2], PhD

[1]Dr. B. Lal Institute of Biotechnology, Jaipur, India

[2]Amity University Rajasthan, Jaipur, India

Corresponding Author:
Ravi Ranjan Kumar Niraj, PhD
Amity University Rajasthan
SP-1, Kant Kalwar
RIICO Industrial Area
Jaipur, 303002
India
Phone: 91 9729559580
Email: rrkniraj@gmail.com

Abstract

Background: The recent emergence of COVID-19 has caused an immense global public health crisis. The etiological agent of COVID-19 is the novel coronavirus SARS-CoV-2. More research in the field of developing effective vaccines against this emergent viral disease is indeed a need of the hour.

Objective: The aim of this study was to identify effective vaccine candidates that can offer a new milestone in the battle against COVID-19.

Methods: We used a reverse vaccinology approach to explore the SARS-CoV-2 genome among strains prominent in India. Epitopes were predicted and then molecular docking and simulation were used to verify the molecular interaction of the candidate antigenic peptide with corresponding amino acid residues of the host protein.

Results: A promising antigenic peptide, GVYFASTEK, from the surface glycoprotein of SARS-CoV-2 (protein accession number QIA98583.1) was predicted to interact with the human major histocompatibility complex (MHC) class I human leukocyte antigen (HLA)-A*11-01 allele, showing up to 90% conservancy and a high antigenicity value. After vigorous analysis, this peptide was predicted to be a suitable epitope capable of inducing a strong cell-mediated immune response against SARS-CoV-2.

Conclusions: These results could facilitate selecting SARS-CoV-2 epitopes for vaccine production pipelines in the immediate future. This novel research will certainly pave the way for a fast, reliable, and effective platform to provide a timely countermeasure against this dangerous virus responsible for the COVID-19 pandemic.

(JMIR Bioinform Biotech 2022;3(1):e32401) doi:10.2196/32401

KEYWORDS

COVID-19; SARS-CoV-2; reverse vaccinology; molecular docking; molecular dynamics simulation; vaccine candidates; vaccine; simulation; virus; peptide; antigen; immunology; biochemistry; genetics

Introduction

COVID-19 began in December 2019 with an outbreak of a novel virus in Wuhan city of China [1]. The disease gained a rapid foothold worldwide, resulting in the World Health Organization (WHO) declaring it a global pandemic by March 2020. As of March 10, 2021, there has been a worldwide total of 118,159,602 cases and 2,622,101 deaths due to COVID-19 reported by the WHO. The virus causing COVID-19, SARS-CoV-2, spreads primarily through saliva, droplets, or discharges from the nose of an infected person after coughing or sneezing. Coronaviruses are enveloped RNA viruses with the largest genome among all RNA viruses [2]. As continuous transmission of the virus across borders increases, imposing a major health burden on the global scale, more studies are urgently required to understand SARS-CoV-2. Moreover, in the absence of effective cures and drugs, vaccination or immunization therapy is imperative to target the entire population. In particular, immunoinformatics tools have proven to be crucial to move the vaccine development pipeline forward

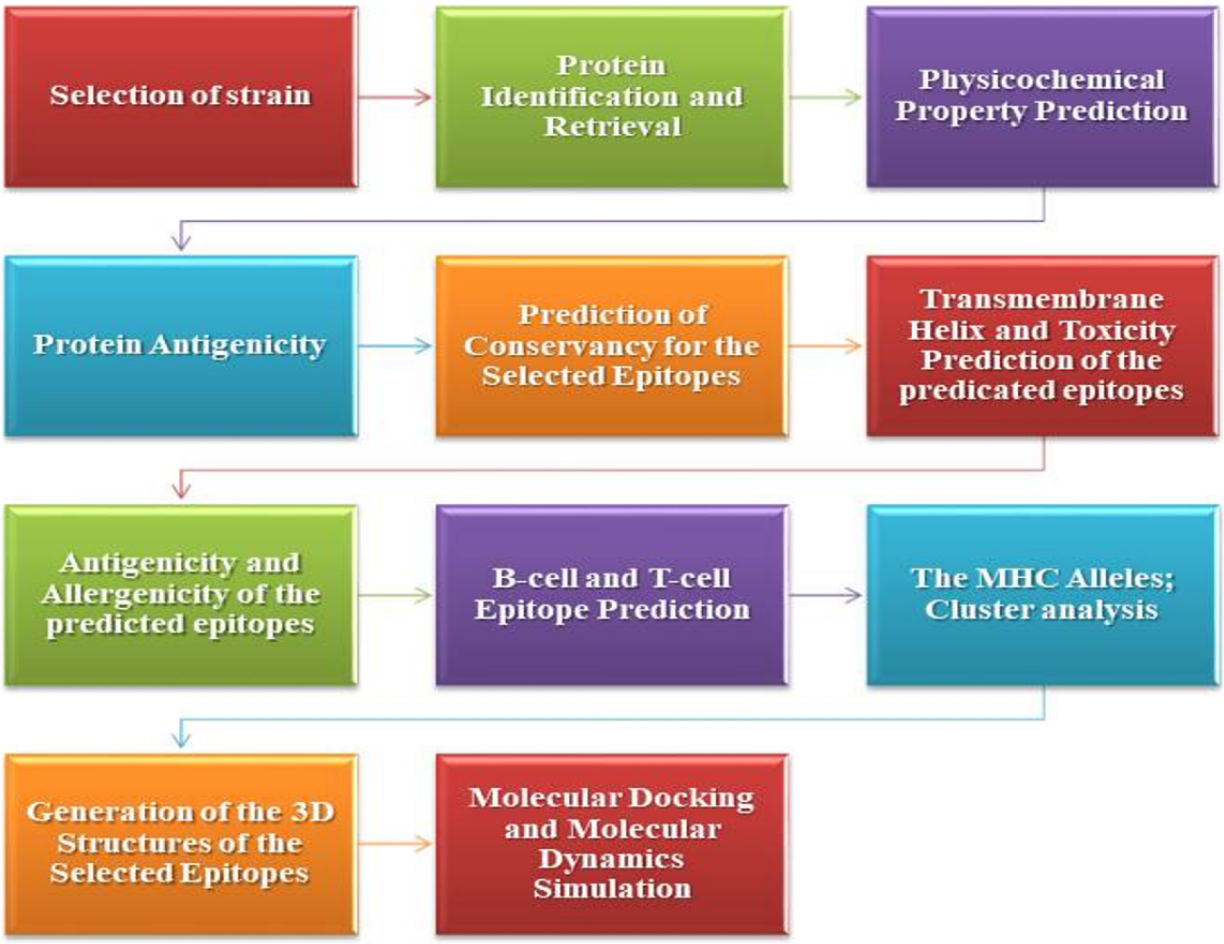
[3]. Since there is relatively little knowledge about the pathogenesis of the virus, an immunoinformatics-based approach to investigate the immunogenic epitopes for further vaccine development is required [4].

Since COVID-19 has affected almost the entire world's population, binding of promiscuous epitopes to a variety of human leukocyte antigen (HLA) alleles is vital for larger dissemination. Toward this end, in silico approaches will be remarkably useful in helping to develop a cure as quickly as possible [5]. Antibody generation by the activation of B cells as well as acute viral clearance by T cells along with virus-specific memory generation by CD8+ T cells are analogously important to develop immunity against the virus [6]. The SARS-CoV-2 spike (S) protein is considered to be highly antigenic, and thereby can evoke strong immune responses and generate neutralizing antibodies that can block attachment of the virus to host cells [7].

In reverse vaccinology, various in silico biology tools are used to discover novel antigens by studying the genetic makeup of a pathogen and the genes that could lead to identification of good epitopes. The reverse vaccinology approach thus offers a fast and cost-effective vaccine discovery platform [8]. With this approach, a novel antigen is identified using omics analysis of the target organism. In silico analysis combined with the reverse vaccinology approach facilitates an easier and time- and labor-saving process of antigen discovery [9].

Herein, we explored the proteome of SARS-CoV-2 strains prominent in the Indian geographical region against the human host to identify potential antigenic proteins and epitopes that can effectively elicit a cellular-mediated immune response against the virus. With this approach, we identified a promising antigenic peptide, GVYFASTEK, from a surface glycoprotein (protein accession number QIA98583.1) of SARS-CoV-2, which was predicted to interact with human major histocompatibility complex (MHC) alleles and displayed up to 90% conservancy and significant antigenicity. Molecular docking analysis further confirmed the molecular interaction of the prime antigenic peptide with the residues of the HLA-A*11-01 allele for MHC class I. An overview of the study design is provided in Figure 1. After careful evaluation, this peptide was determined to be an appropriate epitope for eliciting a strong cell-mediated immune response against SARS-CoV-2. The outcomes from this significant analysis could help to select appropriate SARS-CoV-2 epitopes for multiepitope vaccine production pipelines in the near future. This novel research will certainly pave the way for a fast, reliable, and effective platform to provide a timely countermeasure against this dangerous pandemic disease.

Figure 1. Diagrammatic representation of the methodology. MHC: major histocompatibility complex.



Methods

Strain Selection

The highly virulent strain of SARS-CoV-2 was selected for the in silico analysis. The complete genome of SARS-CoV-2 is available in the National Center for Biotechnology Information database under reference NC_045512.2.

Protein Identification and Retrieval

The following 12 viral protein sequences of SARS-CoV-2 were retrieved from the ViPR database (Host: Human, Country: India) [10]: Orf10 protein (QIA98591.1), Orf8 protein (QIA98589.1), Orf7a protein (QIA98588.1), Orf6 protein (QIA98587.1), Orf3a protein (QIA98584.1), membrane glycoprotein (QIA98586.1), envelope protein (QIA98585.1), surface glycoprotein (QIA98583.1), surface glycoprotein (QHS34546.1), nucleocapsid protein (QII87776.1), nucleocapsid protein (QII87775.1), and nucleocapsid phosphoprotein (QIA98590.1).

Physicochemical Property Prediction

The online tool ProtParam of ExPASy [11] was used to predict various physicochemical properties of the selected protein sequences.

Protein Antigenicity

VaxiJen v2.0 [12] was used to predict the antigenicity of the selected proteins. This software uses the FASTA file format of amino acid sequences as input and then predicts antigenicity based on the physicochemical properties of proteins. The output is denoted according to an antigenic score [13]. During analysis, the threshold was maintained at 0.4 [9].

B Cell and T Cell Epitope Prediction

The B cell and T cell epitopes of the selected surface glycoprotein sequence were predicted via the Immune Epitope Database (IEDB), which contains a large amount of experimental data on epitopes and antibodies [14]. The IEDB enables performing a robust analysis on several epitopes in the context of various tools, including conservation across antigens, population coverage, and clusters with similar sequences [15]. To obtain MHC class I–restricted CD8+ cytotoxic T lymphocyte epitopes of the selected surface glycoprotein sequence, the NetMHCpan EL 4.0 prediction method was applied for the HLA-A*11-01 allele. MHC class II–restricted CD4+ helper T lymphocyte epitopes were obtained for the HLA DRB1*04-01 allele using the Sturniolo prediction method. The top 10 MHC class I and top 10 MHC class II epitopes were randomly selected based on their percentile scores and antigenicity scores. Five random B cell lymphocyte epitopes were selected based on their greater lengths using the Bipipered linear epitope prediction method [8].

Antigenicity and Allergenicity of the Predicted Epitopes

VaxiJen v2.0 was utilized to predict protein antigenicity. During antigenicity analysis, the threshold was maintained at 0.4 [9]. The allergenicity of the selected epitopes was predicted via AllerTOP v2 [16].

Transmembrane Helix and Toxicity Prediction of the Predicted Epitopes

The transmembrane helix of the selected epitopes was predicted using the TMHMM v2.0 server [17], which predicts whether the epitope would be in the transmembrane region, or remain inside or outside of the membrane. The toxicity prediction of the selected epitopes was carried out via the ToxinPred server [18].

Prediction of Conservation of the Selected Epitopes

The conservation analysis of the epitopes was performed via the epitope conservancy analysis tool of the IEDB server [15]. During analysis, the sequence identity threshold was maintained at ≥50 [8].

Cluster Analysis of MHC Alleles

Cluster analysis was carried out by MHCcluster 2.0 [19,20]. During cluster analysis, the number of peptides to be included was kept at 50,000 and the number of bootstrap calculations was set to 100. For cluster analysis, the NetMHCpan-2.8 prediction method was used.

Generation of 3D Structures of Selected Epitopes

The PEP-FOLD3 online tool [21] was used to predict the 3D structures of the selected best epitopes [22-24].

Molecular Docking and Molecular Dynamics Simulation

Molecular docking was carried out to depict the binding pattern of inhibitors with respective proteins. Predocking was carried out by UCSF Chimera [25]. The peptide-protein docking of the selected epitopes was carried out by the online docking tool PatchDock [26]. The results of PatchDock were refined and rescored by the FireDock server [27]. Docking was then performed by the HPEPDOCK server [28]. Docking pose analysis was performed using Ligplot [29]. The molecular simulation was executed with the GROMACS 2018.1 package using the Gromos43a1 force field [9]. Protein solvation was performed with the SPC water model in a cubic box ($10.8 \times 10.8 \times 10.8$ nm^3). The solvated protein system was processed for energy minimization using the steepest algorithm up to a maximum of 25,000 steps or until the maximum force was not greater than 1000 kJ/mol/nm, which is the default threshold. The NVT and NPT ensembles for 50,000 steps (100 ps) were run at 300 K and 1 atm. The system was first equilibrated using the NVT ensemble followed by the NPT ensemble. The final molecular dynamic simulation was performed for the dock complex of the GVYFASTEK epitope docked against the HLA-A*11-01 allele (Protein Data Bank [PDB] ID 5WJL). Finally, the simulations were evaluated according to the root mean square deviation (RMSD) and root mean square fluctuation (RMSF) of atomic positions for the complete episode of simulations. All steps were similar across simulations, except that the final molecular dynamics simulation was carried out for 50 ns.

Results

Selection and Retrieval of Viral Protein Sequences

The SARS-CoV-2 strain was identified and 12 viral protein sequences against the human host in India were retrieved from the ViPR database and selected for possible vaccine candidate identification (Table 1). The FASTA sequences of the proteins are given in Multimedia Appendix 1.

Table 1. SARS-CoV-2 (Host: Human, Country: India) viral protein sequence identification and retrieval via the ViPR database.

Gene symbol	Protein name	GenBank nucleotide accession	GenBank protein accession
orf10	Orf10 protein	MT050493	QIA98591.1
orf8	Orf8 protein	MT050493	QIA98589.1
orf7a	Orf7a protein	MT050493	QIA98588.1
orf6	Orf6 protein	MT050493	QIA98587.1
orf3a	Orf3a protein	MT050493	QIA98584.1
M	Membrane glycoprotein	MT050493	QIA98586.1
E	Envelope protein	MT050493	QIA98585.1
S	Surface glycoprotein	MT050493	QIA98583.1
S	Surface glycoprotein	MT012098	QHS34546.1
N	Nucleocapsid protein	MT163715	QII87776.1
N	Nucleocapsid protein	MT163714	QII87775.1
N	Nucleocapsid phosphoprotein	MT050493	QIA98590.1

Physicochemical Property Analysis and Protein Antigenicity

Analysis of physicochemical properties of the 12 proteins, including amino acids, molecular weight, theoretical isoelectric point (pI), extinction coefficient (M^{-1} cm^{-1}), estimated half-life (in mammalian cells), instability index, aliphatic index, and grand average of hydropathicity (GRAVY), were predicted (Table 2). With a fixed threshold of 0.4, all proteins were predicted to be antigenic (Table 3). The physicochemical analysis revealed that the surface glycoprotein (QIA98583.1) had the highest extinction coefficient of 148,960 M^{-1} cm^{-1} and the lowest GRAVY value of –0.077 among the proteins. In addition, the surface glycoprotein was stable and antigenic; therefore, we selected this protein for further analysis.

Table 2. Physiochemical properties of SARS-CoV-2 viral proteins.

Gene symbol	Amino acids	Molecular weight	Theoretical pI[a]	Extinction coefficient (M^{-1} cm^{-1})	Half-life in mammalian cells (hours)	Instability index	Aliphatic index	GRAVY[b]
orf10	38	4449.23	7.93	4470	30	16.06 (stable)	107.63	0.637
orf8	121	13,804.93	5.42	16,305	30	46.24 (unstable)	94.13	0.181
orf7a	121	13,744.17	8.23	7825	30	48.66 (unstable)	100.74	0.318
orf6	61	7272.54	4.60	8480	30	31.16 (stable)	130.98	0.233
orf3a	275	31,122.94	5.55	58,705	30	32.96 (stable)	103.42	0.275
M	222	25,146.62	9.51	52,160	30	39.14 (stable)	120.86	0.446
E	75	8365.04	8.57	6085	30	38.68 (stable)	144.00	1.128
S	1273	141,206.52	6.24	148,960	30	33.01 (stable)	84.82	–0.077
S	1272	140,972.27	6.16	147,470	30	32.78 (stable)	85.05	–0.071
N	88	9827.08	10.23	8480	4.4	36.54 (stable)	61.14	–1.067
N	133	14,363.88	11.37	8480	1	58.97 (unstable)	44.21	–1.170
N	419	45,625.70	10.07	43,890	30	55.09 (unstable)	52.53	–0.971

[a]pI: isoelectric point.

[b]GRAVY: grand average of hydropathicity.

Table 3. Antigenicity prediction of SARS-CoV-2 viral proteins (threshold value: 0.4).

Protein name	Antigenicity score	Antigenicity
Orf10 protein	0.7185	Antigenic
Orf8 protein	0.6063	Antigenic
Orf7a protein	0.6441	Antigenic
Orf6 protein	0.6131	Antigenic
Orf3a protein	0.4945	Antigenic
Membrane glycoprotein	0.5102	Antigenic
Envelope protein	0.6025	Antigenic
Surface glycoprotein	0.4654	Antigenic
Surface glycoprotein	0.4687	Antigenic
Nucleocapsid protein	0.5767	Antigenic
Nucleocapsid protein	0.6235	Antigenic
Nucleocapsid phosphoprotein	0.5059	Antigenic

T Cell and B Cell Epitope Prediction

The T cell epitopes of MHC class I were determined by the NetMHCpan EL 4.0 prediction method of the IEDB server with the sequence length set to 9. The server-generated epitopes were further analyzed based on the antigenicity scores and percentile scores, and the top 10 potential epitopes were selected randomly for antigenicity, allergenicity, toxicity, and conservancy tests. The server ranks the predicted epitopes in ascending order of percentile scores (Table 4). The T cell epitopes of MHC class II (HLA-DRB1*04-01 allele) of the protein were also determined by the IEDB server (Table 5) using Sturniolo prediction methods. The top 10 ranked epitopes of the protein were selected randomly for further analysis. Additionally, the B cell epitopes of the protein were selected using the Bipipered linear epitope prediction method of the IEDB server, with the selection of epitopes based on greater lengths (Figure 2).

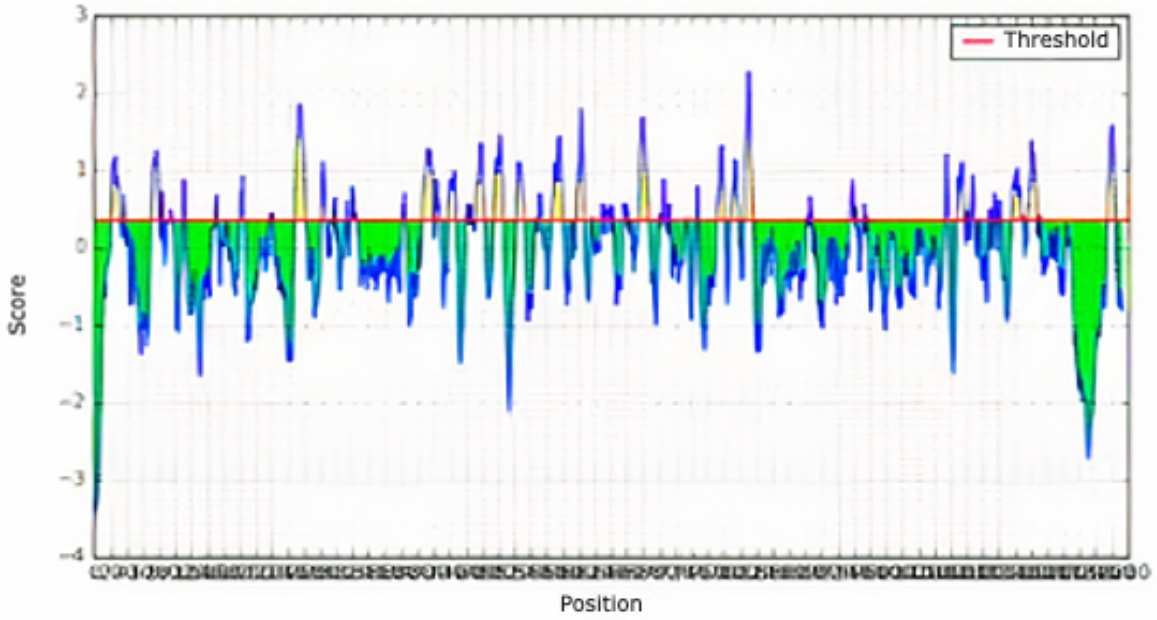
Table 4. Major histocompatibility complex class I epitopes of SARS-CoV-2 surface glycoprotein (QIA98583.1).

Epitope	Start	End	Topology	Antigenicity	Antigenicity score	Allergenicity	Toxicity	Minimum identity (%)	Conservancy (%)
GVYFASTEK	19	27	Inside	Yes	0.7112	Nonallergen	Nontoxic	11.11	100
VTYVPAQEK	15	23	Inside	Yes	0.8132	Allergen	Nontoxic	22.22	100
ASANLAATK	40	48	Inside	Yes	0.7041	Allergen	Nontoxic	22.22	100
TLADAGFIK	57	65	Inside	Yes	0.5781	Nonallergen	Nontoxic	22.22	100
TLKSFTVEK	22	30	Inside	No	0.0809	Allergen	Nontoxic	11.11	100
NSASFSTFK	20	28	Inside	No	0.1232	Allergen	Nontoxic	11.11	100
TEILPVSMTK	24	33	Inside	Yes	1.4160	Allergen	Nontoxic	10.00	100
SSTASALGK	29	37	Outside	Yes	0.6215	Allergen	Nontoxic	22.22	100
GTHWFVTQR	49	57	Inside	No	0.0723	Allergen	Nontoxic	11.11	100
EILPVSMTK	25	33	Inside	Yes	1.6842	Allergen	Nontoxic	11.11	100

Table 5. Major histocompatibility class II epitopes of SARS-CoV-2 surface glycoprotein (QIA98583.1).

Epitope	Start	End	Topology	Antigenicity	Antigenicity score	Allergenicity	Toxicity	Minimum identity (%)	Conservancy (%)
SNFRVQPTESI	36	46	Inside	Yes	0.9897	Allergen	Nontoxic	11.11	100
NFRVQPTESIV	37	47	Inside	Yes	1.0669	Nonallergen	Nontoxic	22.22	100
FRVQPTESIVR	38	48	Inside	No	0.3493	Allergen	Nontoxic	9.09	100
VYYHKNNKSWM	3	13	Inside	No	0.3726	Allergen	Nontoxic	18.18	100
LGVYYHKNNKS	1	11	Inside	Yes	0.8696	Allergen	Nontoxic	9.09	100
GVYYHKNNKSW	2	12	Inside	Yes	0.6685	Allergen	Nontoxic	9.09	100
LLIVNNATNVV	47	57	Inside	Yes	0.4166	Nonallergen	Nontoxic	9.09	100
LIVNNATNVVI	48	58	Inside	No	0.2045	Nonallergen	Nontoxic	9.09	100
IVNNATNVVIK	49	59	Inside	No	0.2274	Allergen	Nontoxic	9.09	100
VFVSNGTHWFV	44	54	Outside	No	0.0957	Allergen	Nontoxic	18.18	100

Figure 2. B cell epitope prediction for the surface glycoprotein of SARS-CoV-2 (QIA98583.1).



Topology Identification of Epitopes

The topology of the selected epitopes was determined by the TMHMM v2.0 server. Table 4 and Table 5 represent the potential T-cell epitopes of selected surface glycoprotein. Table 6 shows the potential B cell epitopes with their respective topologies.

Table 6. B cell epitopes of SARS-CoV-2 surface glycoprotein (QIA98583.1).

Epitope	Topology	Antigenicity	Allergenicity
RTQLPPAYTNS	Inside	Antigen	Allergen
SGTNGTKRFDN	Inside	Antigen	Allergen
LTPGDSSSGWTAG	Outside	Antigen	Nonallergen
VRQIAPGQTGKIAD	Inside	Antigen	Nonallergen
YQAGSTPCNGV	Inside	Nonantigen	Nonallergen
QIAPGQTGKIAD	Inside	Antigen	Nonallergen
YGFQPTNGVGYQ	Outside	Antigen	Allergen
RDIADTTDAVRDPQ	Inside	Antigen	Allergen
QTQTNSPRRARSV	Inside	Nonantigen	Nonallergen
ILPDPSKPSKRS	Outside	Antigen	Nonallergen

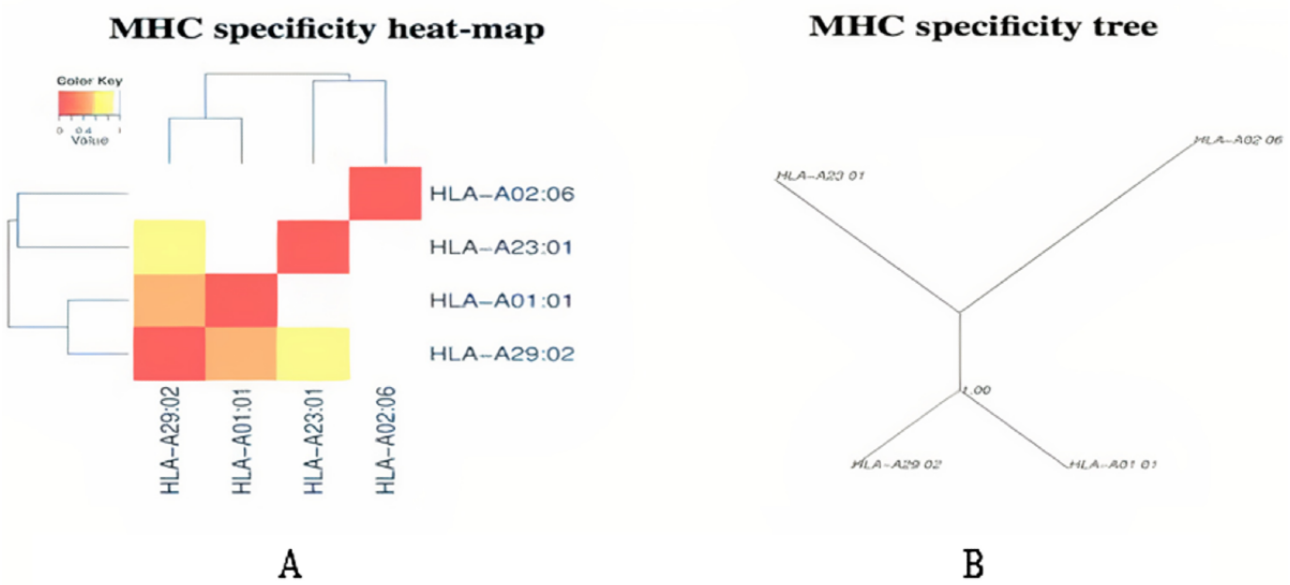
Antigenicity, Allergenicity, Toxicity, and Conservancy Analysis of Epitopes

The selected T cell epitopes were found to be highly antigenic as well as nonallergenic, nontoxic, and had a conservancy greater than 90%. Among the 10 selected MHC class I epitopes and 10 selected MHC class II epitopes, a total of four epitopes were selected based on the above-mentioned criteria: GVYFASTEK, TLADAGFIK, NFRVQPTESI, and LLIVNNATNV.

Cluster Analysis of MHC Alleles

The cluster analysis of the MHC class I alleles that possibly interact with the predicted epitopes was carried out by the online tool MHCcluster 2.0, which generates clusters of alleles phylogenetically. The results are shown in Figure 3, in which the red zone indicates a strong interaction and the yellow zone corresponds to a weaker interaction.

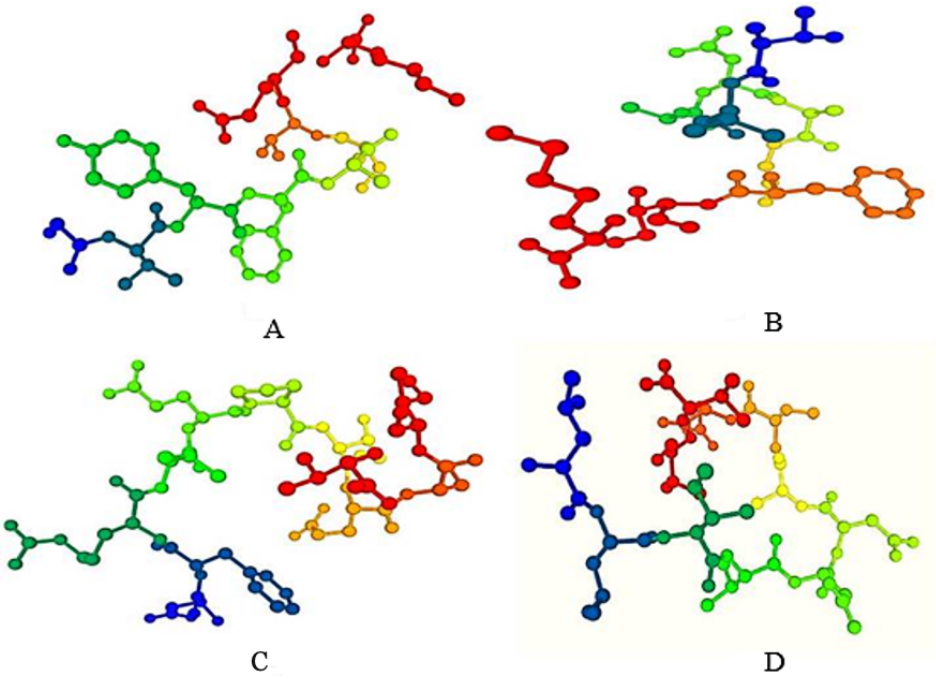
Figure 3. Major histocompatibility complex (MHC) class cluster analysis. (A) Heat map. (B) Specificity tree. The red zone indicates a strong interaction and the yellow zone corresponds to a weaker interaction.



Three-Dimensional Structure Prediction (Modeling) of Epitopes

All T cell epitopes were subjected to 3D structure prediction with the PEP-FOLD3 server, which were used for peptide-protein docking (Figure 4).

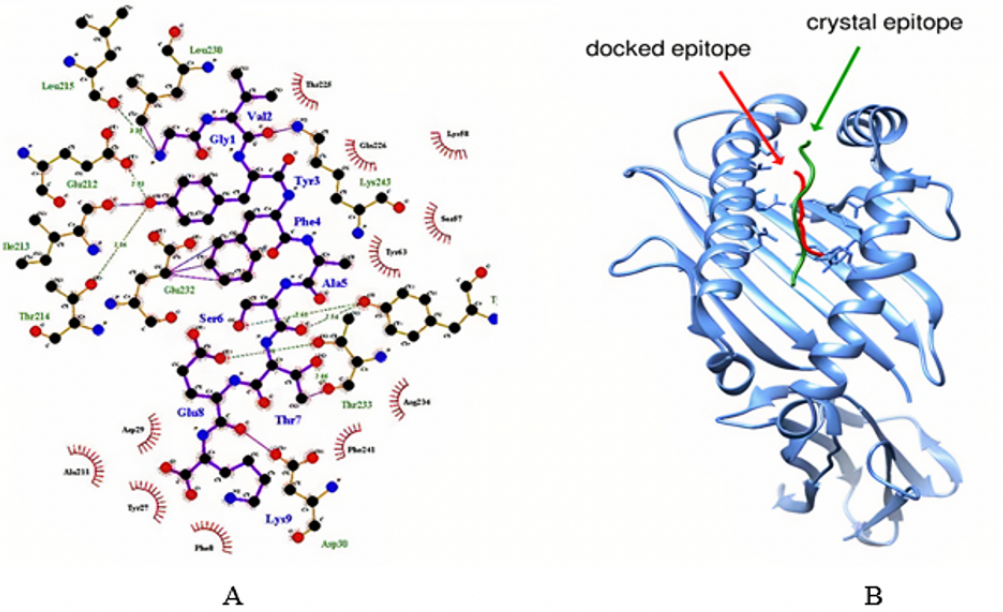
Figure 4. Three-dimensional structure generation of T-cell epitopes by the PEP-FOLD3 server. Epitope representation: (A) GVYFASTEK, (B) TLADAGFIK, (C) NFRVQPTESI, and (D) LLIVNNATNV.



Peptide-Protein Docking and Vaccine Candidate Prioritization

Molecular docking was performed to determine whether all of the identified epitopes could bind with MHC class I and MHC class II molecules. The selected epitopes docked with the HLA-A*11-01 allele (PDB ID 5WJL) and HLA-DRB1*04-01 allele (PDB ID 5JLZ). The docking was performed using the PatchDock online docking tool and refined by the FireDock online server. Results were also analyzed by the HPEPDOCK server (see Figure S1 in Multimedia Appendix 1). Among the four epitopes, the selected glycoprotein QIA98583.1, GVYFASTEK (MHC class I epitope), showed the best result with the lowest global energy of –52.82. Further, the docking pose was analyzed via Ligplot (Figure 5a) and the docking site can be visualized in Figure 5b. We also identified highly antigenic and nonallergenic B cell vaccine candidates LTPGDSSSGWTAG and VRQIAPGQTGKIAD from the selected surface glycoprotein (QIA98583.1).

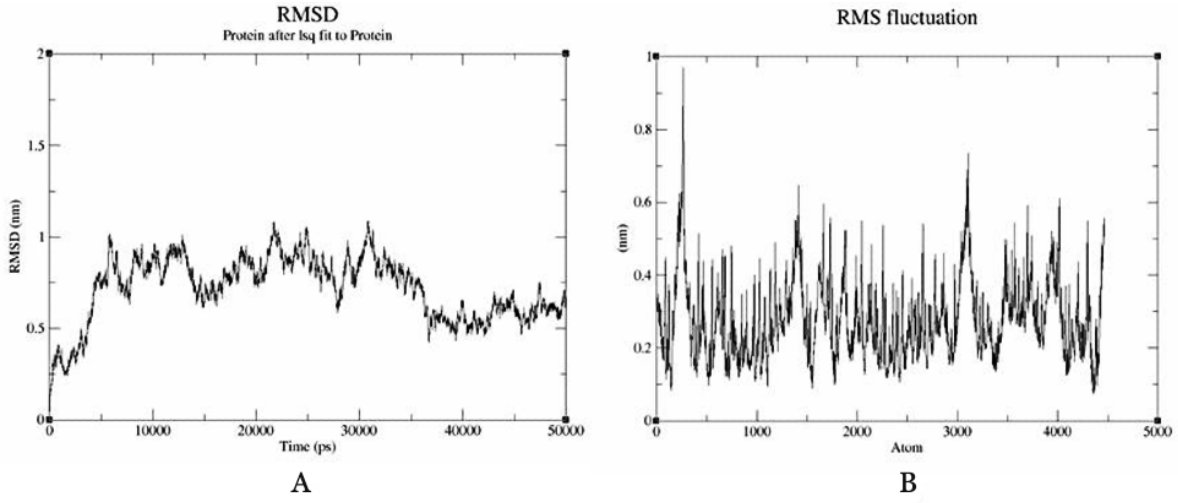
Figure 5. (A) Docking pose analysis via LigPlot (GVYFASTEK epitope docking against the HLA-A*11-01 allele [PDB ID: 5WJL]). Molecular docking result showing protein-ligand interaction. Oxygen (O), nitrogen (N), and carbon (C) atoms are represented by red, blue, and black circles, respectively. (B) Molecular docking analysis showing that the docking site of the ligand (GVYFASTEK epitope) in our study is similar to the ligand used in the crystal structure of the HLA-A*11-01 allele (PDB ID: 5WJL).



Molecular Dynamics Simulation

Molecular dynamics simulation of the dock complex of the GVYFASTEK epitope docked against the HLA-A*11-01 allele (PDB ID 5WJL) was successfully executed for 50 ns. The complex became stable throughout the simulation with an RMSD fluctuation of 0.3-1.0 nm from the original position (Figure 6a). In most cases, residues lying in the core protein regions have low RMSF values while exposed loops have high RMSF values (Figure 6b). The peaks in the graph show a value between 0.1 and 0.6 nm. Both these results indicate that the protein complexes were stable throughout the molecular docking simulations, demonstrating that the proteins possess good ability for stability.

Figure 6. Molecular dynamics simulation. (A) Root mean square deviation (RMSD) and (B) root mean square fluctuation (RMSF) graphs of the dock complex (GVYFASTEK epitope docked against the HLA-A*11-01 allele [PDB ID: 5WJL]).



Discussion

Principal Findings

A vaccine is an enormously imperative and expansively formed therapeutic product. Millions of infants, children, and adults are vaccinated every year. However, the development and research processes of vaccines are expensive and occasionally require countless months to prepare and advance an appropriate vaccine candidate toward eliminating a pathogen. There are currently innumerable tools and approaches of immunoinformatics, computer-aided drug design, bioinformatics, and converse/reverse vaccinology to extensively progress vaccine design and preparations, which in turn help to reduce the duration and cost investment for vaccine expansion [8,30].

In this study, physicochemical analysis revealed that the SARS-CoV-2 surface glycoprotein QIA98583.1 exhibited the highest extinction coefficient of 148,960 M^{-1} cm^{-1} and the lowest GRAVY value of –0.077 among the identified viral proteins. In addition, this selected surface glycoprotein was highly stable (instability index <40) and antigenic. The antigenicity of the protein was determined by the VaxiJen V2.0 server. If a compound has a variability index greater than 40, it means that the product is considered to be unbalanced [31]. The extinction coefficient refers to the quantity of light that is captured by a complex at a particular wavelength [32,33]. Various physicochemical properties, including the number of amino acids, molecular mass/weight, theoretical pI, extinction coefficient, uncertainty index, aliphatic index, and GRAVY, were resolved by the ProtParam server [34].

The two major functioning immune cells are B and T lymphocytic cells, which are responsible for several defensive roles in the body. Once identified by an antigen-presenting cell (APC; eg, dendritic cells and macrophages), the antigen is accessible by the MHC class II molecule existing on the surface of APCs to helper T cells. Subsequently, the helper T cell acquires a CD4+ fragment on its surface, designated as a CD4+ T cell. Once stimulated by an APC, helper T cells subsequently stimulate B cells, yielding antibody-producing plasma B cells alongside memory B cells. Plasma B cells harvest several antibodies and memory B cells function in long-term immunological memory. Moreover, macrophages and CD8+ cytotoxic T cells are also triggered by helper T cells to ultimately abolish the target antigen [35-39].

The possible B and T cell epitopes of the selected SARS-CoV-2 viral protein were identified by the IEDB server [14], which generates and ranks the epitopes based on their antigenicity scores and percentile scores. The top 10 MHC class I and class II epitopes were engaged for this investigation. The topology of the precise epitopes was resolved by the TMHMM v2.0 server [17]. In all inflammatory situations such as allergenicity, antigenicity, toxicity, and conservancy examinations, the T cell epitopes were found to be exceedingly antigenic with a higher immune response without allergenicity or toxicity, and showed a conservancy of over 90%. Among the 10 certain MHC class I and 10 selected MHC class II epitopes of the protein, four epitopes were designated based on the revealed properties, GVYFASTEK, TLADAGFIK, NFRVQPTESI, and LLIVNNATNVV, along with antigenic and nonallergenic B cell epitopes that were selected for additional vaccine candidate investigation. Cluster examination of the conceivable MHC class I and MHC class II alleles that might interact with the

predicted epitopes was performed by the online tool MHC cluster 2.0 [20]. The antigenicity, demarcated as the capability of an extraneous ingredient to act as an antigen and stimulate B and T cell responses over their epitope, correspondingly identifies the antigenic determinant portion [40]. The allergenicity is defined as the capability of that ingredient to act as an allergen and induce latent allergic responses in the host [41].

Moreover, cluster analysis of the MHC class I and II alleles was similarly performed to categorize their association with each other and group them based on their functionality and predicted specificity [19]. In the following steps, peptide-protein docking was performed among the selected epitopes and MHC alleles. The MHC class I epitopes remained docked to the MHC class I molecule (PDB ID 5WJL) and the MHC class II epitopes were docked to the MHC class II molecule (PDB ID 5JLZ) correspondingly. The peptide-protein docking was performed to evaluate the capability of the epitopes to interact with the MHC molecules. Predocking was performed by UCSF Chimera and then 3D structure generation of the epitopes was performed. The docking was executed by the PatchDock and FireDock servers and analyzed by the HPEPDOCK server constructed on global energy. The GVYFASTEK epitope demonstrated the best scores in the peptide-protein docking. All of the vaccine candidates proved to be potentially antigenic and nonallergenic, indicating that they should not cause any allergenic reaction within the host. However, more in vitro and in vivo examinations should be performed to confirm the safety, usefulness, and potential of the predicted vaccine candidates.

Conclusion

In the face of the enormous tragedy of suffering, demise, and social adversity caused by the COVID-19 pandemic, it is of extreme importance to develop an effective and safe vaccine against this disease. Bioinformatics, reverse vaccinology, and related technologies are widely used in vaccine design and development, since these technologies reduce costs and time. In this study, we first identified proteins belonging to SARS-CoV-2 against the human host from strains in India. The potential B cell and T cell epitopes that can effectively elicit cellular-mediated immune responses related to these selected proteins were then determined through robust processes. The potential T cell epitope (GVYFASTEK) and B cell epitopes (LTPGDSSSGWTAG, VRQIAPGQTGKIAD, QIAPGQTGKIAD, and ILPDPSKPSKRS) can play major roles in the development of new subunit and multiepitope vaccines. In brief, reverse vaccinology is confirmed as a reliable means to recognize novel vaccine candidates and their consequential application. This study can motivate further research in an innovative and efficient direction to deliver a fast, reliable, and significant platform in search of an effective and timely cure of COVID-19 caused by SARS-CoV-2.

Acknowledgments

RM acknowledges the financial support and award of the Ramalingaswami fellowship from the Department of Biotechnology, New Delhi, India. RN and EG acknowledge the Amity Institute of Biotechnology, Amity University Rajasthan, Jaipur, and Dr. B. Lal Institute of Biotechnology, Jaipur.

Authors' Contributions

EG: study protocol, data curation, software, analysis and validation, writing of original draft; RKM: writing, reviewing, and editing original draft; RRKN: conceptualization, protocol design, supervision, reviewing, editing, and finalizing original draft.

Conflicts of Interest

None declared.

Multimedia Appendix 1
SARS-CoV-2 protein sequences in FASTA format and HPEPDOCK server docking results (Figure S1).
[DOCX File , 157 KB - bioinform_v3i1e32401_app1.docx]

References

1. Holshue ML, DeBolt C, Lindquist S, Lofy KH, Wiesman J, Bruce H, Washington State 2019-nCoV Case Investigation Team. First case of 2019 novel coronavirus in the United States. N Engl J Med 2020 Mar 05;382(10):929-936 [FREE Full text] [doi: 10.1056/NEJMoa2001191] [Medline: 32004427]
2. Sinha SK, Shakya A, Prasad SK, Singh S, Gurav NS, Prasad RS, et al. An evaluation of different Saikosaponins for their potency against SARS-CoV-2 using NSP15 and fusion spike glycoprotein as targets. J Biomol Struct Dyn 2021 Jun 13;39(9):3244-3255 [FREE Full text] [doi: 10.1080/07391102.2020.1762741] [Medline: 32345124]
3. Mishra S, Sinha S. Immunoinformatics and modeling perspective of T cell epitope-based cancer immunotherapy: a holistic picture. J Biomol Struct Dyn 2009 Dec;27(3):293-306. [doi: 10.1080/07391102.2009.10507317] [Medline: 19795913]
4. Enayatkhani M, Hasaniazad M, Faezi S, Gouklani H, Davoodian P, Ahmadi N, et al. Reverse vaccinology approach to design a novel multi-epitope vaccine candidate against COVID-19: an study. J Biomol Struct Dyn 2021 May 02;39(8):2857-2872 [FREE Full text] [doi: 10.1080/07391102.2020.1756411] [Medline: 32295479]

5. Mishra S. T Cell epitope-based vaccine design for pandemic novel coronavirus 2019-nCoV. ChemRxiv. URL: https://chemrxiv.org/engage/chemrxiv/article-details/60c749b4469df4724cf43c17 [accessed 2022-03-22]
6. Enjuanes L, Zuñiga S, Castaño-Rodriguez C, Gutierrez-Alvarez J, Canton J, Sola I. Molecular basis of coronavirus virulence and vaccine development. Adv Virus Res 2016;96:245-286 [FREE Full text] [doi: 10.1016/bs.aivir.2016.08.003] [Medline: 27712626]
7. Du L, He Y, Zhou Y, Liu S, Zheng B, Jiang S. The spike protein of SARS-CoV--a target for vaccine and therapeutic development. Nat Rev Microbiol 2009 Mar 9;7(3):226-236 [FREE Full text] [doi: 10.1038/nrmicro2090] [Medline: 19198616]
8. Ullah MA, Sarkar B, Islam SS. Exploiting the reverse vaccinology approach to design novel subunit vaccines against Ebola virus. Immunobiology 2020 May;225(3):151949. [doi: 10.1016/j.imbio.2020.151949] [Medline: 32444135]
9. Gupta E, Gupta SRR, Niraj RRK. Identification of drug and vaccine Target in Mycobacterium leprae: a reverse vaccinology approach. Int J Pept Res Ther 2019 Oct 03;26(3):1313-1326. [doi: 10.1007/s10989-019-09936-x]
10. Pickett BE, Sadat EL, Zhang Y, Noronha JM, Squires RB, Hunt V, et al. ViPR: an open bioinformatics database and analysis resource for virology research. Nucleic Acids Res 2012 Jan;40(Database issue):D593-D598 [FREE Full text] [doi: 10.1093/nar/gkr859] [Medline: 22006842]
11. Walker J. The proteomics protocols handbook. Switzerland: Springer Nature; 2005.
12. Doytchinova IA, Flower DR. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. BMC Bioinformatics 2007 Jan 05;8(1):4 [FREE Full text] [doi: 10.1186/1471-2105-8-4] [Medline: 17207271]
13. Meunier M, Guyard-Nicodème M, Hirchaud E, Parra A, Chemaly M, Dory D. Identification of novel vaccine candidates against Campylobacter through reverse vaccinology. J Immunol Res 2016;2016:5715790. [doi: 10.1155/2016/5715790] [Medline: 27413761]
14. Immune Epitope Database and Analysis Resource. URL: https://www.iedb.org/ [accessed 2022-03-22]
15. Vita R, Mahajan S, Overton J, Dhanda S, Martini S, Cantrell J, et al. The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res 2019 Jan 08;47(D1):D339-D343 [FREE Full text] [doi: 10.1093/nar/gky1006] [Medline: 30357391]
16. AllerTOP v. 2.0. Bioinformatics tool for allergenicity prediction. URL: https://www.ddg-pharmfac.net/AllerTOP/ [accessed 2022-03-22]
17. TMHMM-2.0 Prediction of transmembrane helices in proteins. DTU Health Tech. URL: https://services.healthtech.dtu.dk/service.php?TMHMM-2.0 [accessed 2022-03-22]
18. ToxinPred. URL: https://webs.iiitd.edu.in/raghava/toxinpred/protein.php [accessed 2022-03-22]
19. Thomsen M, Lundegaard C, Buus S, Lund O, Nielsen M. MHCcluster, a method for functional clustering of MHC molecules. Immunogenetics 2013 Sep 18;65(9):655-665 [FREE Full text] [doi: 10.1007/s00251-013-0714-9] [Medline: 23775223]
20. MHC Cluster 2.0. DTU Health Tech. URL: https://services.healthtech.dtu.dk/service.php?MHCcluster-2.0 [accessed 2022-03-22]
21. PEP-FOLD 3 De novo peptide structure prediction. URL: http://bioserv.rpbs.univ-paris-diderot.fr/services/PEP-FOLD3/ [accessed 2022-03-22]
22. Thévenet P, Shen Y, Maupetit J, Guyon F, Derreumaux P, Tufféry P. PEP-FOLD: an updated de novo structure prediction server for both linear and disulfide bonded cyclic peptides. Nucleic Acids Res 2012 Jul 11;40(Web Server issue):W288-W293 [FREE Full text] [doi: 10.1093/nar/gks419] [Medline: 22581768]
23. Shen Y, Maupetit J, Derreumaux P, Tufféry P. Improved PEP-FOLD approach for peptide and miniprotein structure prediction. J Chem Theory Comput 2014 Oct 14;10(10):4745-4758. [doi: 10.1021/ct500592m] [Medline: 26588162]
24. Lamiable A, Thévenet P, Rey J, Vavrusa M, Derreumaux P, Tufféry P. PEP-FOLD3: faster de novo structure prediction for linear peptides in solution and in complex. Nucleic Acids Res 2016 Jul 08;44(W1):W449-W454 [FREE Full text] [doi: 10.1093/nar/gkw329] [Medline: 27131374]
25. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem 2004 Oct;25(13):1605-1612. [doi: 10.1002/jcc.20084] [Medline: 15264254]
26. PatchDock. URL: https://bioinfo3d.cs.tau.ac.il/PatchDock/patchdock.html [accessed 2022-03-22]
27. FireDock. URL: http://bioinfo3d.cs.tau.ac.il/FireDock/firedock.html [accessed 2022-03-22]
28. Zhou P, Jin B, Li H, Huang SY. HPEPDOCK: a web server for blind peptide-protein docking based on a hierarchical algorithm. Nucleic Acids Res 2018 Jul 02;46(W1):W443-W450 [FREE Full text] [doi: 10.1093/nar/gky357] [Medline: 29746661]
29. Wallace AC, Laskowski RA, Thornton JM. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. Protein Eng 1995 Feb;8(2):127-134. [doi: 10.1093/protein/8.2.127] [Medline: 7630882]
30. Ribas‐Aparicio RM, Castelán‐Vega JA, Jiménez‐Alberto A, Monterrubio‐López GP, Aparicio‐Ozores G. The impact of bioinformatics on vaccine design and development. In: Afrin F, Hemeg H, Ozbak H, editors. Vaccines. Rijeka, Croatia: InTech; Sep 06, 2017.
31. Guruprasad K, Reddy B, Pandit MW. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. Protein Eng 1990 Dec;4(2):155-161. [doi: 10.1093/protein/4.2.155] [Medline: 2075190]

32. Ikai A. Thermostability and aliphatic index of globular proteins. J Biochem 1980 Dec;88(6):1895-1898 [FREE Full text] [Medline: 7462208]
33. Pace CN, Vajdos F, Fee L, Grimsley G, Gray T. How to measure and predict the molar absorption coefficient of a protein. Protein Sci 1995 Nov;4(11):2411-2423. [doi: 10.1002/pro.5560041120] [Medline: 8563639]
34. ProtParam. ExPASy. URL: https://web.expasy.org/protparam/ [accessed 2022-03-22]
35. Goerdt S, Orfanos CE. Other functions, other genes. Immunity 1999 Feb;10(2):137-142. [doi: 10.1016/s1074-7613(00)80014-x]
36. Tanchot C, Rocha B. CD8 and B cell memory: same strategy, same signals. Nat Immunol 2003 May;4(5):431-432. [doi: 10.1038/ni0503-431] [Medline: 12719734]
37. Pavli P, Hume DA, Van De Pol E, Doe WF. Dendritic cells, the major antigen-presenting cells of the human colonic lamina propria. Immunology 1993 Jan;78(1):132-141. [Medline: 8436399]
38. Arpin C, Déchanet J, Van Kooten C, Merville P, Grouard G, Brière F, et al. Generation of memory B cells and plasma cells in vitro. Science 1995 May 05;268(5211):720-722. [doi: 10.1126/science.7537388] [Medline: 7537388]
39. Cano R, Lopera H. Introduction to T and B lymphocytes. In: Anaya JM, Shoenfeld Y, Rojas-Villarraga A, Levy RA, Cervera R, editors. From bench to bedside. Bogota, Colombia: Rosario University Press; 2013.
40. Fishman J, Wiles K, Wood K. The acquired immune system response to biomaterials, including both naturally occurring and synthetic biomaterials. In: Badylak SF, editor. Host Response to Biomaterials. Cambridge, MA: Academic Press; 2015:151-187.
41. Andreae D, Nowak-Węgrzyn A. The effect of infant allergen/immunogen exposure on long-term health. In: Saavedra JM, Dattilo AM, editors. Early nutrition and long-term health. Sawston, Cambridge: Woodhead Publishing; 2017:131-173.

Abbreviations

APC: antigen-presenting cell
GRAVY: grand average of hydropathicity
HLA: human leukocyte antigen
IEDB: Immune Epitope Database
MHC: major histocompatibility complex
PDB: Protein Data Bank
pI: isoelectric point
RMSD: root mean square deviation
RMSF: root mean square fluctuation
WHO: World Health Organization

Edited by A Mavragani; submitted 26.07.21; peer-reviewed by K Rathi, P Nandigrami; comments to author 26.10.21; revised version received 02.11.21; accepted 27.12.21; published 26.04.22.

Please cite as:
Gupta E, Mishra RK, Kumar Niraj RR
Identification of Potential Vaccine Candidates Against SARS-CoV-2 to Fight COVID-19: Reverse Vaccinology Approach
JMIR Bioinform Biotech 2022;3(1):e32401
URL: https://bioinform.jmir.org/2022/1/e32401
doi:10.2196/32401
PMID:35506029

Original Paper

In Silico Comparative Analysis of the Functional, Structural, and Evolutionary Properties of SARS-CoV-2 Variant Spike Proteins

Renukaradhya K Math[1], PhD; Nayana Mudennavar[1], MSc; Palaksha Kanive Javaregowda[1], PhD; Ambuja Savanur[1], MSc

SDM Research Institute for Biomedical Sciences, Shri Dharmasthala Manjunatheshwara University, Dharwad, India

Corresponding Author:
Renukaradhya K Math, PhD
SDM Research Institute for Biomedical Sciences
Shri Dharmasthala Manjunatheshwara University
5th floor, Manjushree building
SDM College of Medical Sciences & Hospital Campus
Dharwad, 580009
India
Phone: 91 7019982929
Email: aradhya.swamy@gmail.com

Abstract

Background: A recent global outbreak of COVID-19 caused by the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) created a pandemic and emerged as a potential threat to humanity. The analysis of virus genetic composition has revealed that the spike protein, one of the major structural proteins, facilitates the entry of the virus to host cells.

Objective: The spike protein has become the main target for prophylactics and therapeutics studies. Here, we compared the spike proteins of SARS-CoV-2 variants using bioinformatics tools.

Methods: The spike protein sequences of wild-type SARS-CoV-2 and its 6 variants—D614G, alpha (B.1.1.7), beta (B.1.351), delta (B.1.617.2), gamma (P.1), and omicron (B.1.1.529)—were retrieved from the NCBI database. The ClustalX program was used to sequence multiple alignment and perform mutational analysis. Several online bioinformatics tools were used to predict the physiological, immunological, and structural features of the spike proteins of SARS-CoV-2 variants. A phylogenetic tree was constructed using CLC software. Statistical analysis of the data was done using jamovi 2 software.

Results: Multiple sequence analysis revealed that the P681R mutation in the delta variant, which changed an amino acid from histidine (H) to arginine (R), made the protein more alkaline due to arginine's high pKa value (12.5) compared to histidine's (6.0). Physicochemical properties revealed the relatively higher isoelectric point (7.34) and aliphatic index (84.65) of the delta variant compared to other variants. Statistical analysis of the isoelectric point, antigenicity, and immunogenicity of all the variants revealed significant correlation, with *P* values ranging from <.007 to .04. The generation of a 2D gel map showed the separation of the delta spike protein from a grouping of the other variants. The phylogenetic tree of the spike proteins showed that the delta variant was close to and a mix of the *Rousettus* bat coronavirus and MERS-CoV.

Conclusions: The comparative analysis of SARS-CoV-2 variants revealed that the delta variant is more aliphatic in nature, which provides more stability to it and subsequently influences virus behavior.

(JMIR Bioinform Biotech 2022;3(1):e37391) doi:10.2196/37391

KEYWORDS

spike protein variants; NCBI; bioinformatics tools; pI; isoelectric point; 2D map; phylogenetic tree; COVID-19; COVID therapy; SARS-CoV-2 treatment; therapeutic; spike protein; protein; prophylactic; sequence analysis; genomic; bioinformatics; viral protein

Introduction

The constant evolution of new variants of SARS-CoV-2 is a substantial challenge to people from every sector, particularly those in health care and research and development in the areas of diagnostics, prophylactics, and therapeutics development, as well as policy makers and administrators. The virus was first observed in December 2019 in China and later spread throughout the globe causing a pandemic. However, continuous mutations

in the genome of the virus have created several new variants among the circulating viruses in different geographical regions worldwide. These mutations have led to increased transmissibility, antibody evasion, and severity in patients. Several research studies have cited that the spike protein of the coronavirus is responsible for interacting with human cells to penetrate inside [1,2]; however, adaptive mutations accumulated in the gene responsible for the spike protein have allowed the virus to acclimatize and escape the host immune system. A number of variants have been isolated and identified worldwide; the World Health Organization classified them as variants of concern (alpha, beta, gamma, delta, and omicron) and variants of interest (VOIs; eta, iota, kappa, and lambda) [3].

A single mutation in a gene can lead to a change of an amino acid in a protein, which can drastically affect an organism's ability in many ways, such as transmissibility, quick adaptability, stability evading the host immune system, and pathogenicity [4]. The availability of high throughput second generation sequencing technologies like next-generation sequencing in the last decade has helped in the timely sequencing of the genomes of SARS-CoV-2 variants to identify mutations occurring among new genetic variants. Rapid sequencing technology has not only facilitated sequencing of the viral genome but also helped accelerate the development of diagnostic tools, vaccines, and therapeutics for COVID-19. Furthermore, the availability of genome sequence data and databases helped the scientific community understand the virus through the molecular, biochemical, genome, and proteome analyses of the virus. However, the continuous evolution of SARS-CoV-2 compels the scientific community to continuously strive toward new and thorough understanding of the properties of the virus variants and develop counter strategies against the virus to safeguard humankind.

The isoelectric point (pI) of a protein is crucial in determining the physicochemical properties of the protein [5]. The exposure of charged amino acids on the protein surface to solvents, hydration, and dehydration also influences the pI of a protein [6,7]. Thus, evaluation of the pI of the spike proteins of wild-type and mutated variants is quintessential in understanding the influence of the spike protein on virus behavior and the transmission rate of viruses, as well as developing prophylactic and therapeutic agents. Additionally, the roles of posttranslational modification, phosphorylation, methylation, and alkylation also influence the pI of a protein, which cannot be ignored. Similarly, the genome of SARS-CoV-2 is prone to genetic evolution or genetic shift while adapting to a human host's microenvironment. Such mutations result in the emergence of new variants that might have different characteristics compared to its ancestral strains. Therefore, in this study, we aim to adopt an in silico method to analyze the spike protein sequences of wild-type and other variants of SARS-CoV-2 to know their physicochemical, functional, and evolutionary properties, which might help in the surveillance of SARS-CoV-2 through a systematic understanding of the continuously evolving properties as well as to develop the diagnostic tools and standardization of prophylactics and therapeutics strategies.

Methods

Retrieval of Spike Protein Sequences and Multiple Sequence Alignment

The protein sequences of wild-type SARS-CoV-2 and its 6 variants—D614G, alpha (B.1.1.7), beta (B.1.351), delta (B.1.617.2), gamma (P.1), and omicron (B.1.1.529)—were obtained from the NCBI database. The list of all the variants with the NCBI accession numbers are mentioned in Table 1. Multiple sequence alignment of all the protein was done using ClustalX2 software [8] to identify and confirm common and specific mutations among all the sequences listed in Table 1.

Table 1. List of SARS-CoV-2 spike proteins and the comparison of mutational analysis data of the wild type and its variants.

SARS-CoV-2 variants	First identified	NCBI accession	Linear SeqVrl	Pango lineage	Mutation in spike protein variants sequence	Specific mutation in spike protein variants
Wild type	Wuhan, China	YP_009724390	__[a]	—	—	—
D614G	United States	QTA38988	21-Mar-21	D614G	D614G	D614G
Alpha (B.1.1.7)	United Kingdom	P0DTC2	02-Jun-21	B.1.1.7	E484K, A570D, D614G, P681H	N501Y
Beta (B.1.351)	South Africa	7LYQ_C	09-Jul-21	B.1.351	K417N, E484K, N501Y, D614G	R246I
Delta (B.1.617.2)	India	QWP92316	12-Jun-21	B.1.617.2	E156, K417N, D614G, N501Y	T19R, L452R, T478K, P581R, P681R, D950N
Gamma (P.1)	Brazil	7M8K_C	28-May-21	P.1	K417N, E484K, N501Y, D614G	T20N, P26S, D138Y, R190S, T1027I
Omicron (B.1.1.529)	South Africa	7QO7_C	19-Jan-2022	B.1.1.529	N501Y, Y505H, T547K, D614G, H655Y, etc	A67V, del69-70, T95I, del142-144, Y145D, etc

[a]Not available.

Physicochemical Properties and Posttranslational Modification Prediction

Physicochemical properties such as total amino acids, molecular weight, pI, grand average of hydropathicity (GRAVY), aliphatic index, etc, were predicted using the ProtParam tool [9]. Posttranslational modifications for all spike protein variants were predicted as follows: phosphorylation sites using the NetPhos 2.0 server [10]; glycosylation sites using the NetNGlyc 1.0 server [11]; and disulfide bonds using the Scratch Protein Predictor server [12].

Prediction of Immunoproperties

B-cell epitopes of all the variants were predicated using the ABCpred server [13] and exposed B-cell epitopes were predicted using the BepiPred 2.0 server [14]. T-cell epitopes and their immunogenicity were predicted using the IEBD Analysis Resource server [15], strong binding T-cells were identified using the NetMHCpan 4.1 server [16], and predications of cytotoxic T-lymphocytes were identified by the NetCTL 4.0 server [17].

Secondary and Tertiary Structure Prediction

Predications of secondary structures were identified by the PHYRE2 server [18] and the percentage of the following parameters were assessed: alpha helix, beta stand, transmembrane helix, and disorder. Subsequently, the tertiary structure was predicted using the Swiss model server [19] and the global model quality estimation, confidence, and coverage scores were collected. The structure was also analyzed for Ramachandran plot–allowed regions.

Generation of Phylogenetic Tree and 2D Gel Reference Map

The spike protein sequences of the wild type and variants were collected from NCBI, and the protein sequences of MERS-CoV and *Rousettus* bat coronavirus GCCDC1 were also obtained from NCBI with accession numbers AHY22525 and QKF94914, respectively. The phylogenetic tree was constructed using CLC sequence viewer and DNAMAN software [20] by neighbor joining method, and a virtual 2D protein map of all the spike protein variants was generated using the JVirGel V2.0 software [21].

Ethical Considerations

The study is registered with the institutional ethics committee of the Shri Dharmasthala Manjunatheshwara University (registration no. ECR/950/Inst/KA/2017/RR-21), Sattur, India. According to the institutional ethics committee guidelines, in compliance with the National guidelines/regulation on ethics in Biomedical Research, ethical clearance is not required for studies not involving human subjects/animals/patient medical records/tissues/biologicals fluids/pathogenic microorganism.

Results

In Silico Analysis

The results of an in silico analysis of the wild-type and variant protein sequences of spike protein revealed several common and specific mutations as listed in Table 1. Interestingly, the mutation P681H in the alpha variant [22] changed an amino acid from histidine (H) to arginine (R). Further, the analysis of physicochemical properties revealed that the pI and antigenicity of the delta variant were relatively high compared to other variants, and immunoproperties like B- and T-cell epitope sequences were different in the beta and delta variant compared to others. The phylogenetic tree and 2D gel maps clearly separated the delta variant from others.

Physicochemical Properties and Posttranslational Modification Prediction

The analysis of the physicochemical properties of all the spike proteins was interesting, especially the pI and aliphatic index, which ranged from 6.28 to 7.34 and from 82.04 to 84.65, respectively (Table 2). Among all the proteins, the delta variant had high pI (7.34) and high aliphatic index (84.65) when compared to other proteins (Table 2). However, all the proteins were stable. Meanwhile, the predicted total number of phosphorylation sites in the wild type was 133 whereas this prediction relatively decreased in the delta variant protein (Table 3). In addition, the number of serine, threonine, and tyrosine in phosphorylation sites varied among the variants compared to the wild type.

Predictions of the total number of N-glycosylation and disulfide bonds among the spike protein variants revealed numbers ranging from 16 to 20 and from 14 to 15 sites, respectively (Table 3).

Table 2. Comparison of predicted physicochemical property values of the spike protein of SARS-CoV-2 and its variants.

SARS-CoV-2 variants	Total amino acids	MW[a]	pI[b]	Extinction coefficient (M^{-1} cm^{-1})	EC/A[c]	Half-life (h)	Instability index	Classification of protein	Aliphatic index	GRAVY[d]
Wild type	1273	141178	6.24	148960	1.055	30	33.01	stable	84.67	–0.079
D614G	1252	138712	6.49	147345	1.062	30	32.06	stable	85.01	–0.067
Alpha (B.1.1.7)	1273	141178	6.24	148960	1.055	30	33.01	stable	84.67	–0.079
Beta (B.1.351)	1288	142201	6.38	148335	1.043	30	31.29	stable	82.03	–0.142
Delta (B.1.617.2)	1271	140895	7.34	148960	1.057	30	32.58	stable	84.65	–0.08
Gamma (P.1)	1257	139207	6.18	140315	1.008	30	31.15	stable	83.43	–0.127
Omicron (B.1.1.529)	1285	142424	6.63	146845	1.018	30	33.10	stable	81.84	–0.164

[a]MW: molecular weight.

[b]pI: isoelectric point.

[c]EC/A: extinction coefficient/absorbance for 1% solutions.

[d]GRAVY: grand average of hydropathicity.

Table 3. Comparison of phosphorylation, N-glycosylation, and disulfide bonds values of the spike protein of SARS CoV-2 and its variants.

SARS-CoV-2 variants	Phosphorylation			N-glycosylation		Disulfide bonds	
	Phosphorylation sites	Predicted Ser[a], Thr[b], and Tyr[c] sites	Predicted number	N-glycosylation sites	Predicted number	Total number of cysteine	Predicted number
Wild type	Ser, Thr, Tyr	Ser-64, Thr-44, Tyr-22	133	Asn[d]-Ser/Thr	17	40	15
D614G	Ser, Thr, Tyr	Ser-64, Thr-45, Tyr-22	131	Asn-Ser/Thr	17	38	15
Alpha (B.1.1.7)	Ser, Thr, Tyr	Ser-67, Thr-44, Yyr-22,	133	Asn-Ser/Thr	17	40	15
Beta (B.1.351)	Ser, Thr, Tyr	Ser-71, Thr-43, Tyr-24	136	Asn-Ser/Thr	17	30	14
Delta (B.1.617.2)	Ser, Thr, Tyr	Ser-65, Thr-43, Tyr-22	130	Asn-Ser/Thr	16	40	15
Gamma (P.1)	Ser, Thr, Tyr	Ser-66, Thr-43, Tyr-24	133	Asn-Ser/Thr	20	30	14
Omicron (B.1.1.529)	Ser, Thr, Tyr	Ser-68, Thr-43, Tyr-21	132	Asn-Ser/Thr	18	30	14

[a]Ser: serine.

[b]Thr: threonine.

[c]Tyr: tyrosine.

[d]Asn: asparagine.

Prediction of Immunoproperties

Predictions of the number of exposed B-cell epitopes for spike protein varied from 38 to 41 (Table 4). The B-cell epitope sequence for the wild-type, alpha, and gamma variants was QTQTNSPRRARSV, whereas this sequence changed for the D614G, beta, and delta variants. Among all the variants, the delta variant showed the highest score for protective antigen (0.4709) and antigenicity (0.7440; Table 4). Predictions for C-cell epitopes revealed that the number of epitopes ranged from 35 to 38 and the number of strong binders in T-cell were from 21 to 23. The predicted T-cell epitopes for the wild-type and gamma spike variants were identical to IGINITRFQTLLALH but changed in other spike protein variants. However, except in the alpha variant, the sequence QTLLALH was supposed to be conserved (Table 4). The immunogenicity prediction scores for the spike protein variants varied. Meanwhile, statistical analysis of pI, antigenicity, and immunogenicity scores revealed the significance of the data, especially between alpha and wild type (P=.007), delta and wild type (P=.02), and delta and alpha (P=.02). The correlation matrix was positive with P values ranging from <.007 to .04 (Table 5).

Table 4. Comparison of the immunological properties of the spike protein of SARS-CoV-2 and its variants.

SARS-CoV-2 variants	Exposed B-cell epitopes	B-cell epitopes	Protective antigen prediction score	Predicted probability of antigenicity score	Number of epitopes identified in CTL[a]	Number of strong binders in T-cell	Predicted epitopes in T-cell	Immunogenicity predication score
Wild type	40	QTQTNSPRRARSV	0.4646	0.717053	37	21	IGINITRFQ**TLLA**LH	0.3751
D614G	40	YHKNNKS	0.4583	0.741478	35	22	ITRFQ**TLLA**	0.96257
Alpha (B.1.1.7)	40	QTQTNSPRRARSV	0.4646	0.717053	37	21	NGTHWFVTQRN-FYEP	0.3019
Beta (B.1.351)	40	HPQFEKGGGSGGGGSG	0.4542	0.643558	38	23	QPYRVVVLSFE**LL**HA	1.23216
Delta (B.1.617.2)	38	SLGAENSVAYSN	0.4709	0.744007	35	22	IRAAEIRASAN**LAA**T	0.0304
Gamma (P.1)	41	QTQTNSPRRARSV	0.4583	0.596261	36	22	IGINITRFQ**TLLAL**H	1.07515
Omicron (B.1.1.529)	33	QTQTKSHGSASSVA	0.4646	0.717053	31	20	IGINITRFQTLLALH	0.49637

[a]CTL: cytotoxic T-lymphocyte.

Table 5. The *P* values of isoelectric point, antigenicity, and immunogenicity of the spike protein of SARS-CoV-2 and its variants.

SARS-CoV-2 variants	Wild type	D614G	Alpha	Beta	Delta	Gamma	Omicron
Wild type							
r	—[a]	0.996	1.000	0.989	0.999	0.992	1.000
P value	—	.06	.007	.09	.02	.08	.01
D614G							
r	0.996	—	0.995	0.998	0.992	0.999	0.998
P value	.06	—	.06	.04	.08	.03	.04
Alpha							
r	1.000	0.995	—	0.988	1.000	0.990	1.000
P value	.007	.06	—	.10	.02	.09	.02
Beta							
r	0.989	0.998	0.988	—	0.983	1.000	0.992
P value	.09	.04	.10	—	.12	.01	.08
Delta							
r	0.999	0.992	1.000	0.983	—	0.986	0.998
P value	.02	.08	.02	.12	—	.11	.04
Gamma							
r	0.992	0.999	0.990	1.000	0.986	—	0.994
P value	.08	.03	.09	.01	.11	—	.07
Omicron							
r	1.000	0.998	1.000	0.992	0.998	0.994	—
P value	.01	.04	.02	.08	.04	.07	—

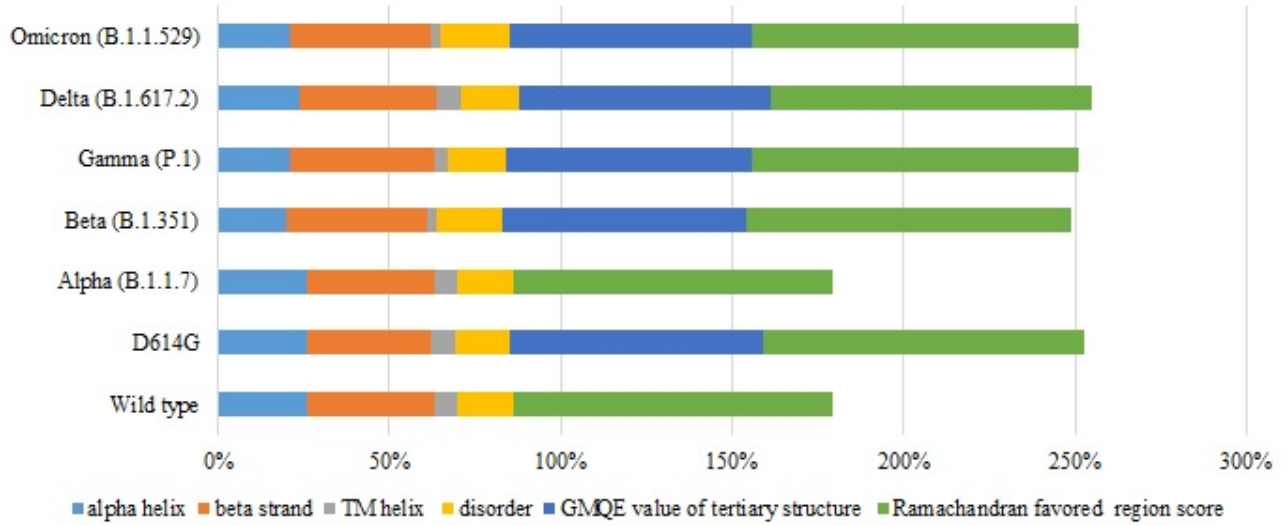
[a]Not applicable.

Secondary and Tertiary Structure Prediction

Secondary and tertiary prediction of all the spike protein variants showed that all protein structures were stable (Figure 1). The predicted proportion of alpha helices, beta strands, transmembrane helices, and disorders of all the variants are shown in Figure 1. The alpha helix percentages varied from 26% in the wild type to 21% in the gamma variant, whereas the beta strand percentages varied from 37% in the wild type to 42% in the gamma variant; however, analysis of alpha and beta percentages revealed that the proteins were stable. The global

model quality estimate scores and Ramachandran favored regions—of which all the spike protein variants' percentages were similar to the wild type's—indicated that there was no significant change in the stability of the variants compared to the wild type (Figure 1).

Figure 1. Graphical illustration of the predicted percentages of the secondary and tertiary structure and disorder of the spike proteins of SARS-CoV-2 wild type and its variants. GMQE: global model quality estimate; TM: transmembrane.



Generation of 2D Gel Reference Map and Phylogenetic Tree

The 2D reference map of the spike protein and its variants revealed a grouping of the wild-type, D614G, alpha, beta, gamma, and omicron variants whereas the delta variant was clearly separated (Figure 2). The phylogenetic tree of the spike protein of the wild-type and variant proteins along with MERS-CoV and bat coronavirus was constructed as seen in Figure 3. Construction of the phylogenetic tree grouped the wild-type and alpha variants as one cluster and the beta, gamma, and omicron variants as another cluster. The D614G variant stood in between these 2 groups. Interestingly, the delta variant was closely grouped with MERS-CoV and *Rousettus* bat coronavirus. Meanwhile, the omicron variant showed 99% bootstrap values with the D614G variant, showing the closest relationship among all others. However, the tree has 2 branches: one with the wild type and variants (except delta) and another with the delta variant along with MERS-CoV and *Rousettus* bat coronavirus (Figure 3).

Figure 2. The 2D gel map of SARS-CoV-2 and its variants.

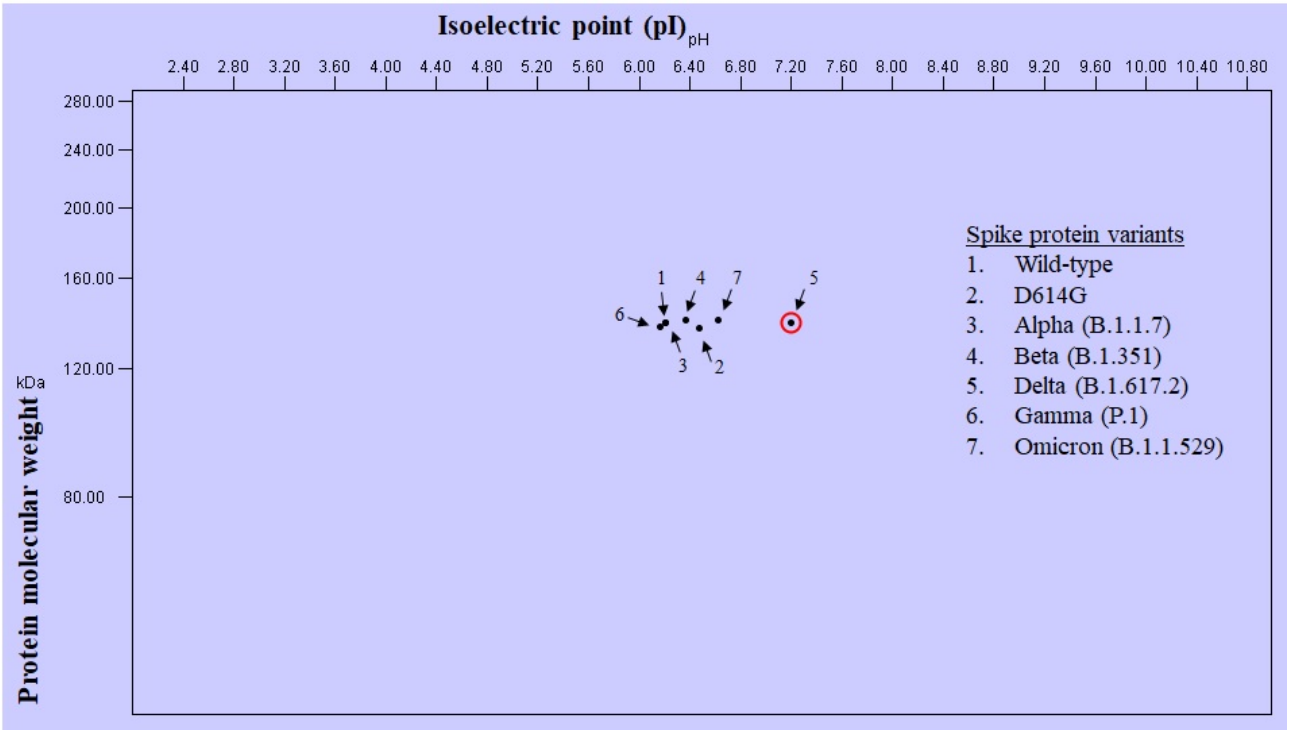
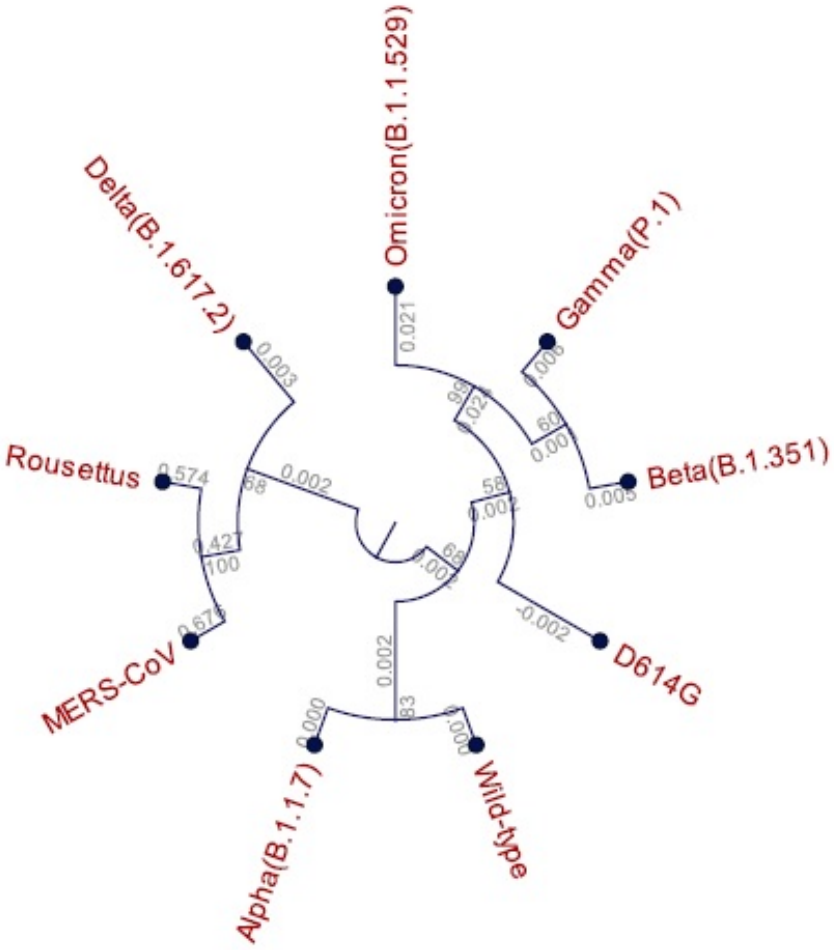


Figure 3. The phylogenetic tree of SARS-CoV-2 and its variants represented in circular cladogram. Protein sequences of MERS-CoV and *Rousettus* bat coronavirus were used as nearest relatives.



Discussion

The current COVID-19 pandemic caused by SARS-CoV-2 has clearly demonstrated the potential of the coronavirus to continuously evolve in the wild and transmit to new species including humans with varied physicochemical and virulence properties. The world has witnessed several mutants of this virus, with “VOIs” and “variants of concern” taking several millions lives. However, the situation seems to be under control because of the availability of vaccines recently from the beginning of the year 2021. However, it is quintessential to understand the constantly mutating property of viruses to establish effective and timely health care strategies to avoid the consequential economic loss or burden and safeguard humankind. Understanding life-threatening organism such as SARS-CoV-2 is a continuous and challenging process for the health care and scientific community as well as for policy makers and administrators. There are several methods to study and understand the disease-causing agents, of which the in silico analysis method has contributed enormously to make the scientific community’s tasks easier by reducing the required effort, time, and costs in recent years. Therefore, in this study, we have made an effort to understand the variants of concerns, especially how their spike proteins spread around the globe. The analysis of the physicochemical properties of spike proteins of the wild-type and mutated variants of SARS-CoV-2 revealed interesting features of the delta variant, such as pI, molecular weight, instability index, GRAVY, and aliphatic index (Table 2).

The pI of a protein is crucial to understanding its biochemical function [23] and dependent on the dissociation and constant of the ionizable amino acid groups. The major ionizable amino acid groups present in SARS-CoV-2 are arginine, aspartate, histidine, glutamate, cysteine, and lysine, which play an important role in defining the pI of a spike protein [24-26]. However, posttranslational modification influences the pI of a protein. The mutation of P681R in the delta variant [22] changed an amino acid from histidine to arginine, which sparked our curiosity to try to understand the potential impact of the change in amino acid on the properties of the spike protein in the alpha and delta variants. Interestingly, the pKa values of both amino acids greatly differed—histidine has a value of 6.0 (low basic) and arginine has a value 12.5 (high basic). Other specific mutations observed in the delta variant were also basic or aliphatic in nature (Table 1). Thus, if this property is substantial, then that would make the protein more basic. The mutation P681H in the alpha variant did not cause substantial changes in the behavior of virus, but the mutation P681R in the delta variant did cause substantial changes in the behavior of the virus in terms of transmissibility, pathogenicity, immune evasion, and the severity of the infection [22]. Conversely, a neutral mutation

was observed in the D614G variant spike protein [27,28], where an acidic amino acid (D) was replaced with neutral one (G).

To see the effect of the pI on the biochemical properties, the wild-type and variant spike proteins were evaluated by predicting the 2D gel reference map, which interestingly demonstrated grouping except the separation of the delta protein on the map (Figure 2). The results of the physicochemical properties and 2D gel analysis revealed that the pI of a protein plays an important role in the behavior of proteins, especially where high pI turns protein more positive or aliphatic. Additionally, it is interesting to know that this property of protein is used in antibody preparation—to make a protein or its subunit more immunogenic [29,30]. Hence, we suppose that this property of proteins should be taken into consideration when designing and preparing prophylactic and therapeutic agents. Similarly, a recent study on the VOIs of SARS-CoV-2 virus particles that measured pI using chemical force microscope revealed VOIs have lower surface charge and hydrophobicity than the wild type, which might have played a role in VOIs having increased transmission ability [29]. Therefore, the pI of protein and the surface charge and hydrophobicity of viral particles are important factors to consider when designing prophylactics and therapeutics for any viral diseases [4,28,30,31].

Substantial changes in the physicochemical properties of the spike protein of wild-type SARS-CoV-2 and its variants has not been observed; however, a high number of phosphorylation sites (136) was observed in the beta variant and a low number (130) was observed in the delta variant. Meanwhile, a high number of N-glycosylation sites (20) was observed in the gamma variant and a low number (16) was observed in the delta variant. The total number of disulfide bonds among the wild type and variants did not vary much; however, it ranged from 14 to 15 and the number bonds were enough to give structural stability. Overall, the physicochemical properties did not differ markedly; however, the differences found were enough to influence the change in protein behavior and, subsequently, the behavior of the virus.

The antigenicity and immunogenicity scores of wild-type SARS-CoV-2 and its variants were interesting, especially the delta variant which showed high and low scores, respectively. Comparison of pI, antigenicity, and immunogenicity was made to evaluate the significance of the data; the *P* values were .007 between the alpha variant and wild type, .02 between the delta variant and wild type, and .02 between the delta and alpha variants. Overall, these *P* values suggest that the predicted values and the antigenic and immunogenic sequences are reliable. The number of exposed B-cell and cytotoxic T-lymphocyte epitopes were low in the delta variant compared to the wild-type and other variants of SARS-CoV-2. The exposed B-cell epitope sequence for the delta variant was SLGAENSVAYSN, and the change in epitope sequence showed that mutations did have an effect on the morphology of the virus.

The evolutionary relationship among the wild type and its variants was evaluated by constructing a phylogenetic tree using protein sequences. Interestingly, we found a clear grouping of the D614G, beta, and gamma variants, with bootstrap values of 100% for beta and gamma and 93% between D614G and beta/gamma. The delta variant was close to *Rousettus* bat coronavirus and MERS-CoV (Figure 3) with a bootstrap value of 68%. The alpha variant was close to the wild type with an 83% bootstrap value. Interestingly, we could observe a hybrid position of the delta variant between the wild type and the MERS-CoV and *Rousettus* bat coronavirus. Therefore, we suppose that the delta variant might be carrying characteristics from the wild type and other wild bat coronaviruses (Figure 3). In another interesting observation, the omicron variant seems to have evolved from the D614G variant with a bootstrap value 99%, as well as having imported some characteristics from the beta and gamma variants since omicron was also seen branching from these 2 variants with a 60% bootstrap value (Figure 3).

In conclusion, our study highlights that the accumulation of adaptive mutations in SARS-CoV-2 influenced the change in pI of the spike protein and, subsequently, the behavior of the virus. The prediction and comparison of the physicochemical properties of spike proteins of the wild type and its variants revealed that the delta variant displayed unique changes compared to the wild type. Evolutionary features showed a clear separation of the wild-type, alpha, and delta spike proteins and the grouping of the D614G, beta, and gamma spike proteins. Nevertheless, the continuous evolution of new SARS-CoV-2 strains demands further systematic understanding of its variants, which would not only help in developing the improvised rapid “antigen test” market but also in developing vaccines and therapeutics. Thus, more similar studies would unravel the important biochemical properties, evolutionary history, immunological behavior, and physiological properties of viruses.

Conflicts of Interest

None declared.

References

1. Brielle ES, Schneidman-Duhovny D, Linial M. The SARS-CoV-2 exerts a distinctive strategy for interacting with the ACE2 human receptor. Viruses 2020 Apr 30;12(5):497 [FREE Full text] [doi: 10.3390/v12050497] [Medline: 32365751]
2. Walls AC, Park Y, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell 2020 Apr 16;181(2):281-292.e6 [FREE Full text] [doi: 10.1016/j.cell.2020.02.058] [Medline: 32155444]
3. Tracking SARS-CoV-2 variants. World Health Organization. 2022 May 18. URL: https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/ [accessed 2022-05-19]

4. Mastriani E, Rakov AV, Liu SL. Isolating SARS-CoV-2 strains from countries in the same meridian: genome evolutionary analysis. JMIR Bioinform Biotech 2021 Jan 22;2(1):e25995 [FREE Full text] [doi: 10.2196/25995] [Medline: 33497425]
5. Ebrahim-Saraie HS, Dehghani B, Mojtahedi A, Shenagari M, Hasannejad-Bibalan M. Functional and structural characterization of SARS-Cov-2 spike protein: an in silico study. Ethiop J Health Sci 2021 Mar 01;31(2):213-222 [FREE Full text] [doi: 10.4314/ejhs.v31i2.2] [Medline: 34158771]
6. Youden WJ, Denny FE. Factors influencing the pH equilibrium known as the isoelectric point of plant tissue. Am J Bot 1926 Dec 01;13(10):743-753. [doi: 10.1002/j.1537-2197.1926.tb05907.x]
7. Shaw KL, Grimsley GR, Yakovlev GI, Makarov AA, Pace CN. The effect of net charge on the solubility, activity, and stability of ribonuclease Sa. Protein Sci 2001 Jun;10(6):1206-1215 [FREE Full text] [doi: 10.1110/ps.440101] [Medline: 11369859]
8. Higgins DG. CLUSTAL V: multiple alignment of DNA and protein sequences. Methods Mol Biol 1994;25:307-318. [doi: 10.1385/0-89603-276-0:307] [Medline: 8004173]
9. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, et al. Protein Identification and Analysis Tools on the ExPASy Server. In: Walker JM, editor. The Proteomics Protocols Handbook. Totowa, NJ: Springer Humana Press; 2005:571-607.
10. Blom N, Gammeltoft S, Brunak S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol 1999 Dec 17;294(5):1351-1362. [doi: 10.1006/jmbi.1999.3310] [Medline: 10600390]
11. Gupta R, Brunak S. Prediction of glycosylation across the human proteome and the correlation to protein function. Singapore: World Scientific; 2001 Dec Presented at: Pacific Symposium on Biocomputing 2002; January 3-7, 2002; Kauai, HI p. 310-322. [doi: 10.1142/9789812799623_0029]
12. Cheng J, Randall AZ, Sweredoski MJ, Baldi P. SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res 2005 Jul 01;33(Web Server issue):W72-W76 [FREE Full text] [doi: 10.1093/nar/gki396] [Medline: 15980571]
13. Saha S, Raghava GPS. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. Proteins 2006 Oct 01;65(1):40-48. [doi: 10.1002/prot.21078] [Medline: 16894596]
14. Jespersen MC, Peters B, Nielsen M, Marcatili P. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. Nucleic Acids Res 2017 Jul 03;45(W1):W24-W29 [FREE Full text] [doi: 10.1093/nar/gkx346] [Medline: 28472356]
15. Nielsen M, Lundegaard C, Lund O, Keşmir C. The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. Immunogenetics 2005 Apr;57(1-2):33-41. [doi: 10.1007/s00251-005-0781-7] [Medline: 15744535]
16. Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucleic Acids Res 2020 Jul 02;48(W1):W449-W454 [FREE Full text] [doi: 10.1093/nar/gkaa379] [Medline: 32406916]
17. Larsen MV, Lundegaard C, Lamberth K, Buus S, Lund O, Nielsen M. Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. BMC Bioinformatics 2007 Oct 31;8(1):424 [FREE Full text] [doi: 10.1186/1471-2105-8-424] [Medline: 17973982]
18. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 2015 Jun;10(6):845-858 [FREE Full text] [doi: 10.1038/nprot.2015.053] [Medline: 25950237]
19. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res 2018 Jul 02;46(W1):W296-W303 [FREE Full text] [doi: 10.1093/nar/gky427] [Medline: 29788355]
20. Wang W. Erratum to: the molecular detection of Corynespora Cassiicola on cucumber by PCR assay using DNAman software and NCBI. In: Li D, Li Z, editors. Computer and Computing Technologies in Agriculture IX. CCTA 2015. IFIP Advances in Information and Communication Technology, vol 479. Cham, Switzerland: Springer; 2016.
21. Hiller K, Schobert M, Hundertmark C, Jahn D, Münch R. JVirGel: calculation of virtual two-dimensional protein gels. Nucleic Acids Res 2003 Jul 01;31(13):3862-3865 [FREE Full text] [doi: 10.1093/nar/gkg536] [Medline: 12824438]
22. Liu Y, Liu J, Johnson BA, Xia H, Ku Z, Schindewolf C, et al. Delta spike P681R mutation enhances SARS-CoV-2 fitness over Alpha variant. bioRxiv Preprint posted online on September 05,2021. [FREE Full text] [doi: 10.1101/2021.08.12.456173] [Medline: 34462752]
23. Math RK, Kambiranda D, Yun HD, Ghebreiyessus Y. Binding of cloned Cel enzymes on clay minerals related to the pI of the enzymes and database survey of cellulases of soil bacteria for pI. Biosci Biotechnol Biochem 2020 Feb;84(2):238-246. [doi: 10.1080/09168451.2019.1679613] [Medline: 31625450]
24. Grimsley GR, Scholtz JM, Pace CN. A summary of the measured pK values of the ionizable groups in folded proteins. Protein Sci 2009 Jan;18(1):247-251 [FREE Full text] [doi: 10.1002/pro.19] [Medline: 19177368]
25. Pace CN, Grimsley GR, Scholtz JM. Protein ionizable groups: pK values and their contribution to protein stability and solubility. J Biol Chem 2009 May 15;284(20):13285-13289 [FREE Full text] [doi: 10.1074/jbc.R800080200] [Medline: 19164280]

26. Masunov A, Lazaridis T. Potentials of mean force between ionizable amino acid side chains in water. J Am Chem Soc 2003 Feb 19;125(7):1722-1730. [doi: 10.1021/ja025521w] [Medline: 12580597]
27. Isabel S, Graña-Miraglia L, Gutierrez JM, Bundalovic-Torma C, Groves HE, Isabel MR, et al. Evolutionary and structural analyses of SARS-CoV-2 D614G spike protein mutation now documented worldwide. Sci Rep 2020 Aug 20;10(1):14031 [FREE Full text] [doi: 10.1038/s41598-020-70827-z] [Medline: 32820179]
28. Math RK, Goni M, Javaregowda PK, Khandagale AS, Oli A. SARS-CoV19-2 spike protein mutation patterns: a global scenario. Preprints Preprint posted on July 7, 2020. [doi: 10.20944/preprints202007.0132.v1]
29. Areo O, Joshi PU, Obrenovich M, Tayahi M, Heldt CL. Single-particle characterization of SARS-CoV-2 isoelectric point and comparison to variants of interest. Microorganisms 2021 Jul 28;9(8):1606 [FREE Full text] [doi: 10.3390/microorganisms9081606] [Medline: 34442686]
30. Apple RJ, Domen PL, Muckerheide A, Michael JG. Cationization of protein antigens. IV. Increased antigen uptake by antigen-presenting cells. J Immunol 1988 May 15;140(10):3290-3295. [Medline: 3258879]
31. Domen PL, Muckerheide A, Michael JG. Cationization of protein antigens. III. abrogation of oral tolerance. J Immunol 1987 Nov 15;139(10):3195-3198. [Medline: 3680943]

Abbreviations

GRAVY: grand average of hydropathicity
pI: isoelectric point
VOI: variant of interest

Edited by A Mavragani; submitted 18.02.22; peer-reviewed by BS Chrisman, M Giri; comments to author 10.04.22; revised version received 25.04.22; accepted 16.05.22; published 30.05.22.

Please cite as:
Math RK, Mudennavar N, Javaregowda PK, Savanur A
In Silico Comparative Analysis of the Functional, Structural, and Evolutionary Properties of SARS-CoV-2 Variant Spike Proteins
JMIR Bioinform Biotech 2022;3(1):e37391
URL: https://bioinform.jmir.org/2022/1/e37391
doi:10.2196/37391
PMID:35669291

Original Paper

Development of a Multiepitope Vaccine Against SARS-CoV-2: Immunoinformatics Study

Fatemeh Ghafouri[1], BSc; Reza Ahangari Cohan[2], PhD; Hilda Samimi[3], PhD; Ali Hosseini Rad S M[4], PhD; Mahmood Naderi[5], PhD; Farshid Noorbakhsh[6], MD, PhD; Vahid Haghpanah[3,7], MD, MPH, PhD

[1]Department of Biotechnology, Faculty of Life Sciences and Biotechnology, Shahid Beheshti University, Tehran, Iran

[2]Department of Nanobiotechnology, New Technologies Research Group, Pasteur Institute of Iran, Tehran, Iran

[3]Endocrinology and Metabolism Research Center, Endocrinology and Metabolism Clinical Sciences Institute, Tehran University of Medical Sciences, Tehran, Iran

[4]Department of Microbiology and Immunology, University of Otago, Otago, New Zealand

[5]Digestive Diseases Research Center, Digestive Diseases Research Institute, Tehran University of Medical Sciences, Tehran, Iran

[6]Department of Immunology, School of Medicine, Tehran University of Medical Sciences, Tehran, Iran

[7]Personalized Medicine Research Center, Endocrinology and Metabolism Clinical Sciences Institute, Tehran University of Medical Sciences, Tehran, Iran

Corresponding Author:
Vahid Haghpanah, MD, MPH, PhD
Endocrinology and Metabolism Research Center
Endocrinology and Metabolism Clinical Sciences Institute
Tehran University of Medical Sciences
Dr Shariati Hospital
North Kargar Avenue
Tehran, 14114
Iran
Phone: 98 21 88220037
Email: v.haghpanah@gmail.com

Abstract

Background: Since the first appearance of SARS-CoV-2 in China in December 2019, the world witnessed the emergence of the SARS-CoV-2 outbreak. Due to the high transmissibility rate of the virus, there is an urgent need to design and develop vaccines against SARS-CoV-2 to prevent more cases affected by the virus.

Objective: A computational approach is proposed for vaccine design against the SARS-CoV-2 spike (S) protein, as the key target for neutralizing antibodies, and envelope (E) protein, which contains a conserved sequence feature.

Methods: We used previously reported epitopes of S protein detected experimentally and further identified a collection of predicted B-cell and major histocompatibility (MHC) class II–restricted T-cell epitopes derived from E proteins with an identical match to SARS-CoV-2 E protein.

Results: The in silico design of our candidate vaccine against the S and E proteins of SARS-CoV-2 demonstrated a high affinity to MHC class II molecules and effective results in immune response simulations.

Conclusions: Based on the results of this study, the multiepitope vaccine designed against the S and E proteins of SARS-CoV-2 may be considered as a new, safe, and efficient approach to combatting the COVID-19 pandemic.

(JMIR Bioinform Biotech 2022;3(1):e36100) doi:10.2196/36100

KEYWORDS

SARS-CoV-2; envelope protein; spike protein; COVID-19 vaccine; bioinformatics; COVID-19; informatics; immunoinformatics; computational model; vaccine design; pandemic

Introduction

The recent outbreak of the new virus in Wuhan City, China, contributed to the discovery of a new coronavirus strain, labeled SARS-CoV-2, of the Coronaviridae family. This virus has caused severe damage and anxiety, leading to the loss of myriad individuals, impacting more than 535,863,950 people to date. SARS-CoV-2 causes the disease named COVID-19, which is associated with symptoms such as a flu-like illness, acute respiratory distress syndrome, and clinical or radiological evidence of pneumonia in individuals needing hospitalization [1]. Patients diagnosed with COVID-19 are reported to have high levels of interleukin (IL)1β, interferon (IFN)γ, interferon-inducible protein 10 (IP10), and monocyte chemoattractant protein 1 (MCP1), likely leading to activated T helper-1 cell responses. In comparison, patients requiring intensive care unit admission had higher concentrations of granulocyte-colony stimulating factor, IP10, MCP1, MIP1A, and tumor necrosis factor-α than those not requiring intensive care, suggesting a possible correlation of cytokine storm and disease intensity. Nonetheless, SARS-CoV-2 infection also resulted in the enhanced production of T helper-2 cell cytokines such as IL4 and IL10, which inhibit inflammation that varies from that induced by SARS-CoV infection [2]. The persistent rise in patients and the high contagious rate of SARS-CoV-2 infection illustrate the immediate need to develop a safe and effective vaccine.

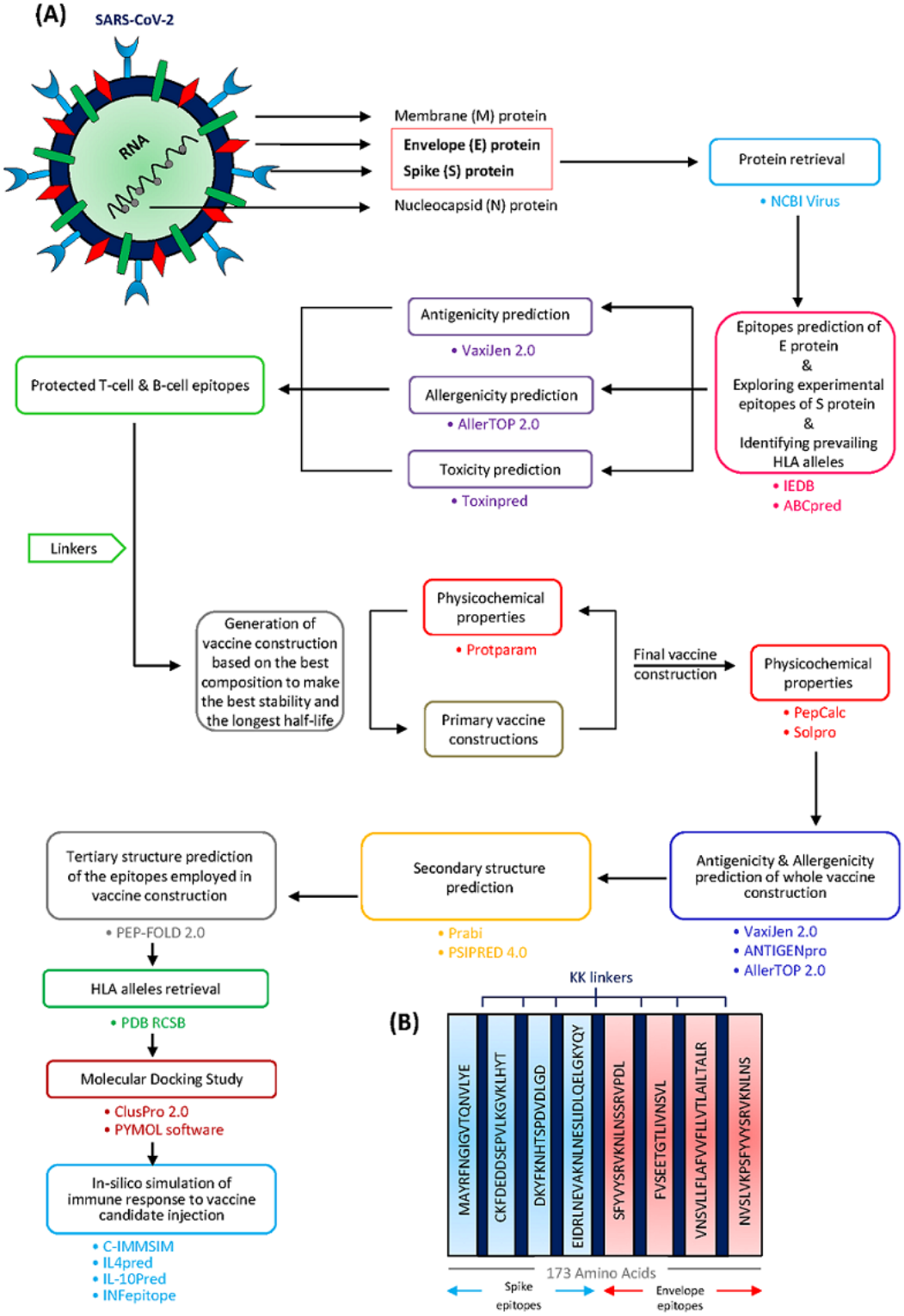
Vaccines are mostly comprised of whole pathogens, either destroyed or attenuated. However, it may be beneficial to use protein vaccines that are capable of generating an immune response against a specific pathogen. Epitope-based vaccines (EVs) utilize immunogenic proteins (epitopes) to induce an immune response. The performance of an EV is calculated by the number of epitopes to be used as the foundation. Nevertheless, the experimental identification of candidate epitopes is costly in terms of both time and money. Moreover, different immunological requirements need to be considered for the final choice of epitopes [3].

The properties of coronaviruses can be determined by electron microscopy. Coronaviruses are enveloped viruses with single-stranded positive-sense RNA. The coronavirus genome size varies from 26 to 32 kb [4]. Like all coronaviruses, SARS-CoV-2 comprises four viral proteins, namely spike (S) protein, a type of glycoprotein; membrane (M) protein, covering the membrane; envelope (E) protein, a strongly hydrophobic protein that covers the entire coronavirus structure; and nucleocapsid (N) protein, a structural protein that suppresses RNA interference to overcome the host defense response [5,6] (Figure 1A). Such accessory proteins are not only essential for virion assembly but might also play additional roles in disrupting the host immune responses to promote viral replication [7]. SARS-CoV-2 requires the S glycoprotein, as the key target for neutralizing antibodies, to bind to the receptor and facilitate membrane fusion and virus entry. Every trimeric S protein monomer is roughly 180 kDa in size and comprises two subunits, S1 and S2, mediating binding and membrane fusion, respectively [8]. Therefore, S protein, but not other structural proteins, is the main antigen that causes the production of defensive neutralizing antibodies that stop viruses from attaching to their specific receptor, thereby preventing viral infection [9,10]. The S and M structural proteins have also been shown to have substantial mutational modifications, whereas the E and N proteins are highly conserved (Figure 1A), indicating differential selection pressures imposed on SARS-CoV-2 during evolution [11]. E protein is a small intrinsic membrane protein that is actively engaged in several stages of the life cycle of the virus, such as assembling, propagation, enveloping, and pathogenesis [12]. This protein also slows the transport of proteins through the secretive pathway by adjusting the concentrations of Ca^{2+} and H^{+} in the Golgi and endoplasmic reticulum compartments, which has been suggested as a mechanism for immune avoidance [13].

In this study, the S and E protein sequences were collected from a protein database and analyzed with various bioinformatics tools to identify protective epitopes. The toxicity of whole E protein as a second antigen was analyzed, and toxic epitopes were identified. The predicted B-cell and major histocompatibility complex (MHC) class II–restricted T-cell epitopes were checked in terms of not coinciding with these regions. The presence of less toxic epitopes in E protein in comparison with S protein served as a motivation to design an effective vaccine against these two antigens.

In particular, we sought to design a vaccine against the two structural antigens, S and E proteins, without using built-in adjuvants to obtain a vaccine that could be effective against all current and potential mutations of SARS-CoV-2, along with the advantage of a low molecular weight to avoid the complexity of future manufacturing. Since E protein is more highly conserved and the candidate vaccine showed all of the desired properties in simulations without using adjuvants, the study goals were achieved.

Figure 1. Schematic of the overall study design for development of a SARS-CoV-2 multiepitope vaccine. (A) Study workflow of the in silico design of a multiepitope vaccine against the envelope (E) protein of SARS-CoV-2. (B) Overlaps of 21 selected epitopes merged showing the final construct consisting of 8 epitopes. HLA: human leukocyte antigen; NCBI: National Center for Biotechnology Information.



Methods

Protein Sequence Retrieval

Based on the vaxquery database [14], the S and E proteins of SARS-CoV-2 were selected as targets for vaccine design because there are already vaccines in development or produced based on these proteins. The amino acid sequences of the E and S proteins of SARS-CoV-2 were collected from the National Center for Biotechnology Information (NCBI) virus database [15] with accession number QHD43418 and QHR63280, respectively.

B-Cell Epitopes Prediction

Prediction methods are both time- and cost-effective, and are reliable approaches for predicting linear B-cell epitopes as the

first step in the genome-wide quest for identifying B-cell antigens in a pathogenic organism [16]. The ABCpred database [17] includes the full-length E protein sequence for the prediction of linear B-cell epitopes. In this study, we set the "threshold" to 0.51 (default value), "length" to 16 (default value), and "overlapping filter" to "NO." ABCpred uses a machine-learning methodology that requires fixed-length patterns for training or research, and the B-cell epitopes range from 5 to 30 residues in length. To overcome this issue, the server sought to create data sets of fixed-length patterns from B-cell epitopes by removing or linking residues to terminals. With a single hidden layer, the ABCpred server employs a partly recurrent neural network (Jordan network). The networks contain a single hidden layer with 35 residues and selectable window lengths of 10, 12, 14, 16, 18, and 20. The result is a single binary value that is either 1 or 0 (epitope or nonepitope).

The performance of prediction algorithms was evaluated using three parameters [17,18]: sensitivity, specificity, and accuracy. Sensitivity was calculated as the percentage of epitopes correctly identified as epitopes with the formula (TP/[TP+FN])×100, where TP and FN are the numbers of true positives and false negatives, respectively. Specificity was calculated as the percentage of correctly predicted nonepitopes with the formula (TN/[TN+FP])×100, where TN and FP are the numbers of true negatives and false positives, respectively. Accuracy is calculated as the total number of correct predictions (which includes both TP and TN) divided by the total number of forecasts made, multiplied by 100.

The Immune Epitope Database (IEDB) [19] was utilized to browse the available experimental B-cell assays on S protein of SARS-CoV-2. IEDB documents experimental evidence on antibody and T-cell epitopes examined in humans, nonhuman primates, and other animal species in the sense of infectious diseases, allergy, autoimmunity, and transplantation. We used the following parameters for browsing S epitopes: "Epitope" was set to "Linear peptide," "Organism" was set to "SARS-CoV," "Antigen" was set to "Spike glycoprotein," "Assay" was set to "T cell" and "B Cell," "MHC restriction" was set to "Class II," and "Host" was set to "Human."

For the B-cell epitope prediction of E protein using IEDB, we used the Bepipred Linear Epitope Prediction 2.0 service, which is based on a random forest algorithm trained on epitopes and nonepitope amino acids obtained from reported crystal structures to predict B-cell epitopes from a protein sequence, followed by sequential prediction smoothing [20].

T-Cell Epitopes Prediction

T-cell epitopes are a group of proteins that can be detected by T-cell receptors after a given antigen has been processed intracellularly and attached to at least one MHC molecule, which are then expressed on the surface of antigen-presenting cells (APCs) as an MHC-protein complex. For entities that have at least one MHC molecule with strong affinity for binding to allergenic amino acid sequences from an allergen, the T-cell clones that can detect this MHC-protein complex are genetically susceptible to allergic reactions to this allergen. This concept can be investigated in silico by employing advanced statistical and mathematical methods [21]. The helper T lymphocyte (HTL) epitopes of E protein were predicted by the IEDB database [19]. For T-cell MHC class II epitope prediction by IEDB, the "Prediction method" was set to "IEDB recommended 2.22," "species/locus" was set to "human"/"HLA-DRA-DBR1*01:01" and "HLA-DPB1*01:02," and "length" was left as the default setting. IEDB recommends using the consensus method, which compares a variety of methods to predict MHC class II epitopes, including a consensus approach combining NN-align, SMM-align, and combinatorial library methods [19]. The other tool that is available on the IEDB can browse the experimental HTL epitopes through the library based on a relevant antigen [19].

Antigenicity, Allergenicity, and Toxicity Prediction

VaxiJen v2.0 with a threshold of 0.4 was used to predict the antigenicity of both B-cell and T-cell epitopes. VaxiJen is the first alignment-independent antigen predictor server, which was developed to achieve the categorization of antigens solely based on the physicochemical properties of proteins without recourse to sequence alignment. The system can be used either on its own or in conjunction with alignment-based prediction methods [22]. The methodology of this server is based on z descriptors, autocross covariance (ACC) preprocessing, discriminant analysis by partial least squares, and sequence similarity of the training set [23]. The z descriptors reflect the most critical physicochemical features for antigen recognition, including z1, z2, and z3 descriptors to describe the protein sequences. The hydrophobicity of amino acids is represented by the first principal component (z1), their size is represented by the second component (z2), and their polarity is represented by the third component (z3). The auto covariance Ajj(lag) is represented by Equation (1) [22]:

(1)

The z-scales are calculated using index j (j=1, 2, 3), n (number of amino acids in a sequence), I (amino acid position; I=1, 2,...n), and l (lag) (l=1, 2,...L). A small range of lags (L=1,2,3,4,5) was employed to explore the effect of near amino acid proximity on protein antigenicity. Cross covariances Cjk(lag) between two distinct z-scales, j and k, were calculated with Equation (2):

(2)

VaxiJen v2.0 was used to estimate the antigenicity of the whole-protein chimera. Based on this server, the antigenicity score of the final protein was 0.5830 (Probable ANTIGEN) with a threshold of 0.4. Likewise, ANTIGENpro [24] was utilized to predict the antigenicity of the protein chimera. ANTIGENpro is an alignment-free, sequence-based, and pathogen-independent protein antigenicity predictor. ANTIGENpro is the first indicator of protein antigenicity that is trained to employ reactivity data from the protein microarray analysis of five pathogens.

AllerTOP v2.0 was used to predict the allergenicity of both B-cell and HTL epitopes. Protein sequences are sent to this

server in simple text. The results page then provides the identity of an allergen as “probable allergen” or “probable nonallergen.” The whole-protein chimera was predicted as a “probable nonallergen” using this tool [20]. This server was chosen because of its high sensitivity (94%) and higher rate of accurate prediction (94%-100%) in comparison to other similar servers to predict allergenicity [20]. Similar to VaxiJen, this database analyzes the presentation of protein sequences by z-descriptors and ACC transformation [17,18].

ToxinPred [25] with a protein fragment length of 10 was used to predict the toxicity of both B-cell and HTL epitopes. ToxinPred is a computational tool that was built to anticipate and design toxic versus nontoxic proteins. The primary data set used for this approach is comprised of 1805 toxic proteins (≤35 residues). This server also was used to predict the toxicity of the whole-protein chimera and no fragment was predicted as a toxin.

Construction of the Chimeric Protein

Selected B-cell and HTL epitopes were used to construct the protein chimera as a multiepitope vaccine. Overlaps of B-cell and HTL epitopes were merged. Bilysine (KK) linkers, as flexible linkers, were used to connect the epitopes. The KK linker was implanted between separate epitopes to maintain their independent immunological functions (Figure 1B). KK is the target sequence of cathepsin B, which is one of the essential antigen-processing proteases in MHC class II antigen presentation [26].

Amino Acid Composition, Physicochemical Properties, and Solubility Prediction

The Protparam database [27] was used to calculate and predict the molecular weight, isoelectric point (pI), in vivo and in vitro half-life, instability index II, and grand average of hydropathicity (GRAVY). ProtParam from the ExPASy server is a reliable algorithm to compute physicochemical properties. However, it uses a single sequence per analysis through the interface. The instability index is calculated using weight values, as shown in Equation (3) [28]:

(3)

where L is the length of the sequence and DIWV(x[i]x[i+1]) is the instability weight value for the dipeptide starting in position i. A protein with an instability index less than 40 is anticipated to be stable, whereas one with an index greater than 40 is predicted to be unstable.

The relative volume occupied by aliphatic side chains (alanine, valine, isoleucine, and leucine) is known as the aliphatic index, which is considered a potentially beneficial element in the enhancement of globular protein thermostability. The aliphatic index is calculated by the formula X(Ala)+aX(Val)+b(X[Ile]+X[Leu]) [29], where X(Ala), X(Val), X(Ile), and X(Leu) represent the mole percent (100×mole fraction) of alanine, valine, isoleucine, and leucine, respectively, and the coefficients a and b are the relative volumes of the valine side chain (a=2.9) and Leu/Ile side chains (b=3.9) to the side chain of alanine.

The SOLpro program from ANTIGENpro [24] was used to predict the solubility of protein chimera upon overexpression. SOLpro predicts the tendency of a protein to be soluble when overexpressed in *Escherichia coli* using a two-stage support vector machine model based on multiple representations of the primary sequence.

The PepCalc server [30] was used to predict the solubility of the final protein, which provides only a very rough estimation of water solubility.

Secondary Structure Prediction

The Prabi server [31] was used to predict the secondary structure of the final sequence of the protein chimera. All PRABI components provide services in their various areas of expertise (eg, molecular, phylogeny, genomics, transcriptomics, proteomics, protein structure, and medical biostatistics). “GOR IV” was selected as the secondary structure prediction method. The program outputs two files: one with the sequence and anticipated secondary structure in rows (H=helix, E=extended or beta strand, and C=coil), and the other with the probability values for each secondary structure at each amino acid position (H=helix, E=extended or beta strand, and C=coil) [32].

The PSIPRED 4.0 [33] server was also used to predict the secondary structure, which provides more details of residues’ configurations. This is a very simple system of secondary prediction based on a simple neural network evaluation of PSI-BLAST–generated profiles, which is capable of generating findings that place the process at the very top of the prediction system crop [33].

Molecular Docking of Final Vaccine Epitopes With MHC Molecules

PEP-FOLD 2.0 from the RPBS Web Portal server [32] was used to predict the tertiary structure of the vaccine construct epitopes. PEP-FOLD is an online tool that was designed to model 3D protein conformation structures in aqueous solutions for proteins 9-25 amino acids in length (de novo modeling). PEP-FOLD conducts a series of 50 simulations beginning with an amino acid sequence, and returns the most critical energy and population-related conformations found [32].

The ClusPro 2.0 server [34,35] rotates the ligands of each of the final epitopes of a vaccine protein with 70,000 rotations. The ligand rotations are translated relative to the MHC receptor alleles in three axes (x, y, z) on a grid. The top 1000 lowest energy docked structures from 70,000 rotations are then chosen and processed in turn. This set might have the potential to consist of at least some models that are close to the native structure of the complex. The server then clusters the 1000 rotations by finding the structure with the most “neighbors” within a 9 Å interface root mean square deviation radius as the distance measure. This ligand and its neighbors are then considered as the “cluster center” and the “members” of the cluster, respectively. This process was repeated for the remainder of the ligands to find the next clusters. Finally, the server provides a score for the models and reports the top scoring models based on the cluster size (10 most populated clusters) [34,35]. One of the main advantages of ClusPro 2.0 as an

automated protein docking server is its ability to generate protein-protein complexes with high accuracy [36].

PyMOL software was used to analyze the docking results. PyMOL is mostly utilized for molecular visualization by crystallographic, molecular dynamic simulation, and protein modeling software packages [37].

Immune Response Simulation

IL4-, IL10-, and IFNγ-inducing proteins from the 8 epitopes in the final vaccine construct were predicted via IL4pred server [38], IL-10Pred server [39], and IFNepitope server [40], respectively.

The immune response to vaccine injection was simulated using the C-ImmSim 10.1 server [41]. C-ImmSim 10.1 is an agent-based computational immune-response simulator that utilizes a position-specific score matrix and machine-learning methods for predicting epitope and immune interactions, respectively [42]. We regulated the parameters based on the predominant human leukocyte antigen (HLA) alleles of predictions. The host HLA selection parameters for MHC class I were set to A1010, A1101, and B0702; the parameters for DR MHC class II were set to DBR1_0101; and the time step to injection was set to 1, 84, and 100 (maximum allowed value), respectively. We randomly shuffled the vaccine protein sequence (without adjuvants) using Stothard *P* 2000 from the Sequence Manipulation Suite server [43] to create a control group. The overall immunogenicity of the generic protein sequence associated with its amino acid sequence was assessed by this immune system simulation server [41]. The entire simulation was focused on three events, (1) B-cell epitopes binding, (2) HLA class I and II epitopes binding, and (3) T-cell receptor binding, in which the HLA–protein complex interaction should be present. Such processes are independently carried out by cells through various agents and the consumption of specific simulated biological quantities [41].

Results

Selection of Protein Sequences

The amino acid sequences of E protein and S protein of SARS-CoV-2 were collected from the NCBI virus database with accession numbers QHD43418 and QHR63280, respectively, which were released January 13, 2020, and have nucleotide completeness. The FASTA sequences were used to construct a multiepitope vaccine against SARS-CoV-2.

B-Cell Epitopes Analysis

The ABCpred database reviewed the full-length E protein sequence for the analysis of linear B-cell epitopes. Among the results that passed the three filters of antigenicity, allergenicity, and toxicity, two epitopes (NVSLVKPSFYVYSRVK and YVYSRVKNLNSSRVPD) were chosen as protective epitopes (Table 1). The IEDB was then utilized to investigate the B-cell linear epitopes, resulting in 37 epitopes that were experimentally identified for S protein [44-54]. By contrast, there were no experimental B-cell epitopes for E protein of SARS-CoV-2.

T-Cell Epitopes Analysis

The binding epitopes to MHC class II molecules of E protein were analyzed by the IEDB. We used the prediction method to identify T-cell epitopes of E protein since there were no corresponding experimental epitopes in this database, whereas we used the available experimental T-cell epitopes of S protein [55]. The same three filters of antigenicity, allergenicity, and toxicity were applied to identify the protective antigens. Based on the number of alleles, the predominant HLA alleles were HLA-DRA-DBR1*01:01 and HLA-DPB1*01:02 among MHC class II alleles.

Antigenicity of Potential Epitopes

The antigenicity of both the B-cell and T-cell epitopes was predicted by VaxiJen 2.0, with a threshold of 0.4. The predicted epitopes with an antigenicity score above the threshold were considered as “antigen” epitopes. Screenings for the other two filters (allergenicity and toxicity) were not carried out on “nonantigen” epitopes (Table 1 and Table 2).

Table 1. Predicted T-cell and B-cell epitopes of SARS-CoV-2 envelope protein.[a]

Epitope	B-cell	MHC[b] II	Antigenicity	Allergenicity	Toxicity
ABCpred					
Not selected for vaccine construction					
TLAILTALRLCAYCCN	+[c]	–[d]	Antigen	Nonallergen	Toxin
LCAYCCNIVNVSLVKP	+	–	Antigen	Nonallergen	Toxin
FVSEETGTLIVNSVLL	+	–	Nonantigen	Discontinued	Discontinued
Selected for vaccine construction					
NVSLVKPSFYVYSRVK	+	–	Antigen	Nonallergen	Nontoxin
YVYSRVKNLNSSRVPD	+	+	Antigen	Nonallergen	Nontoxin
IEDB[e]					
Not selected for vaccine construction					
IVNSVLLFLAFVVFL	–	+	Antigen	Allergen	Discontinued
EETGTLIVNSVLLFL	–	+	Antigen	Allergen	Discontinued
GTLIVNSVLLFLAFV	–	+	Nonantigen	Discontinued	Discontinued
IVNSVLLFLAFVVFL	–	+	Nonantigen	Discontinued	Discontinued
MYSFVSEETGTLIVN	–	+	Nonantigen	Discontinued	Discontinued
NIVNVSLVKPSFYVY	–	+	Antigen	Allergen	Discontinued
RVKNLNSSRVPDLLV	–	+	Antigen	Allergen	Discontinued
SEETGTLIVNSVLLF	–	+	Nonantigen	Discontinued	Discontinued
TGTLIVNSVLLFLAF	–	+	Nonantigen	Discontinued	Discontinued
YSFVSEETGTLIVNS	–	+	Nonantigen	Discontinued	Discontinued
Selected for vaccine construction					
SFYVYSRVKNLNSSR	–	+	Antigen	Nonallergen	Nontoxin
FYVYSRVKNLNSSRV	–	+	Antigen	Nonallergen	Nontoxin
YVYSRVKNLNSSRVP	–	+	Antigen	Nonallergen	Nontoxin
FLAFVVFLLVTLAIL	–	+	Antigen	Nonallergen	Nontoxin
FVSEETGTLIVNSVL	–	+	Antigen	Nonallergen	Nontoxin
KPSFYVYSRVKNLNS	–	+	Antigen	Nonallergen	Nontoxin
YSRVKNLNSSRVPDL	–	+	Antigen	Nonallergen	Nontoxin
NSVLLFLAFVVFLLV	–	+	Antigen	Nonallergen	Nontoxin
VKPSFYVYSRVKNLN	–	+	Antigen	Nonallergen	Nontoxin
VNSVLLFLAFVVFLL	–	+	Antigen	Nonallergen	Nontoxin
VSLVKPSFYVYSRVK	–	+	Antigen	Nonallergen	Nontoxin
VVFLLVTLAILTALR	–	+	Antigen	Nonallergen	Nontoxin
LLFLAFVVFLLVTLA	–	+	Antigen	Nonallergen	Nontoxin

[a]T-cell epitopes were identified as the best epitopes based on the number of alleles.

[b]MHC: major histocompatibility complex.

[c]+: Related.

[d]–: Unrelated.

[e]IEDB: Immune Epitope Database.

Table 2. Experimental T-cell and B-cell epitopes of SARS-CoV-2 spike protein from Immune Epitope Database.

Epitope	Selected for vaccine construction	B-cell	MHC[a] II	Allergenicity	Toxicity
AATKMSECVLGQSKRVD	No	+[b]	–[c]	Allergen	Discontinued
CKFDEDDSEPVLKGVKLHYT	Yes	+	–	Nonallergen	Nontoxin
DDSEPVLKGVKLHYT	Yes	+	–	Nonallergen	Nontoxin
DKYFKNHTSPDVDLGD	Yes	+	–	Nonallergen	Nontoxin
DLGDISGINASVVNIQK	No	+	–	Allergen	Discontinued
EIDRLNEVAKNLNESLIDLQELGKYEQY	Yes	+	–	Nonallergen	Nontoxin
EVAKNLNESLIDLQELG	No	+	–	Allergen	Discontinued
KNHTSPDVDLGDISGIN	No	+	–	Allergen	Discontinued
LYQDVNC	No	+	–	Allergen	Discontinued
LYQDVNCT	No	+	–	Allergen	Discontinued
MAYRFNGIGVTQNVLY	Yes	+	–	Nonallergen	Nontoxin
MAYRFNGIGVTQNVLYE	Yes	+	–	Nonallergen	Nontoxin
RASANLAATKMSECVLG	No	+	–	Allergen	Discontinued
SPDVDLGDISGINAS	No	+	–	Allergen	Discontinued
MAYRFNGIGVTQNVLY	Yes	–	+	Nonallergen	Nontoxin
QLIRAAEIRASANLAATK	No	–	+	Allergen	Discontinued

[a]MHC: major histocompatibility complex.

[b]+: Related.

[c]–: Unrelated.

Allergenicity of Potential Epitopes

The allergenicity of both B-cell and HTL epitopes was estimated by AllerTOP v. 2.0 (Table 1).

Toxicity of Potential Epitopes

The toxicity of both B-cell and HTL epitopes was predicted by ToxinPred, with a protein fragment length of 10 (Tables 1 and 2).

Construction of the Chimeric Protein

The screened epitopes were chosen for the design of a chimeric protein as a multiepitope vaccine. As shown in Tables 1 and 2, we selected 21 epitopes (including 6 B-cell epitopes and 1 HTL epitope of S protein, and 2 B-cell epitopes and 12 HTL epitopes of E protein), which all filled the criteria of "antigen," "nonallergen," and "nontoxin." To establish a contiguous sequence in the final construction, the overlapping sequences of B-cell and T-cell epitopes were merged. In detail, MAYRFNGIGVTQNVLYE was obtained from MAYRFNGIGVTQNVLY (S protein HTL epitope), MAYRFNGIGVTQNVLY, and MAYRFNGIGVTQNVLYE (S protein B-cell epitopes); CKFDEDDSEPVLKGVKLHYT was obtained from CKFDEDDSEPVLKGVKLHYT and DDSEPVLKGVKLHYT (S protein B-cell epitopes); SFYVYSRVKNLNSSRVPDL was obtained from YVYSRVKNLNSSRVPD (E protein B-cell epitope), SFYVYSRVKNLNSSR, YVYSRVKNLNSSRVP, and YSRVKNLNSSRVPDL (E protein HTL epitopes); VNSVLLFLAFVVFLLVTLAILTALR was obtained from VVFLLVTLAILTALR, LLFLAFVVFLLVTLAILTA, FLAFVVFLLVTLAIL, NSVLLFLAFVVFLLV, and VNSVLLFLAFVVFLL (E protein HTL epitopes); and NVSLVKPSFYVYSRVKNLNS was obtained from NVSLVKPSFYVYSRVK (E protein B-cell epitope), VSLVKPSFYVYSRVK, VKPSFYVYSRVKNLN, and KPSFYVYSRVKNLNS (E protein HTL epitopes). Predicted linear B-cell epitopes and T-cell epitopes were connected utilizing KK linkers as flexible connectors (Figure 1B).

The arrangement of epitopes in the final vaccine construct had a substantial effect on the physicochemical properties such as half-life and instability, with the half-life varying from 5.5 hours to 30 hours in mammalian cells, and the stability varying from an unstable to a completely stable protein simply by changing the order of epitopes. Therefore, we further investigated the properties of more than 40 possible permutations considering the overlaps of the selected epitopes to find the best formulation of this vaccine candidate.

Antigenicity, Allergenicity, and Toxicity Estimation of the Candidate Multiepitope Vaccine

The antigenicity of the final protein chimera (Figure 1B) was estimated by the VaxiJen 2.0 server to be 0.5830 with a threshold of 0.4. The ANTIGENpro platform was also utilized to estimate the antigenicity of the final protein. Based on this server, the whole protein (Figure 1B) is predicted as an antigen with a probability of 0.415508. The AllerTOP v.2.0 server indicated that the final protein is predicted as a "nonallergen"

and the ToxinPred server predicted the final protein as a "nontoxin."

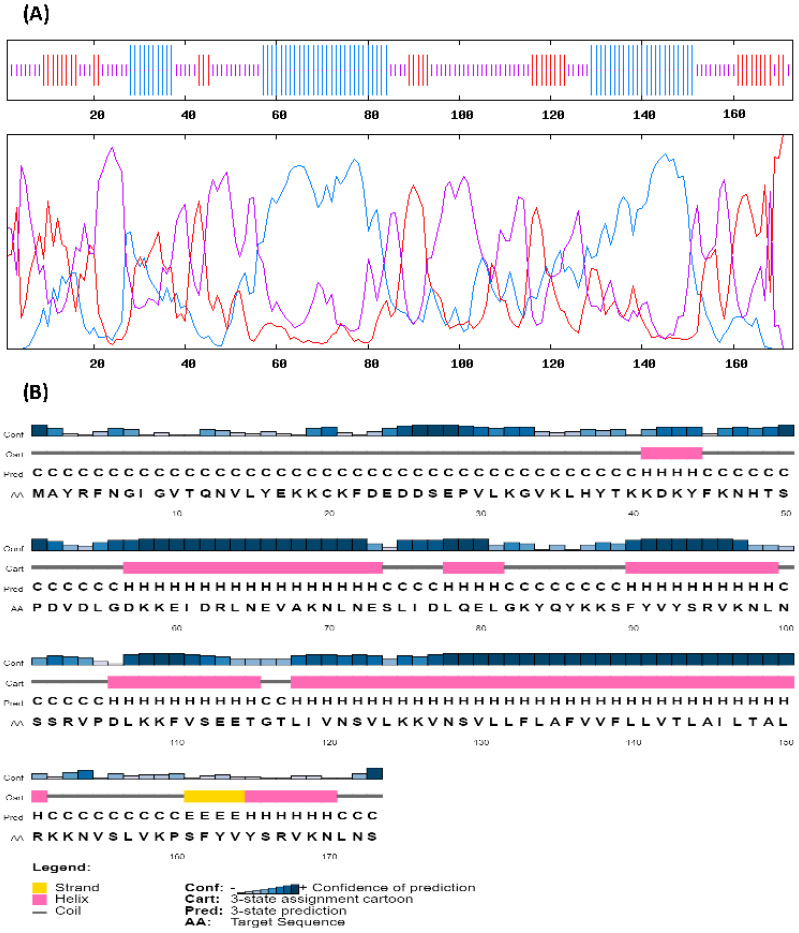
Amino Acid Composition, Physicochemical Properties, and Solubility Prediction

Based on the Protparam database, the final protein chimera comprised 173 amino acids (Figure 1B) with a molecular weight of 19.9 kDa. The pI value was predicted to be 9.57. The half-life was estimated to be 30 hours in mammalian reticulocytes in vitro, more than 20 hours in yeast, and over 10 hours in *E. coli* in vivo. Instability index II was predicted to be 26.45, classifying the protein as stable (an index>40 indicates instability). The aliphatic index was 101.21, indicating high thermostability. The estimated GRAVY value was –0.293. This negative attribute indicates that the protein is hydrophilic and can react with water molecules. Furthermore, based on the PepCalc server, the solubility was predicted to be "good" in water. Based on SOLpro from ANTIGENpro, the protein chimera was expected to be soluble with a probability of 0.765767.

Secondary Structure Prediction

The secondary structure of the final protein (Figure 1B) was analyzed by the Prabi server. The final chimeric protein was estimated to include 35.26% alpha helices, 20.81% extended strands, and 43.93% random coils (Figure 2A). The detailed secondary structure predicted by the PSIPRED 4.0 server, including the residues and their configurations, is shown in Figure 2B.

Figure 2. Properties of the final vaccine protein candidate. (A) Secondary structure of the final protein predicted by the Prabi server. Red: extended strands; blue: alpha helix configurations; purple: random coils. (B) The residues and their arrangements analyzed by the PSIPRED 4.0 server. Yellow and pink regions indicate strands and helices, respectively; gray linkers indicate coil configurations.



Molecular Docking of Final Epitopes of the Vaccine With MHC Molecules

The crystal structures of HLA-DRA-DBR1*01:01 and HLA-DPB1*01:02 were retrieved from the PDB RCSB database (PDB ID: 1AQD and 3LQZ, respectively). The PDB files were edited and cleaned from heteroatoms. PEP-FOLD 2.0 from the RPBS web portal server was used to predict the tertiary structure of 8 epitopes of the vaccine construct individually. Molecular docking was performed on the epitopes and the whole vaccine construct with relevant MHC alleles using the ClusPro 2.0 online server. PyMOL software was used to perform a detailed analysis of the interface of protein-protein interactions (Figure 3). The weighted score of the lowest energy docked complexes are reported in Table 3. The best way to rank the model is according to the cluster size (number of members) [34,35]. The most populated clusters were found in MAYRFNGIGVTQNVLYE and HLA-DPA1*01:03, DKYFKNHTSPDVDLGD and HLA-DPA1*01:03, CKFDEDDSEPVLKGVKLHYT and HLA-DPA1*01:03, and EIDRLNEVAKNLNESLIDLQELGKYQY and HLA-DRB1*01:01, with 784, 577, 350, and 261 cluster members, respectively.

Figure 3. Molecular docking analysis. (A) VNSVLLFLAFVVFLLVTLAILTALR epitope (green) and HLA-DPA1*01:03 protein (blue). (B) SFYVYSRVKNLNSSRVPDL epitope (yellow) and HLA-DPA1*01:03 protein (blue). (C) MAYRFNGIGVTQNVLYE epitope (red) and HLA-DPA1*01:03 protein (blue). (D) NVSLVKPSFYVYSRVKNLNS epitope (dark blue) and HLA-DPA1*01:03 protein (blue). (E) FVSEETGTLIVNSVL epitope (pink) and HLA-DPA1*01:03 protein (blue). (F) VNSVLLFLAFVVFLLVTLAILTALR epitope (green) and HLA-DRB1*01:01 protein (light pink). (G) EIDRLNEVAKNLNESLIDLQELGKYQY epitope (light blue) and HLA-DRB1*01:01 protein (light pink). (H) NVSLVKPSFYVYSRVKNLNS epitope (dark blue) and HLA-DRB1*01:01 protein (light pink). (I) MAYRFNGIGVTQNVLYE epitope (red) and HLA-DRB1*01:01 protein (light pink). HLA: human leukocyte antigen.

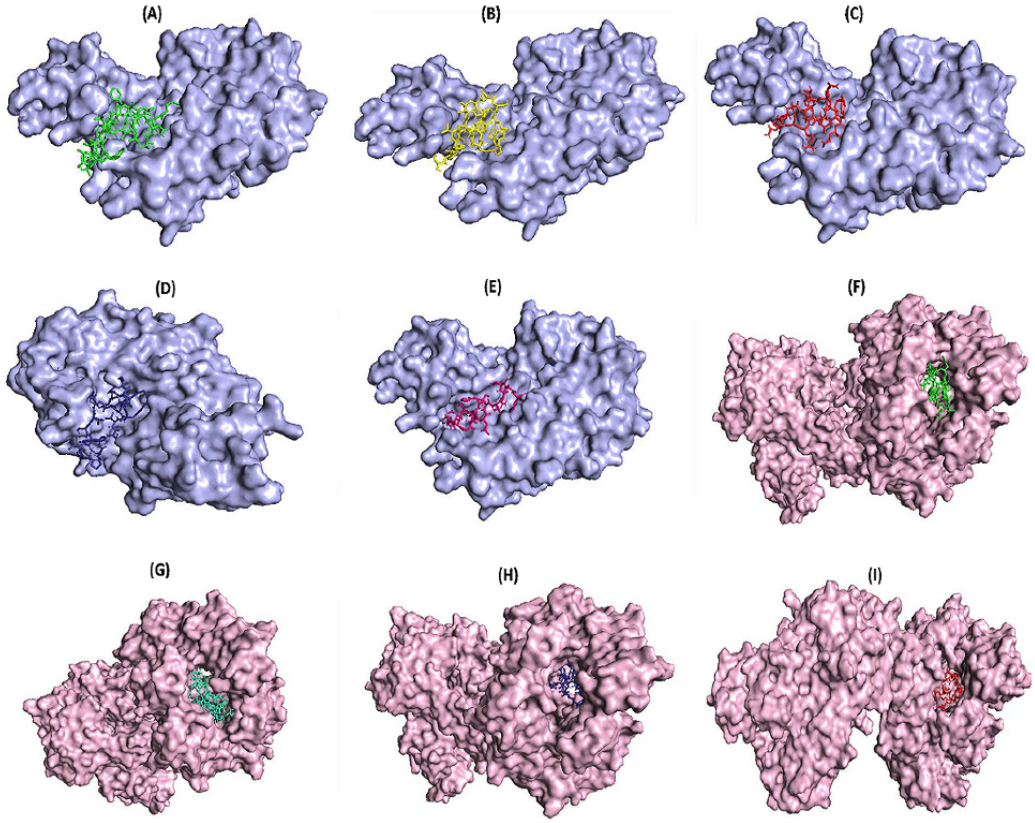


Table 3. Docking results and prediction of the immunity effects of epitopes.

Vaccine epitopes	Weighted scores[a] of the complex docked with				
	HLA[b]-DPA1*01:03	HLA-DRB1*01:01	IL[c]4 inducer	IL10 inducer	IFN[d]γ inducer
MAYRFNGIGVTQNVLYE	–869.2	–962.2	–[e]	–	–
CKFDEDDSEPVLKGVKLHYT	–758.7	–810.6	+[f]	–	+
DKYFKNHTSPDVDLGD	–763.2	–755	+	–	–
EIDRLNEVAKNLNESLIDLQELGKYQY	–715.4	–1050.7	–	+	+
SFYVYSRVKNLNSSRVPDL	–875.5	–992.2	+	+	+
FVSEETGTLIVNSVL	–798.2	–823.9	+	+	–
VNSVLLFLAFVVFLLVTLAILTALR	–1148.3	–1477	–	+	+
NVSLVKPSFYVYSRVKNLNS	–852.7	–1036.6	+	+	+

[a]The weighted scores of the lowest energy docked structures were based on the cluster size of the most populated cluster.

[b]HLA: human leukocyte antigen.

[c]IL: interleukin.

[d]IFN: interferon.

[e]–: Unrelated.

[f]+: Related.

Immune Response Simulation

We predicted the IL4, IL10, and IFNγ inducing proteins from the 6 epitopes in the final vaccine construct via IL4pred server, IL10pred server, and IFNepitope server, respectively. The results are shown in Table 3.

The primary and secondary immune responses were stimulated by the C-ImmSim 10.1 server. This server simulated the immune response of vaccine candidates with three injections in the time steps of 1, 84, and 100; each time step is equal to 8 hours. To perform a relative comparison, we created a shuffled sequence of the vaccine candidate as a control protein, and we analyzed the results of the immune response simulation to the injection of the control. This shuffled sequence was employed to evaluate the significance of the vaccine sequence results, because in immune response simulation by this server, the sequence composition of the final epitopes connected via KK linkers is an important consideration. The results of the vaccine injection clearly varied from those of the controls (Figure 4 and Figure 5).

Figure 4. In silico immune response simulation to the injection of the candidate vaccine and control protein by the C-ImmSim 10.1 server. The simulation was performed with three injections in the time steps of 1, 84, and 100; each time step is equal to 8 hours. (A) B-cell population. (B) B-cell population per state. (C) T helper (TH) cell population. (D) TH cell population per state. (E) T cytotoxic (TC) cell population. (F) TC cell population per state. (G) Macrophage (MA) cell population. (H) Natural killer (NK) cell population. (I) Immunoglobulins. (J) Cytokines. (K) Cytokines after the protein control injection. (L) Immunoglobulins following the protein control injection.

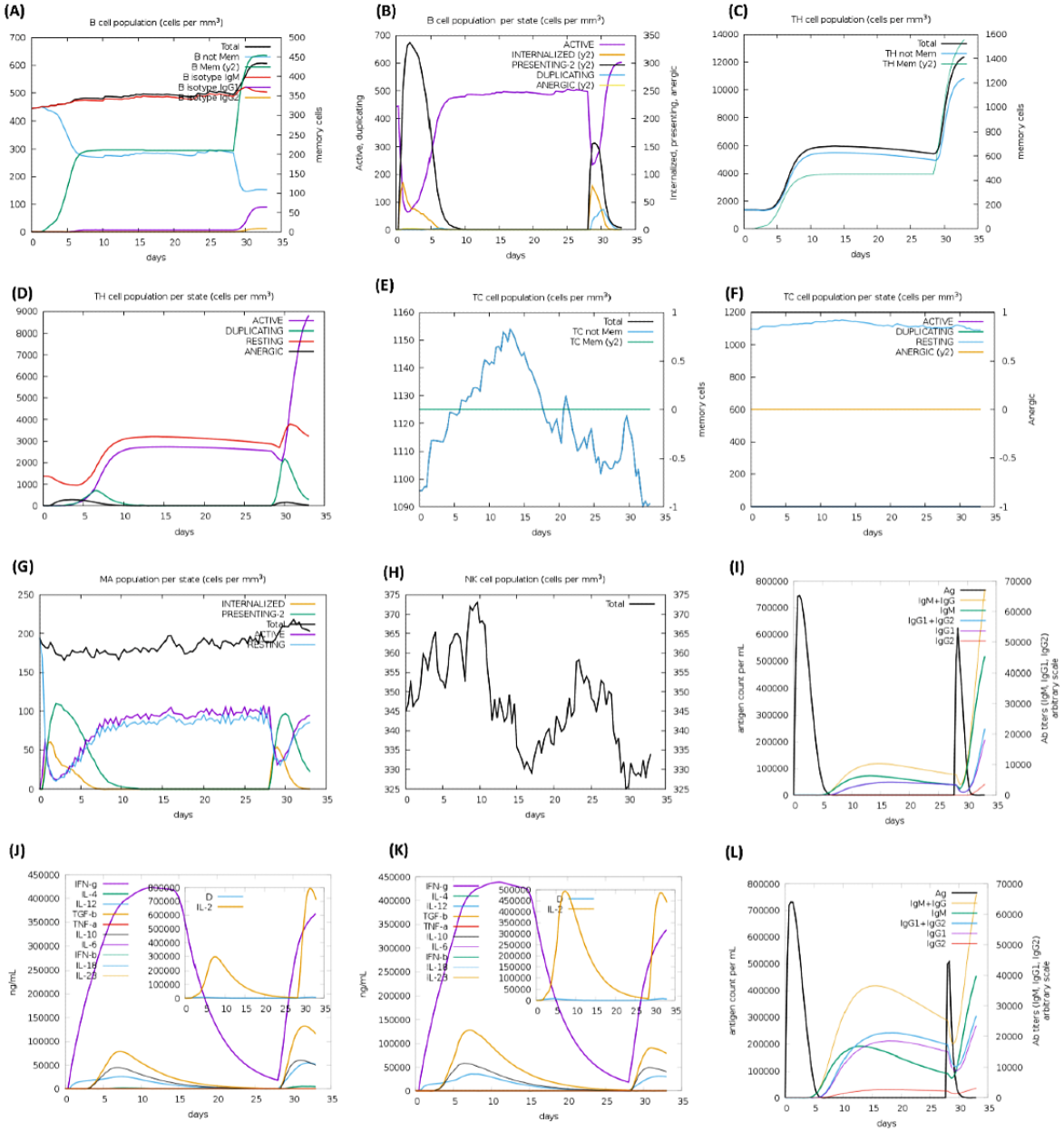
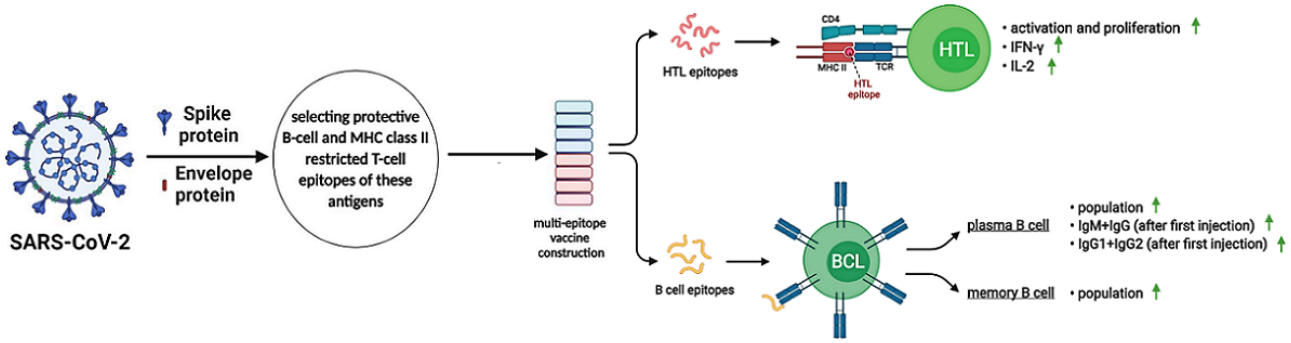


Figure 5. Graphical representation of immune response simulation to the injection of the vaccine candidate against SARS-CoV-2 spike (S) and envelope (E) proteins. HTL: helper T lymphocyte; BCL: B-cell lymphoma; IgM: immunoglobulin M; IgG: immunoglobulin G; IFN-γ: interferon gamma; IL-2: interleukin-2; TCR: T-cell receptor; MHC class II: major histocompatibility complex class II; CD4: cluster of differentiation 4. The image was created using the Biorender illustrator tool.



Discussion

Principal Findings

EVs offer a new strategy for the prophylactic and therapeutic use of pathogen-specific immunity [56]. A multiepitope vaccine consisting of a protein series or overlapping proteins has been proposed as an appropriate solution to the prevention and treatment of viral infections [57-62]. The perfect multiepitope vaccine should be engineered to include epitopes that can activate cytotoxic T lymphocytes, T-cells, and B-cells, and trigger successful responses to specific viruses [57].

We here present the in silico design of a potential multiepitope vaccine against the S and E proteins of SARS-CoV-2, which comprises both B-cell and HTL epitopes and can stimulate the immune system responses impressively. Immune interference is less likely to be a concern for multicomponent vaccines against a specific organism. For multitarget vaccinations, a strong response to one immune agent may reduce the otherwise marginal reaction to the second immunogen, and thus render the individual susceptible to infection with the pathogen corresponding to the second immune agent [63]. Since the SARS-CoV-2 S glycoprotein is surface-exposed and facilitates entry into host cells, it is the major priority of neutralizing antibodies against infection and the target of therapeutic and vaccine development [64,65]. S protein is also a primary focus for the design of subunit vaccines for SARS-CoV and Middle East Respiratory Syndrome (MERS)-CoV [66]. S trimers are widely coated with N-linked glycans, which are crucial for efficient folding and for modulating accessibility to host proteases and neutralizing antibodies [65-70]. E protein is conserved in all coronaviruses and covers the entire surface of SARS-CoV-2 (Figure 1A). There are fewer toxic epitopes of E protein than found for S protein. This finding was verified in the literature, in which E protein was explored in SARS-CoV in 2003 and, more recently, in MERS-CoV, demonstrating the retention of this protein in seven strains using the BioEdit Package tool and less toxic regions than in the S protein [12,57-62]. Several studies have examined the potential of coronaviruses with mutated E protein, focusing specifically on SARS- and MERS-CoV, as live attenuated vaccine candidates associated with hopeful results [12,71-75]. We obtained the FASTA sequence of the S and E proteins of SARS-CoV-2 from the NCBI database. B-cell and HTL epitopes of E protein were predicted by different servers, whereas experimentally confirmed epitopes were utilized for S protein. The epitopes were screened based on the three filters of antigenicity, allergenicity, and toxicity. Therefore, we selected only protective epitopes. We merged the overlaps of B-cell and T-cell epitopes and fused them with appropriate flexible linkers. Previous studies reported that KK linkers preserve independent immune responses when they are inserted between epitopes [26] (Figure 1B).

The absence of allergenic properties of the proposed protein chimera further increases its potential as a vaccine candidate [76]. Finally, the whole-protein chimera was analyzed for antigenicity, allergenicity, and toxicity, which was predicted as an antigen [22], nonallergen [20], and nontoxin [25]. The pI was calculated to be 9.57, which shows that the final protein is alkaline. The vaccine protein construct was predicted as "soluble" upon expression in the *E. coli* host. The structural stability of a vaccine is known to be an essential aspect of its effectiveness, which can ensure the appropriate presentation of antigens and thus efficiently activate the immune system [77,78]. The instability index II of our candidate was calculated to be 26.45, which indicates that this protein is "stable."

Secondary structure analysis predicted that the final protein consists of 35.26% alpha-helices, 20.81% extended strands, and 43.93% random coils. Essential types of "structural antigens" have been identified as natively unfolded protein regions and alpha-helical coil proteins. These two structural types, when examined in synthetic proteins, can fold into their native structure and are therefore recognized by antibodies naturally triggered in response to infection [76,79]. In the context of structural vaccinology, a molecular docking study was needed to predict the binding affinity of epitopes to the crystallized fragment of antibodies or MHC molecules [80,81]. To analyze the affinity of the final multiepitope vaccine to MHC molecules, we performed 16 docking simulations on the 8 epitopes of the final vaccine with MHC class II receptors. The results of docking analyses were notable, demonstrating the high affinity of the final epitopes of the vaccine construct to MHC molecules. The interface of protein-protein interactions was further considered using a visualization tool.

In the next step of designing a multiepitope vaccine, a systems vaccinology approach is beneficial in assessing the human complex immune response at different stages of biological structures [82]. Finally, we utilized an immune simulator server

to predict the primary and secondary responses of the immune system to three injections of the candidate vaccine. From the cytokines simulation plot, we noted an increase in the levels of IL-4 and IFNγ, which is similar to the clinical features of COVID-19 patients reported by Huang et al [2] (Figure 4J). Appropriate activation of APCs, high production of memory cells due to the extensive activation of B-cells and T-cells, control and clearance of antigens due to the creation of cytokines by the participation of T helper memory cells, and the evident long-term memory persistence after three injections could confirm the efficiency of our candidate vaccine [83].

Finally, we selected one of the multiepitope vaccine candidates with the lowest molecular weight (shortest sequence length), which can potentially result in low-cost manufacturing and shorten production times [84].

Comparison With Prior Work

Unlike most of the multiepitope vaccines that have been suggested during the COVID-19 pandemic, we preferred to design a vaccine without built-in adjuvants. Since adjuvants are necessary to increase the dosage efficacy by preventing the rapid degradation of proteins [85], tank-mixed adjuvants can be added to the final formulation. For instance, aluminum salts can be candidate adjuvants as they are used in various viral and bacterial vaccines and would be expected to enhance the antigen stability [86].

Adjuvants are effective in vaccine stability and can contribute to immunization reduction and enhanced antibody responses. Although our suggested vaccine lacks a built-in adjuvant, it demonstrates the same stability and half-life estimation as previously reported candidates. Our vaccine construct also showed roughly the highest AI in comparison with other suggested multiepitope vaccines, along with the highest thermostability [29]. These benefits are only achieved by the careful arrangement of selected epitopes in designing this vaccine construct. This study thus demonstrates the importance of testing various permutations of epitopes in vaccine properties.

Finally, discharging our multiepitope vaccine from built-in adjuvants can demonstrate that the immune simulation for two injections (Figure 4) was induced because of the designed vaccine without interference from any nonspecific immunization against the adjuvants.

Conclusion

The goal of this research was to suggest a computational method for predicting protective B-cell and T-cell epitopes of the E protein of SARS-CoV-2 accompanied by experimental epitopes of S protein to construct a chimeric protein vaccine candidate against this pandemic disease. The results demonstrated the high affinity of this chimeric protein to MHC molecules of the immune system, and the outputs of immune response simulation to the injection of this novel vaccine confirmed our findings. Thus, this multiepitope vaccine designed against the S and E proteins of SARS-CoV-2 utilizing immunoinformatics methods may be considered a new, safe, and efficient approach against SARS-CoV-2.

Acknowledgments

The authors are grateful to Mrs Hajipour for her English language comments on the initial draft of the manuscript.

Authors' Contributions

FG, FN, and VH developed the concept for the study. All authors discussed and designed the study. FG performed all parts of the study. RAC supervised the bioinformatics analyses. FN supervised the immunology experiments. HS, AHRSM, and MN contributed to acquisition and data analysis. VH supervised the whole project. FG wrote the first draft and all authors critically reviewed and approved the final version of the manuscript.

Conflicts of Interest

None declared.

References

1. Razai MS, Doerholt K, Ladhani S, Oakeshott P. Coronavirus disease 2019 (covid-19): a guide for UK GPs. BMJ 2020 Mar 05;368:m800. [doi: 10.1136/bmj.m800] [Medline: 32144127]
2. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020 Feb 15;395(10223):497-506 [FREE Full text] [doi: 10.1016/S0140-6736(20)30183-5] [Medline: 31986264]
3. Bing Z, Sakharkar K, Sakharkar M. In silico design of epitope-based vaccines. In: Dubitzky W, Wolkenhauer O, Cho KH, Yokota H, editors. Encyclopedia of systems biology. New York, NY: Springer; 2013:1003-1015.
4. Wu A, Peng Y, Huang B, Ding X, Wang X, Niu P, et al. Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. Cell Host Microbe 2020 Mar 11;27(3):325-328 [FREE Full text] [doi: 10.1016/j.chom.2020.02.001] [Medline: 32035028]
5. Pillaiyar T, Manickam M, Namasivayam V, Hayashi Y, Jung S. An overview of severe acute respiratory syndrome-coronavirus (SARS-CoV) 3CL protease inhibitors: peptidomimetics and small molecule chemotherapy. J Med Chem 2016 Jul 28;59(14):6595-6628 [FREE Full text] [doi: 10.1021/acs.jmedchem.5b01461] [Medline: 26878082]

6. Cui L, Wang H, Ji Y, Yang J, Xu S, Huang X, et al. The nucleocapsid protein of coronaviruses acts as a viral suppressor of RNA silencing in mammalian cells. J Virol 2015 Sep;89(17):9029-9043 [FREE Full text] [doi: 10.1128/JVI.01331-15] [Medline: 26085159]
7. Pillaiyar T, Meenakshisundaram S, Manickam M. Recent discovery and development of inhibitors targeting coronaviruses. Drug Discov Today 2020 Apr;25(4):668-688 [FREE Full text] [doi: 10.1016/j.drudis.2020.01.015] [Medline: 32006468]
8. Ou X, Liu Y, Lei X, Li P, Mi D, Ren L, et al. Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. Nat Commun 2020 Mar 27;11(1):1620. [doi: 10.1038/s41467-020-15562-9] [Medline: 32221306]
9. Bisht H, Roberts A, Vogel L, Subbarao K, Moss B. Neutralizing antibody and protective immunity to SARS coronavirus infection of mice induced by a soluble recombinant polypeptide containing an N-terminal segment of the spike glycoprotein. Virology 2005 Apr 10;334(2):160-165 [FREE Full text] [doi: 10.1016/j.virol.2005.01.042] [Medline: 15780866]
10. Bukreyev A, Lamirande EW, Buchholz UJ, Vogel LN, Elkins WR, St Claire M, et al. Mucosal immunisation of African green monkeys (Cercopithecus aethiops) with an attenuated parainfluenza virus expressing the SARS coronavirus spike protein for the prevention of SARS. Lancet 2004 Jun 26;363(9427):2122-2127 [FREE Full text] [doi: 10.1016/S0140-6736(04)16501-X] [Medline: 15220033]
11. Ramaiah A, Arumugaswami V. Insights into cross-species evolution of novel human coronavirus 2019-nCoV and defining immune determinants for vaccine development. bioRxiv. 2020. URL: https://www.biorxiv.org/content/10.1101/2020.01.29.925867v3 [accessed 2022-07-12]
12. Schoeman D, Fielding BC. Coronavirus envelope protein: current knowledge. Virol J 2019 May 27;16(1):69 [FREE Full text] [doi: 10.1186/s12985-019-1182-0] [Medline: 31133031]
13. de Jong AS, Visch H, de Mattia F, van Dommelen MM, Swarts HG, Luyten T, et al. The coxsackievirus 2B protein increases efflux of ions from the endoplasmic reticulum and Golgi, thereby inhibiting protein trafficking through the Golgi. J Biol Chem 2006 May 19;281(20):14144-14150 [FREE Full text] [doi: 10.1074/jbc.M511766200] [Medline: 16540472]
14. Xiang Z, Todd T, Ku KP, Kovacic BL, Larson CB, Chen F, et al. VIOLIN: vaccine investigation and online information network. Nucleic Acids Res 2008 Jan 23;36(Database issue):D923-D928 [FREE Full text] [doi: 10.1093/nar/gkm1039] [Medline: 18025042]
15. Brister J, Ako-Adjei D, Bao Y, Blinkova O. NCBI viral genomes resource. Nucleic Acids Res 2015 Jan;43(Database issue):D571-D577 [FREE Full text] [doi: 10.1093/nar/gku1207] [Medline: 25428358]
16. Larsen J, Lund O, Nielsen M. Improved method for predicting linear B-cell epitopes. Immunome Res 2006 Apr 24;2:2 [FREE Full text] [doi: 10.1186/1745-7580-2-2] [Medline: 16635264]
17. Saha S, Raghava GPS. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. Proteins 2006 Oct 01;65(1):40-48. [doi: 10.1002/prot.21078] [Medline: 16894596]
18. Saha S, Raghava G. Prediction methods for B-cell epitopes. In: Flower DR, editor. Immunoinformatics. Methods in Molecular Biology, vol 409. Totowa, NJ: Humana Press; 2007:387-394.
19. Vita R, Mahajan S, Overton J, Dhanda S, Martini S, Cantrell J, et al. The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res 2019 Jan 08;47(D1):D339-D343 [FREE Full text] [doi: 10.1093/nar/gky1006] [Medline: 30357391]
20. Dimitrov I, Bangov I, Flower DR, Doytchinova I. AllerTOP v.2--a server for in silico prediction of allergens. J Mol Model 2014 Jun 31;20(6):2278. [doi: 10.1007/s00894-014-2278-5] [Medline: 24878803]
21. Konstantinou G. T-Cell Epitope Prediction. Methods Mol Biol 2017;1592:211-222. [doi: 10.1007/978-1-4939-6925-8_17] [Medline: 28315223]
22. Doytchinova IA, Flower DR. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. BMC Bioinformatics 2007 Jan 05;8:4 [FREE Full text] [doi: 10.1186/1471-2105-8-4] [Medline: 17207271]
23. Wold S, Jonsson J, Sjörström M, Sandberg M, Rännar S. DNA and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures. Analytica Chimica Acta 1993 May;277(2):239-253. [doi: 10.1016/0003-2670(93)80437-p]
24. Cheng J, Randall AZ, Sweredoski MJ, Baldi P. SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res 2005 Jul 01;33(Web Server issue):W72-W76 [FREE Full text] [doi: 10.1093/nar/gki396] [Medline: 15980571]
25. Gupta S, Kapoor P, Chaudhary K, Gautam A, Kumar R, Open Source Drug Discovery Consortium, et al. In silico approach for predicting toxicity of peptides and proteins. PLoS One 2013 Sep 13;8(9):e73957 [FREE Full text] [doi: 10.1371/journal.pone.0073957] [Medline: 24058508]
26. Sarobe P, Lasarte J, Larrea E, Golvano J, Prieto I, Gullón A, et al. Enhancement of peptide immunogenicity by insertion of a cathepsin B cleavage site between determinants recognized by B and T cells. Res Immunol 1993 May;144(4):257-262. [doi: 10.1016/0923-2494(93)80102-5] [Medline: 7690980]
27. Gasteiger E, Hoogland C, Gattiker A, Wilkins M, Appel R, Bairoch A. Protein identification and analysis tools on the ExPASy server. In: Walker JM, editor. The proteomics protocols handbook. Totowa, NJ: Humana Press; 2005:571-607.
28. Guruprasad K, Reddy B, Pandit MW. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. Protein Eng 1990 Dec;4(2):155-161. [doi: 10.1093/protein/4.2.155] [Medline: 2075190]

29. Ikai A. Thermostability and aliphatic index of globular proteins. J Biochem 1980 Dec;88(6):1895-1898 [FREE Full text] [Medline: 7462208]
30. Lear S, Cobb SL. Pep-Calc.com: a set of web utilities for the calculation of peptide and peptoid properties and automatic mass spectral peak assignment. J Comput Aided Mol Des 2016 Mar 24;30(3):271-277 [FREE Full text] [doi: 10.1007/s10822-016-9902-7] [Medline: 26909892]
31. Garnier J. GOR secondary structure prediction method version IV. Meth Enzym 1996;266:540-553. [doi: 10.1016/S0076-6879(96)66034-0]
32. Maupetit J, Derreumaux P, Tuffery P. PEP-FOLD: an online resource for de novo peptide structure prediction. Nucleic Acids Res 2009 Jul 11;37(Web Server issue):W498-W503 [FREE Full text] [doi: 10.1093/nar/gkp323] [Medline: 19433514]
33. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999 Sep 17;292(2):195-202. [doi: 10.1006/jmbi.1999.3091] [Medline: 10493868]
34. Kozakov D, Beglov D, Bohnuud T, Mottarella SE, Xia B, Hall DR, et al. How good is automated protein docking? Proteins 2013 Dec;81(12):2159-2166 [FREE Full text] [doi: 10.1002/prot.24403] [Medline: 23996272]
35. Kozakov D, Hall DR, Xia B, Porter KA, Padhorny D, Yueh C, et al. The ClusPro web server for protein-protein docking. Nat Protoc 2017 Feb;12(2):255-278 [FREE Full text] [doi: 10.1038/nprot.2016.169] [Medline: 28079879]
36. Kozakov D, Hall DR, Beglov D, Brenke R, Comeau SR, Shen Y, et al. Achieving reliability and high accuracy in automated protein docking: ClusPro, PIPER, SDU, and stability analysis in CAPRI rounds 13-19. Proteins 2010 Nov 15;78(15):3124-3130 [FREE Full text] [doi: 10.1002/prot.22835] [Medline: 20818657]
37. Yuan S, Chan HS, Hu Z. Using PyMOL as a platform for computational drug design. WIREs Comput Mol Sci 2017 Jan 05;7(2):e1298. [doi: 10.1002/wcms.1298]
38. Dhanda SK, Gupta S, Vir P, Raghava GPS. Prediction of IL4 inducing peptides. Clin Dev Immunol 2013;2013:263952. [doi: 10.1155/2013/263952] [Medline: 24489573]
39. Nagpal G, Usmani SS, Dhanda SK, Kaur H, Singh S, Sharma M, et al. Computer-aided designing of immunosuppressive peptides based on IL-10 inducing potential. Sci Rep 2017 Feb 17;7(1):42851. [doi: 10.1038/srep42851] [Medline: 28211521]
40. Dhanda SK, Vir P, Raghava GP. Designing of interferon-gamma inducing MHC class-II binders. Biol Direct 2013 Dec 05;8(1):30 [FREE Full text] [doi: 10.1186/1745-6150-8-30] [Medline: 24304645]
41. Rapin N, Lund O, Castiglione F. Immune system simulation online. Bioinformatics 2011 Jul 15;27(14):2013-2014. [doi: 10.1093/bioinformatics/btr335] [Medline: 21685045]
42. Nain Z, Abdulla F, Rahman M, Karim M, Khan MSA, Sayed SB, et al. Proteome-wide screening for designing a multi-epitope vaccine against emerging pathogen using immunoinformatic approaches. J Biomol Struct Dyn 2020 Oct;38(16):4850-4867. [doi: 10.1080/07391102.2019.1692072] [Medline: 31709929]
43. Stothard P. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. Biotechniques 2000 Jun;28(6):1102, 1104 [FREE Full text] [doi: 10.2144/00286ir01] [Medline: 10868275]
44. Poh CM, Carissimo G, Wang B, Amrun SN, Lee CY, Chee RS, et al. Two linear epitopes on the SARS-CoV-2 spike protein that elicit neutralising antibodies in COVID-19 patients. Nat Commun 2020 Jun 01;11(1):2806. [doi: 10.1038/s41467-020-16638-2] [Medline: 32483236]
45. Wang Q, Zhang L, Kuwahara K, Li L, Liu Z, Li T, et al. Immunodominant SARS coronavirus epitopes in humans elicited both enhancing and neutralizing effects on infection in non-human primates. ACS Infect Dis 2016 May 13;2(5):361-376 [FREE Full text] [doi: 10.1021/acsinfecdis.6b00006] [Medline: 27627203]
46. Buus S, Rockberg J, Forsström B, Nilsson P, Uhlen M, Schafer-Nielsen C. High-resolution mapping of linear antibody epitopes using ultrahigh-density peptide microarrays. Mol Cell Proteomics 2012 Dec;11(12):1790-1800 [FREE Full text] [doi: 10.1074/mcp.M112.020800] [Medline: 22984286]
47. Zhao J, Huang Q, Wang W, Zhang Y, Lv P, Gao X. Identification and characterization of dominant helper T-cell epitopes in the nucleocapsid protein of severe acute respiratory syndrome coronavirus. J Virol 2007 Jun;81(11):6079-6088 [FREE Full text] [doi: 10.1128/JVI.02568-06] [Medline: 17392374]
48. Lien S, Shih Y, Chen H, Tsai J, Leng C, Lin M, et al. Identification of synthetic vaccine candidates against SARS CoV infection. Biochem Biophys Res Commun 2007 Jul 06;358(3):716-721 [FREE Full text] [doi: 10.1016/j.bbrc.2007.04.164] [Medline: 17506989]
49. Chow SCS, Ho CYS, Tam TTY, Wu C, Cheung T, Chan PKS, et al. Specific epitopes of the structural and hypothetical proteins elicit variable humoral responses in SARS patients. J Clin Pathol 2006 May;59(5):468-476 [FREE Full text] [doi: 10.1136/jcp.2005.029868] [Medline: 16461566]
50. Hu H, Li L, Kao RY, Kou B, Wang Z, Zhang L, et al. Screening and identification of linear B-cell epitopes and entry-blocking peptide of severe acute respiratory syndrome (SARS)-associated coronavirus using synthetic overlapping peptide library. J Comb Chem 2005;7(5):648-656. [doi: 10.1021/cc0500607] [Medline: 16153058]
51. Chan WS, Wu C, Chow SCS, Cheung T, To K, Leung W, et al. Coronaviral hypothetical and structural proteins were found in the intestinal surface enterocytes and pneumocytes of severe acute respiratory syndrome (SARS). Mod Pathol 2005 Nov;18(11):1432-1439 [FREE Full text] [doi: 10.1038/modpathol.3800439] [Medline: 15920543]

52. Lai S, Chong PC, Yeh C, Liu LS, Jan J, Chi H, et al. Characterization of neutralizing monoclonal antibodies recognizing a 15-residues epitope on the spike protein HR2 region of severe acute respiratory syndrome coronavirus (SARS-CoV). J Biomed Sci 2005 Oct;12(5):711-727 [FREE Full text] [doi: 10.1007/s11373-005-9004-3] [Medline: 16132115]
53. He Y, Zhou Y, Wu H, Luo B, Chen J, Li W, et al. Identification of immunodominant sites on the spike protein of severe acute respiratory syndrome (SARS) coronavirus: implication for developing SARS diagnostics and vaccines. J Immunol 2004 Sep 15;173(6):4050-4057 [FREE Full text] [doi: 10.4049/jimmunol.173.6.4050] [Medline: 15356154]
54. Guo J, Petric M, Campbell W, McGeer PL. SARS corona virus peptides recognized by antibodies in the sera of convalescent cases. Virology 2004 Jul 01;324(2):251-256 [FREE Full text] [doi: 10.1016/j.virol.2004.04.017] [Medline: 15207612]
55. Yang J, James E, Roti M, Huston L, Gebe JA, Kwok WW. Searching immunodominant epitopes prior to epidemic: HLA class II-restricted SARS-CoV spike protein epitopes in unexposed individuals. Int Immunol 2009 Jan;21(1):63-71 [FREE Full text] [doi: 10.1093/intimm/dxn124] [Medline: 19050106]
56. Khan AM, Miotto O, Heiny A, Salmon J, Srinivasan K, Nascimento EJ, et al. A systematic bioinformatics approach for selection of epitope-based vaccine targets. Cellular Immunology 2006 Dec;244(2):141-147. [doi: 10.1016/j.cellimm.2007.02.005]
57. Zhang L. Multi-epitope vaccines: a promising strategy against tumors and viral infections. Cell Mol Immunol 2018 Feb 11;15(2):182-184 [FREE Full text] [doi: 10.1038/cmi.2017.92] [Medline: 28890542]
58. Buonaguro L, HEPAVAC Consortium. Developments in cancer vaccines for hepatocellular carcinoma. Cancer Immunol Immunother 2016 Jan 21;65(1):93-99. [doi: 10.1007/s00262-015-1728-y] [Medline: 26093657]
59. Brennick CA, George MM, Corwin WL, Srivastava PK, Ebrahimi-Nik H. Neoepitopes as cancer immunotherapy targets: key challenges and opportunities. Immunotherapy 2017 Mar;9(4):361-371 [FREE Full text] [doi: 10.2217/imt-2016-0146] [Medline: 28303769]
60. Kuo T, Wang C, Badakhshan T, Chilukuri S, BenMohamed L. The challenges and opportunities for the development of a T-cell epitope-based herpes simplex vaccine. Vaccine 2014 Nov 28;32(50):6733-6745 [FREE Full text] [doi: 10.1016/j.vaccine.2014.10.002] [Medline: 25446827]
61. He R, Yang X, Liu C, Chen X, Wang L, Xiao M, et al. Efficient control of chronic LCMV infection by a CD4 T cell epitope-based heterologous prime-boost vaccination in a murine model. Cell Mol Immunol 2018 Sep 13;15(9):815-826 [FREE Full text] [doi: 10.1038/cmi.2017.3] [Medline: 28287115]
62. Lu I, Farinelle S, Sausy A, Muller CP. Identification of a CD4 T-cell epitope in the hemagglutinin stalk domain of pandemic H1N1 influenza virus and its antigen-driven TCR usage signature in BALB/c mice. Cell Mol Immunol 2017 Jun 9;14(6):511-520 [FREE Full text] [doi: 10.1038/cmi.2016.20] [Medline: 27157498]
63. Saul A, Fay MP. Human immunity and the design of multi-component, single target vaccines. PLoS One 2007 Sep 05;2(9):e850 [FREE Full text] [doi: 10.1371/journal.pone.0000850] [Medline: 17786221]
64. Walls AC, Park Y, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell 2020 Apr 16;181(2):281-292 [FREE Full text] [doi: 10.1016/j.cell.2020.02.058] [Medline: 32155444]
65. Rossen JWA, de Beer R, Godeke G, Raamsman MJB, Horzinek MC, Vennema H, et al. The viral spike protein is not involved in the polarized sorting of coronaviruses in epithelial cells. J Virol 1998 Jan;72(1):497-503. [doi: 10.1128/jvi.72.1.497-503.1998]
66. Wang N, Shang J, Jiang S, Du L. Subunit vaccines against emerging pathogenic human coronaviruses. Front Microbiol 2020 Feb 28;11:298. [doi: 10.3389/fmicb.2020.00298] [Medline: 32265848]
67. Walls AC, Tortorici MA, Frenz B, Snijder J, Li W, Rey FA, et al. Glycan shield and epitope masking of a coronavirus spike protein observed by cryo-electron microscopy. Nat Struct Mol Biol 2016 Oct 12;23(10):899-905 [FREE Full text] [doi: 10.1038/nsmb.3293] [Medline: 27617430]
68. Walls AC, Xiong X, Park Y, Tortorici MA, Snijder J, Quispe J, et al. Unexpected receptor functional mimicry elucidates activation of coronavirus fusion. Cell 2019 Feb 21;176(5):1026-1039 [FREE Full text] [doi: 10.1016/j.cell.2018.12.028] [Medline: 30712865]
69. Xiong X, Tortorici MA, Snijder J, Yoshioka C, Walls AC, Li W, et al. Glycan shield and fusion activation of a deltacoronavirus spike glycoprotein fine-tuned for enteric infections. J Virol 2018 Feb 15;92(4):e01628-17 [FREE Full text] [doi: 10.1128/JVI.01628-17] [Medline: 29093093]
70. Yang Y, Liu C, Du L, Jiang S, Shi Z, Baric RS, et al. Two mutations were critical for bat-to-human transmission of Middle East Respiratory Syndrome coronavirus. J Virol 2015 Sep;89(17):9119-9123. [doi: 10.1128/jvi.01279-15]
71. Regla-Nava JA, Nieto-Torres JL, Jimenez-Guardeño JM, Fernandez-Delgado R, Fett C, Castaño-Rodríguez C, et al. Severe acute respiratory syndrome coronaviruses with mutations in the E protein are attenuated and promising vaccine candidates. J Virol 2015 Apr;89(7):3870-3887 [FREE Full text] [doi: 10.1128/JVI.03566-14] [Medline: 25609816]
72. Netland J, DeDiego ML, Zhao J, Fett C, Álvarez E, Nieto-Torres JL, et al. Immunization with an attenuated severe acute respiratory syndrome coronavirus deleted in E protein protects against lethal respiratory disease. Virology 2010 Mar 30;399(1):120-128 [FREE Full text] [doi: 10.1016/j.virol.2010.01.004] [Medline: 20110095]
73. Almazán F, DeDiego ML, Sola I, Zuñiga S, Nieto-Torres JL, Marquez-Jurado S, et al. Engineering a replication-competent, propagation-defective Middle East respiratory syndrome coronavirus as a vaccine candidate. mBio 2013 Sep 10;4(5):e00650 [FREE Full text] [doi: 10.1128/mBio.00650-13] [Medline: 24023385]

74. Lamirande EW, DeDiego ML, Roberts A, Jackson JP, Alvarez E, Sheahan T, et al. A live attenuated severe acute respiratory syndrome coronavirus is immunogenic and efficacious in golden Syrian hamsters. J Virol 2008 Aug;82(15):7721-7724 [FREE Full text] [doi: 10.1128/JVI.00304-08] [Medline: 18463152]
75. Fett C, DeDiego ML, Regla-Nava JA, Enjuanes L, Perlman S. Complete protection against severe acute respiratory syndrome coronavirus-mediated lethal respiratory disease in aged mice by immunization with a mouse-adapted virus lacking E protein. J Virol 2013 Jun 15;87(12):6551-6559. [doi: 10.1128/jvi.00087-13]
76. Shey RA, Ghogomu SM, Esoh KK, Nebangwa ND, Shintouo CM, Nongley NF, et al. In-silico design of a multi-epitope vaccine candidate against onchocerciasis and related filarial diseases. Sci Rep 2019 Mar 13;9(1):4409. [doi: 10.1038/s41598-019-40833-x] [Medline: 30867498]
77. Negahdaripour M, Nezafat N, Eslami M, Ghoshoon MB, Shoolian E, Najafipour S, et al. Structural vaccinology considerations for in silico designing of a multi-epitope vaccine. Infect Genet Evol 2018 Mar;58:96-109. [doi: 10.1016/j.meegid.2017.12.008] [Medline: 29253673]
78. Scheiblhofer S, Laimer J, Machado Y, Weiss R, Thalhamer J. Influence of protein fold stability on immunogenicity and its implications for vaccine design. Expert Rev Vaccines 2017 May;16(5):479-489 [FREE Full text] [doi: 10.1080/14760584.2017.1306441] [Medline: 28290225]
79. Bennuru S, Cotton JA, Ribeiro JMC, Grote A, Harsha B, Holroyd N, et al. Stage-specific transcriptome and proteome analyses of the filarial parasite Onchocerca volvulus and its endosymbiont. mBio 2016 Dec 30;7(6):e02028. [doi: 10.1128/mbio.02028-16]
80. Agostino M, Mancera RL, Ramsland PA, Fernández-Recio J. Optimization of protein-protein docking for predicting Fc-protein interactions. J Mol Recognit 2016 Nov;29(11):555-568. [doi: 10.1002/jmr.2555] [Medline: 27445195]
81. Ribas‐Aparicio MR, Castelán‐Vega JA, Jiménez‐Alberto A, Monterrubio‐López GP, Aparicio‐Ozores G. The impact of bioinformatics on vaccine design and development. In: Afrin F, Hemeg H, Ozbak H, editors. Vaccines. London, UK: IntechOpen; 2017.
82. Raeven RHM, van Riet E, Meiring HD, Metz B, Kersten GFA. Systems vaccinology and big data in the vaccine development chain. Immunology 2019 Jan 13;156(1):33-46. [doi: 10.1111/imm.13012] [Medline: 30317555]
83. Six A, Bellier B, Thomas-Vaslin V, Klatzmann D. Systems biology in vaccine design. Microb Biotechnol 2012 Mar;5(2):295-304. [doi: 10.1111/j.1751-7915.2011.00321.x] [Medline: 22189033]
84. Strings DNA fragments. ThermoFisher Scientific. URL: https://www.thermofisher.com/us/en/home/life-science/cloning/gene-synthesis/gene-strings-dna-fragments.html [accessed 2022-07-12]
85. Sarkar I, Garg R, van Drunen Littel-van den Hurk S. Selection of adjuvants for vaccines targeting specific pathogens. Expert Rev Vaccines 2019 May 22;18(5):505-521 [FREE Full text] [doi: 10.1080/14760584.2019.1604231] [Medline: 31009255]
86. De Gregorio E, Caproni E, Ulmer JB. Vaccine adjuvants: mode of action. Front Immunol 2013;4:214. [doi: 10.3389/fimmu.2013.00214] [Medline: 23914187]

Abbreviations

ACC: auto cross covariance
APC: antigen-presenting cell
E protein: envelope protein
EV: epitope-based vaccine
FN: false negative
FP: false positive
GRAVY: grand hydropathic average
HLA: human leukocyte antigen
HTL: helper T lymphocyte
IEDB: Immune Epitope Database
IFN: interferon
IL: interleukin
IP10: interferon-inducible protein 10
KK: bilysine
MCP1: monocyte chemoattractant protein 1
MERS: Middle East respiratory syndrome
MHC: major histocompatibility complex
M protein: membrane protein
NCBI: National Center for Biotechnology Information
N protein: nucleocapsid protein
pI: isoelectric point
S protein: spike protein
TN: true negative

TP: true positive

Edited by A Mavragani; submitted 01.01.22; peer-reviewed by S Rostam Niakan Kalhori, A Banerjee; comments to author 27.04.22; revised version received 16.05.22; accepted 04.07.22; published 19.07.22.

Please cite as:
Ghafouri F, Ahangari Cohan R, Samimi H, Hosseini Rad S M A, Naderi M, Noorbakhsh F, Haghpanah V
Development of a Multiepitope Vaccine Against SARS-CoV-2: Immunoinformatics Study
JMIR Bioinform Biotech 2022;3(1):e36100
URL: https://bioinform.jmir.org/2022/1/e36100
doi:10.2196/36100
PMID:35891920

Original Paper

Mutational Patterns Observed in SARS-CoV-2 Genomes Sampled From Successive Epochs Delimited by Major Public Health Events in Ontario, Canada: Genomic Surveillance Study

David Chen[1], BMedSci; Gurjit S Randhawa[2], MSc, PhD; Maximillian PM Soltysiak[1], BSc; Camila PE de Souza[3], PhD; Lila Kari[4], PhD; Shiva M Singh[1], PhD; Kathleen A Hill[1], MSc, PhD

[1]Department of Biology, Western University, London, ON, Canada

[2]School of Mathematical and Computational Sciences, University of Prince Edward Island, Charlottetown, PE, Canada

[3]Department of Statistical and Actuarial Sciences, Western University, London, ON, Canada

[4]School of Computer Science, University of Waterloo, Waterloo, ON, Canada

Corresponding Author:
Kathleen A Hill, MSc, PhD
Department of Biology
Western University
1151 Richmond Street
London, ON, N6A 5B7
Canada
Phone: 1 519 661 2111 ext 81337
Email: khill22@uwo.ca

Abstract

Background: The emergence of SARS-CoV-2 variants with mutations associated with increased transmissibility and virulence is a public health concern in Ontario, Canada. Characterizing how the mutational patterns of the SARS-CoV-2 genome have changed over time can shed light on the driving factors, including selection for increased fitness and host immune response, that may contribute to the emergence of novel variants. Moreover, the study of SARS-CoV-2 in the microcosm of Ontario, Canada can reveal how different province-specific public health policies over time may be associated with observed mutational patterns as a model system.

Objective: This study aimed to perform a comprehensive analysis of single base substitution (SBS) types, counts, and genomic locations observed in SARS-CoV-2 genomic sequences sampled in Ontario, Canada. Comparisons of mutational patterns were conducted between sequences sampled during 4 different epochs delimited by major public health events to track the evolution of the SARS-CoV-2 mutational landscape over 2 years.

Methods: In total, 24,244 SARS-CoV-2 genomic sequences and associated metadata sampled in Ontario, Canada from January 1, 2020, to December 31, 2021, were retrieved from the Global Initiative on Sharing All Influenza Data database. Sequences were assigned to 4 epochs delimited by major public health events based on the sampling date. SBSs from each SARS-CoV-2 sequence were identified relative to the MN996528.1 reference genome. Catalogues of SBS types and counts were generated to estimate the impact of selection in each open reading frame, and identify mutation clusters. The estimation of mutational fitness over time was performed using the Augur pipeline.

Results: The biases in SBS types and proportions observed support previous reports of host antiviral defense activity involving the SARS-CoV-2 genome. There was an increase in U>C substitutions associated with adenosine deaminase acting on RNA (ADAR) activity uniquely observed during Epoch 4. The burden of novel SBSs observed in SARS-CoV-2 genomic sequences was the greatest in Epoch 2 (median 5), followed by Epoch 3 (median 4). Clusters of SBSs were observed in the spike protein open reading frame, ORF1a, and ORF3a. The high proportion of nonsynonymous SBSs and increasing dN/dS metric (ratio of nonsynonymous to synonymous mutations in a given open reading frame) to above 1 in Epoch 4 indicate positive selection of the spike protein open reading frame.

Conclusions: Quantitative analysis of the mutational patterns of the SARS-CoV-2 genome in the microcosm of Ontario, Canada within early consecutive epochs of the pandemic tracked the mutational dynamics in the context of public health events that instigate significant shifts in selection and mutagenesis. Continued genomic surveillance of emergent variants will be useful for the design of public health policies in response to the evolving COVID-19 pandemic.

(JMIR Bioinform Biotech 2022;3(1):e42243) doi:10.2196/42243

KEYWORDS

SARS-CoV-2; COVID-19; Ontario; virus; genetics; evolution; selection; mutation; epidemiology; variant

Introduction

SARS-CoV-2 is responsible for the global COVID-19 pandemic, and there have been 4,109,931 total confirmed COVID-19 cases in Canada as of August 12, 2022 [1]. As the most populated province in Canada, Ontario reported 1,394,524 confirmed COVID-19 cases and 52,998 hospitalizations as of August 13, 2022 [2], among a population estimated to be 15,007,816 during the second quarter of 2022 [3]. The adaptive evolution of more transmissible and virulent COVID-19 variants associated with different acquired mutations over time may lead to increased case counts, increased mortality rates, and reduced effectiveness of general COVID-19 vaccines [4]. The emergence of novel SARS-CoV-2 variants of concern over time may in part be attributed to both the innate error rate of SARS-CoV-2 replication and the different sources of host somatic mutagenesis that cause nonrandom patterns of mutation types and counts in viral genomes. Some known mechanisms that drive SARS-CoV-2 genomic evolution are commonly associated with host antiviral defenses, including the antiviral activity of (1) apolipoprotein B mRNA editing enzyme catalytic polypeptide-like (APOBEC) family causing C>U nucleotide substitutions, (2) reactive oxygen species (ROS) causing G>U substitutions, and (3) adenosine deaminase acting on RNA (ADAR) causing A>G and U>C substitutions [5,6]. Each host-specific antiviral defense mechanism may generate a unique set of mutation types and abundances over time, known as a mutational signature, which can be used to identify which and to what extent specific mutational processes contribute to all of the mutations observed in each genome [7]. Tracking the abundance of different substitution types over time can provide insights into the contribution of each mechanism of host antiviral defense to nucleotide changes in the SARS-CoV-2 genome [8].

Acquired mutations at specific sites in open reading frames [9] or near N6-methyladenosine (m6A) methylation sites [10] of the SARS-CoV-2 genome may confer advantages in viral transmissibility, host invasion, and reproduction, and modulate the severity of the clinical symptoms of COVID-19. Previous research has identified 8 m6A methylation sites as potential sites of negative regulation of viral infection [10]. Quantifying how the landscape of SARS-CoV-2 mutations in the SARS-CoV-2 genome changes over time can reveal how viral evolution may be associated with specific patterns of mutation burden, mutation types, and genomic locations of mutations, as well as different selection pressures. Moreover, the ratio of nonsynonymous to synonymous mutations in a given open reading frame, known as the dN/dS ratio, can be used to estimate the extent of and change in positive or negative selection of protein-coding regions across the SARS-CoV-2 genome over time.

The initial spread of COVID-19 in China was limited in part due to movement restrictions set in Wuhan in January 2020, followed by the rapid implementation of nonpharmaceutical interventions, including case isolation, physical distancing, wearing face masks, and contact tracing. These interventions were effective at reducing the seropositivity rate below the threshold for an epidemic [11]. However, international travel [12] and a reduction in nonpharmaceutical interventions due to lifting of regional mandates [12,13] have been associated with the spread of novel variants and waves of infections around the world since the initial COVID-19 outbreak in Wuhan. Similarly, changes in public health policy over the course of the pandemic may influence the transmission and emergence of new SARS-CoV-2 variants in Ontario, resulting in fluctuations in the prevalence of variants and genomic diversity reflected in the landscape of novel mutations observed in each epoch.

Previous SARS-CoV-2 genomic studies have characterized the landscape of mutation types and allele frequencies around the world, including but not limited to the United States [14], Qatar [15], the United Kingdom [16], Uruguay [17], and Canada [18]. Public health policies implemented in different countries to reduce COVID-19 transmission may in part influence the emergence of novel SARS-CoV-2 genetic variants [19] and mutation rates that can lead to the evolution of resistance to vaccines [20]. For instance, genomic epidemiology of the first 2 waves of SARS-CoV-2 in Canada revealed that the number of sublineages imported to Canada reduced by over 10-fold when restrictions of foreign nationals were implemented in March 2020. Thus, public health policies and travel restrictions can affect the number of opportunities for novel SARS-CoV-2 lineages to seed new outbreaks or challenge existing lineages. The change in SARS-CoV-2 genomic diversity in the viral population over time is attributed to both the impact of different public health policies and the introduction of novel international variants.

Similar to Canada, Spain employed a multistage nonuniform lockdown, while China employed more immediate and widespread lockdown procedures to reduce COVID-19 transmission compared with Canada [21,22]. Moreover, inconsistent provincial and territorial public health responses to the COVID-19 pandemic in Canada have been reported to be less effective at reducing COVID-19 transmission and SARS-CoV-2 genetic diversity in the Canadian population [23]. Effective epidemiological surveillance requires rigorous and reliable COVID-19 testing of case counts as well as genome sequencing to track the genomic mutations that are associated with increased transmission. A database of SARS-CoV-2 genomic sequences is necessary to track the genomic evolution of SARS-CoV-2 across the world. The Global Initiative on Sharing All Influenza Data (GISAID) is a database of influenza genomic sequences, and associated clinical and epidemiological data, which has facilitated the rapid public sharing of SARS-CoV-2–related data necessary for genomic surveillance during the COVID-19 pandemic [24].

Thus, characterizing the evolution of the SARS-CoV-2 mutational landscape in the Canadian province of Ontario as a

microcosm can reveal in part how Ontario-specific public health decisions and the introduction of novel variants during different time-based epochs may in part be associated with changes in the mutational landscape. This study is the first to characterize the change in the mutational landscape of 24,244 SARS-CoV-2 genomes sampled in Ontario from January 1, 2020, to December 31, 2021, at a large scale and over time, expanding on previous studies by increasing the sample size [25] and reporting the results of SARS-CoV-2 genomic surveillance over successive epochs of time [4]. In line with these studies, we also confirm previous reports of nonneutral selection in open reading frames, such as the spike and nucleocapsid open reading frames, as well as ORF3A in a larger curated data set of SARS-CoV-2 genomes sampled from the Ontario microcosm. Moreover, genomic surveillance is useful for quantifying the diversity of different variants over time as a measure of the effectiveness of public health policies to reduce transmission [26] and make rapid policy changes as well as predict conserved genomic regions undergoing negative selection as promising targets for vaccine design [27].

The surface-exposed SARS-CoV-2 spike glycoprotein, similar to other class I viral fusion proteins, such as influenza virus haemagglutinin and HIV envelope glycoprotein, regulates viral entry into host cells by changing conformation from a metastable unliganded state to a liganded stable state [28,29]. Previous studies have suggested that the spike protein mainly binds to the angiotensin II receptor to enter host cells [30]. In studies of viral challenge by SARS-CoV with similar class I viral fusion proteins as in SARS-CoV-2, polyclonal antibody responses targeting the spike protein were effective in inhibiting viral entry and decreasing viral load [31]. The design of antiviral agents that target the spike protein, a known region of positive selection that influences viral transmission, has been proposed to target the spike protein-angiotensin converting enzyme II binding interface, the auxiliary receptors involved in viral-cell fusion, or the specific epitopes of receptor binding motifs [32].

Recently, conserved pan-variant epitopes across the Alpha, Beta, Gamma, and Epsilon variant spike proteins have been successfully targeted using a neutralizing antibody fragment [33]. The changes in the patterns of mutation types, mutation genomic locations, and selection pressures of open reading frames associated with viral fitness between successive epochs suggest that the prediction of highly confident conserved epitopes and the resulting design of therapeutics with sustained efficacy may remain a significant challenge. The design of specific therapeutic approaches and vaccines for SARS-CoV-2 requires ongoing genomic surveillance to monitor their efficacy in combatting novel as well as increasingly resistant and transmissible SARS-CoV-2 variants [34]. As the COVID-19 pandemic continues, the genomic diversity of SARS-CoV-2 in Ontario may increase owing to the emergence of novel mutations acquired in part due to the host immune response and the increasing pool of infected hosts. Moreover, selection for emergent variants with mutations that confer a fitness advantage, characterized by increased immune escape and transmissibility, may further drive the genomic evolution of SARS-CoV-2 [35,36].

Here, we compared the novel changes in the proportions of single base substitution (SBS) types and the total burden of novel SBSs from SARS-CoV-2 genomes across 4 successive epochs. We tested if observed SBSs in open reading frames cluster near genes associated with increased viral fitness or known m6A sites across 4 successive epochs. Selection using the dN/dS ratios of different coding regions and the diversity of substitution types across the whole genome was compared between the 4 epochs to identify which coding regions were likely conserved. Finally, estimations of the rates of novel SBSs and the change in mutational fitness over time were used to estimate the evolution of the SARS-CoV-2 mutational landscape in Ontario.

This comprehensive study investigating the SARS-CoV-2 mutational landscape in the Ontario microcosm provides a foundation for research into simulating genomic epidemiology parameterized using both genomic and public health factors to inform public health decision-making as well as predict the activity of mutational processes and selection pressures that drive SARS-CoV-2 genomic evolution.

Methods

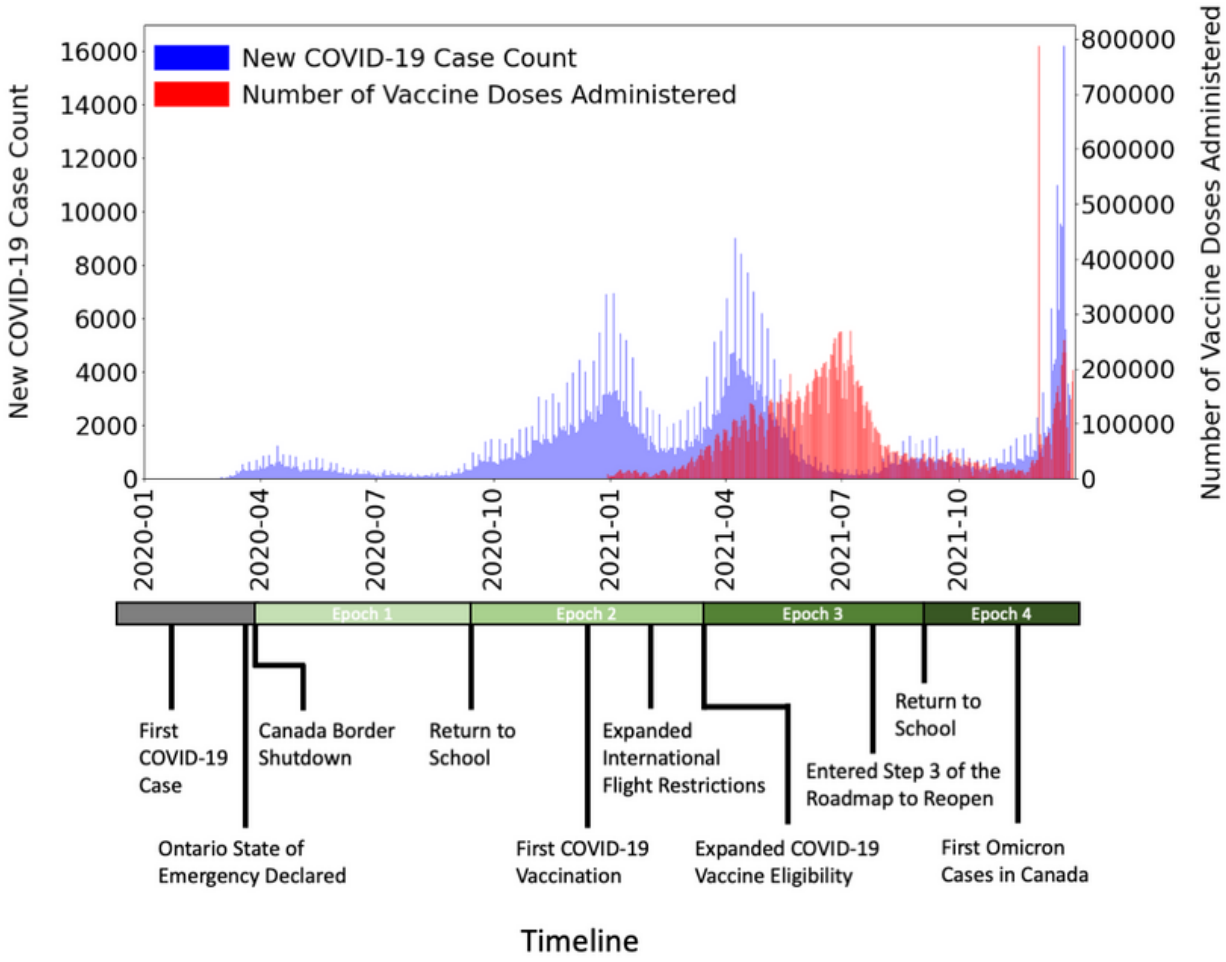
Data Collection

Data for 24,244 complete (>29,000 base pairs), high coverage (<1% N's), and nonduplicate SARS-CoV-2 genomic sequences in FASTA format sampled from January 1, 2020, to December 31, 2021, were retrieved from the GISAID database on January 1, 2022 [20]. Metadata associated with each genomic sequence, including sampling date and GISAID clade assignment [37], were also retrieved from the GISAID database. Duplicate sequences were removed if the FASTA header and associated metadata were the same. Whole-genome alignment of SARS-CoV-2 genomic sequences to the 29,903 base-pair reference genome (GenBank accession number MN908947.3) was conducted using MAFFT (Multiple Alignment with Fast Fourier Transform) version 7 with default parameters and keeping the alignment length between the sequence and the reference genome. Custom code was used to identify SBSs observed in aligned SARS-CoV-2 genomic sequences compared to the reference genome by iteratively checking if the base type at each position of the genomic sequence and reference genome was different, and if so, updating the data table with the observed SBS position, the reference base, and the alternate base for downstream analysis of mutational patterns. Custom code used in this study was implemented using Python version 3.8.0. The final position of each observed SBS is referenced based on the reference genome position. Minor allele frequency, defined as the proportion of the SARS-CoV-2 population sampled during a given epoch with the allele, was calculated for each minority allele at each base position.

Sequences were assigned to 4 different epochs based on time as follows: Epoch 1 from April 1, 2020, to August 31, 2020 (n=2256); Epoch 2 from September 1, 2020, to February 28, 2021 (n=4443); Epoch 3 from March 1, 2021, to August 31, 2021 (n=12,864); and Epoch 4 from September 1, 2021, to December 31, 2021 (n=4102) (Figure 1). The start of each epoch closely coincided with major public health events that may be

associated with reduced opportunities for viral transmission (eg, province-wide lockdowns with closure of nonessential businesses, introduction of international travel restrictions, and warmer weather) or increased viral transmission (eg, reopening of schools, lifting of international travel restrictions, and cooler weather), including the Ontario State of Emergency declaration in Epoch 1, the 2020 return to school in Epoch 2, expanded COVID-19 vaccine eligibility in Epoch 3, and the 2021 return to school in Epoch 4. Epoch-specific genomic mutations, defined as mutations in 1 epoch not observed in any previous epoch, were used for the downstream mutational pattern, clustering, selection, and diversity analyses. Mutations specific to Epoch 1 were determined by identifying mutations not previously observed in SARS-CoV-2 genomic sequences sampled from January 1, 2020, to March 31, 2020, in the collected data set (n=579).

Figure 1. Timeline of the 4 successive epoch time periods annotated with major public health events from April 1, 2020, to December 31, 2021. The public health events are relevant in the timeline given their impact on virus spread. The number of new COVID-19 case counts (blue) and number of vaccine doses administered (red) are plotted over time. The Ontario State of Emergency Declaration (March 17, 2020) instituted a province-wide lockdown imposing severe restrictions on human behavior with the intent to greatly limit the spread of infection. Re-entry from lockdown was phased in steps through the provincial Roadmap to Reopen. Significant public health policies associated with changes in viral transmission in the timeline include the Roadmap to Reopen and return to school.



Ethical Considerations

We confirm that all secondary analyses of research data from the GISAID database were performed in accordance with relevant usage guidelines and regulations. The genomic data retrieved from the GISAID database were deidentified and could not be linked to patients' identities [24].

Generation of SBS-96 Mutational Catalogues

Each SBS for each SARS-CoV-2 genomic sequence was assigned to 1 of 96 possible classes, where each class was defined by 6 base substitutions represented by the pyrimidine of the Watson-Crick base pair (C>A, C>G, C>U, U>A, U>C, and U>G) and the flanking 5′ and 3′ bases of the SBS that forms the local trinucleotide context [7]. For example, both the C>A pyrimidine substitution and G>U purine substitution were referred to by their pyrimidine base pair, C>A. For each SARS-CoV-2 genomic sequence, a mutational catalogue of the count of each of the 96 SBSs observed was generated.

To compare the mutational catalogues between multiple sequences in each epoch, the mean proportion of each of the 96 SBS types for each epoch was generated. First, the proportion of each of the 96 SBS types was calculated by dividing the count of each of the 96 SBS types by the total count of all SBS mutations observed in each SARS-CoV-2 genomic sequence. Second, the mean proportion of each of the 96 SBS types was calculated by summing the proportion of each of the 96 SBS

types across all SARS-CoV-2 genomic sequences assigned to the same epoch and dividing the sum by the number of SARS-CoV-2 genomic sequences in the epoch.

Somatic Mutation Clustering Near Open Reading Frames and m6A Methylation Sites

The distribution of sites where novel SBSs were observed along the whole genome was analyzed to detect the presence of SBS clusters in specific open reading frames or near known m6A methylation sites [10]. Open reading frames with observed clusters of epoch-specific mutations were identified if the positions of SBSs in the open reading frame were different from a random sample of 100 positions in the open reading frame using a 2-sample Kolmogorov-Smirnov test. Eight known m6A methylation sites of the SARS-CoV-2 genome were previously identified [10]. Observed clusters of epoch-specific mutations near m6A methylation sites were identified if the positions of SBSs within a window of ±1500 base pairs of one of the m6A methylation sites were different from a random sample of 100 positions in the same window using a 2-sample Kolmogorov-Smirnov test. The ±1500 base-pair window flanking each m6A methylation site was chosen based on a previous study that first identified the m6A methylation sites and reported clustering statistics near these sites using the same ±1500 base-pair window [10].

dN/dS Ratio of Open Reading Frames

To quantify the change in the selection pressures of each open reading frame over time, the mean dN/dS ratio of each open reading frame was compared between different epochs. First, we calculated the proportion of synonymous (pS) and nonsynonymous SBSs (pN) by dividing the count of pS and pN by the count of synonymous and nonsynonymous sites, respectively, for each open reading frame. Next, we estimated the count of synonymous (dS) and nonsynonymous (dN) substitutions per site using the Nei and Gojobori method shown in Equations 1 and 2 [38,39]. Null values of the dN/dS ratio, such as when zero nonsynonymous substitutions in an open reading frame were observed in a given SARS-CoV-2 genomic sequence, were excluded from the analyses.

Site-Specific Shannon Diversity Index

To compare the differences in genetic diversity at each genomic site over time, the Shannon diversity index was calculated at each genomic site for SARS-CoV-2 genomic sequences in each epoch. The mean diversity index value across the whole genome and site-specific diversity index values were compared between SARS-CoV-2 genomic sequences from different epochs [40].

Nextstrain: Annual Rate of Novel SBSs and Estimation of Mutational Fitness

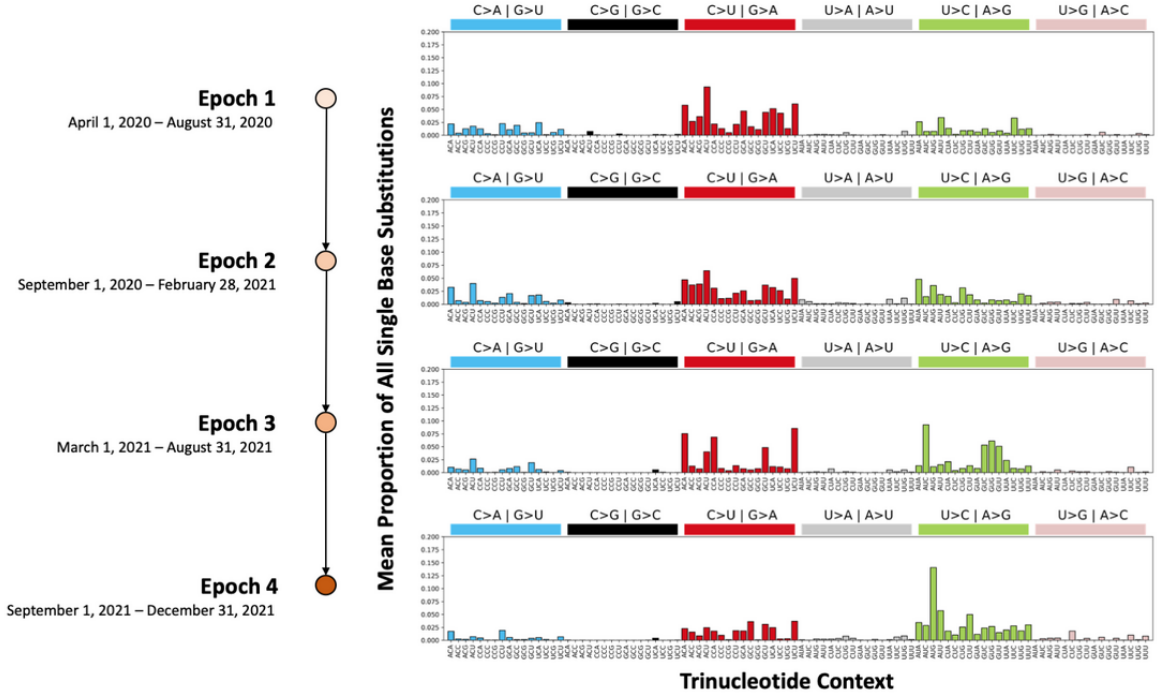
Out of 24,244 SARS-CoV-2 genomic sequences included in this study, 7398 SARS-CoV-2 genomic sequences with complete sampling dates, including day, month, and year, were used in the Nextstrain Augur 18.0.0 and Auspice 2.32.1 pipelines [41] for phylogenetic analysis and visualization, respectively [42]. The Nextstrain set of pipelines was used to estimate the rate of novel substitutions per year and the change in mutation fitness over time. Mutation fitness is defined as a metric that predicts viral reproduction and transmissibility based on the contribution of multiple somatic mutations that have been annotated to affect lineage growth, such as mutations that confer a fitness advantage by evading host immunity or decreasing generation time [43].

Results

Host Antiviral Defense Mechanisms are Associated With Nonrandom Biases in the Prevalence of Different SBS Types

Patterns of mutation types and their proportions commonly associated with mutagenesis due to host antiviral defense mechanisms were quantified by classifying the novel SBSs observed in each epoch into the SBS-96 classification scheme. Next, the SBS-96 mutation types and mean proportions were compared between different epochs (Figure 2). There were 1767 unique epoch-specific SBSs in Epoch 1, while there were 3573, 4822, and 2076 such SBSs in Epochs 2, 3, and 4, respectively (Multimedia Appendix 1).

Figure 2. The mean proportion of single base substitution (SBS)-96 mutation types of each of the 4 different epochs. The mean proportion shows, on average across all SARS-CoV-2 genomic sequences sampled during the epoch, what proportion each SBS type makes up out of all 96 possible SBS types. At least 80% of novel SBSs in each epoch were C>U, C>A, or U>C substitutions that have previously been attributed to the activity of apolipoprotein B mRNA editing enzyme catalytic polypeptide-like, reactive oxygen species, and adenosine deaminase acting on RNA, respectively.

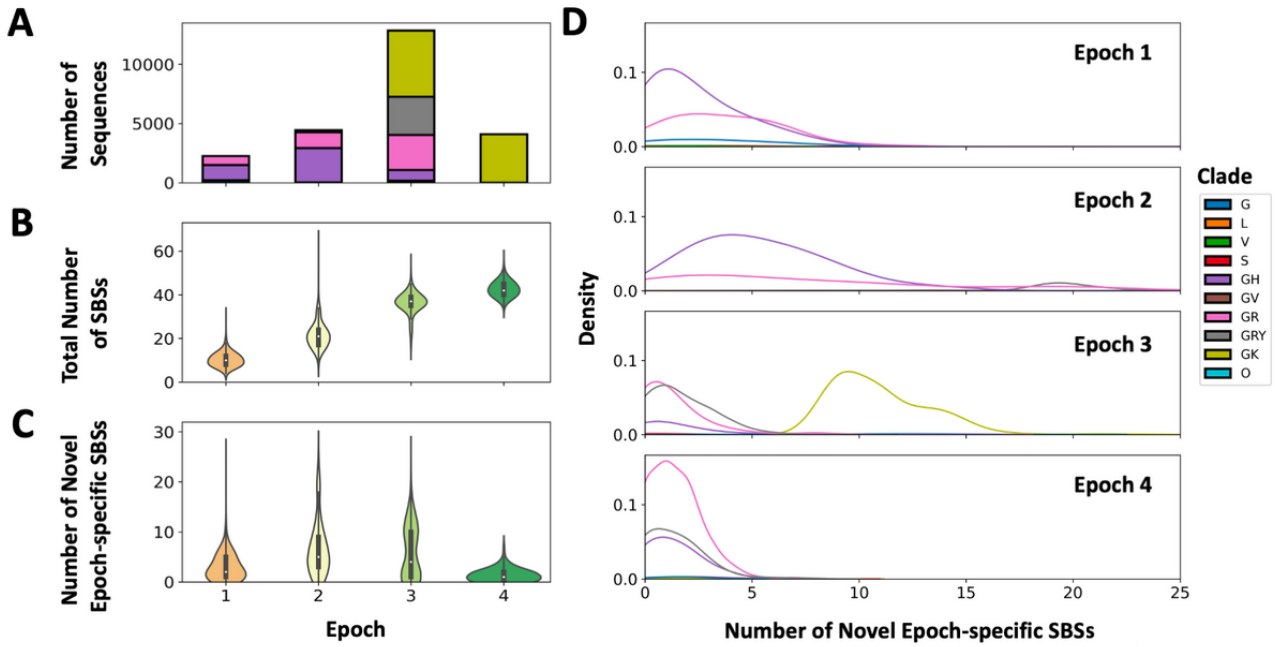


Across each of the 4 epochs, at least 80% of the novel SBSs were C>U, C>A, or U>C substitutions that have previously been attributed to the activity of APOBEC, ROS, and ADAR, respectively. Moreover, we observed that pyrimidine substitutions were more prevalent between each pair of purine and pyrimidine substitutions, constituting between 55.4% and 59.2% of all substitutions. The mean proportion of the C>A substitution type increased from 13.5% of all substitutions in Epoch 1 to 22.4% in Epoch 4. Conversely, the mean proportion of the C>T substitution type decreased from 67.9% in Epoch 1 to 48.4% in Epoch 4. The mean proportion of the T>C substitution type was relatively consistent among Epochs 1 to 3, ranging from 16.8% to 17.1% of all substitutions observed in each respective epoch. During Epoch 4, there was an increase in the mean proportion of the U>C substitution type at the AUG trinucleotide context compared to previous epochs. An increase in the mean proportion of the U>C substitution type to 19.2% of all substitutions was observed in Epoch 4 compared to Epochs 1 to 3. Visualizing each SARS-CoV-2 genomic sequence using a 3D uniform manifold approximation and projection plot of the SBS-96 mutation types and counts showed that genomic sequences sampled from the same epoch tended to uniquely cluster together (Multimedia Appendix 2).

Average Burden of Novel SBSs Differs Between Successive Epochs

To analyze the variation in the number of observed mutations over time, the average cumulative SBS mutational burden and average burden of novel SBSs first observed in each epoch were compared between different epochs. Comparing the clade composition of different epochs, we observed that clade G (Delta) and GR (Gamma) sequences made up the majority of sequences sampled in Epochs 1 and 2, clade GRY (Alpha) and GK sequences made up the majority of sequences sampled in Epoch 3, and clade GK (Delta) sequences were the sole majority of sequences sampled in Epoch 4 (Figure 3A). We observed that across successive epochs, the cumulative number of observed SBSs increased as expected, with a median cumulative mutational burden of 10 SBSs in Epoch 1, 21 SBSs in Epoch 2, 37 SBSs in Epoch 3, and 41 SBSs in Epoch 4 (Figure 3B).

Figure 3. Mutational burden of single base substitutions (SBSs). Global Initiative on Sharing All Influenza Data (GISAID) clade assignments for each SARS-CoV-2 genomic sequence are based on known marker mutations associated with 8 high-level phylogenetic groupings. (A) The number of SARS-CoV-2 genomic sequences colored by GISAID clade assignment and sampled from each of the 4 epochs. (B) The distribution of the total number of SBSs observed in SARS-CoV-2 genomic sequences sampled from each of the 4 epochs. (C) The count distribution of the number of novel epoch-specific SBSs first observed in SARS-CoV-2 genomic sequences sampled from each of the 4 epochs. (D) The density distribution of the number of novel epoch-specific SBSs first observed in SARS-CoV-2 genomic sequences colored by GISAID clade assignment and sampled from each of the 4 epochs.



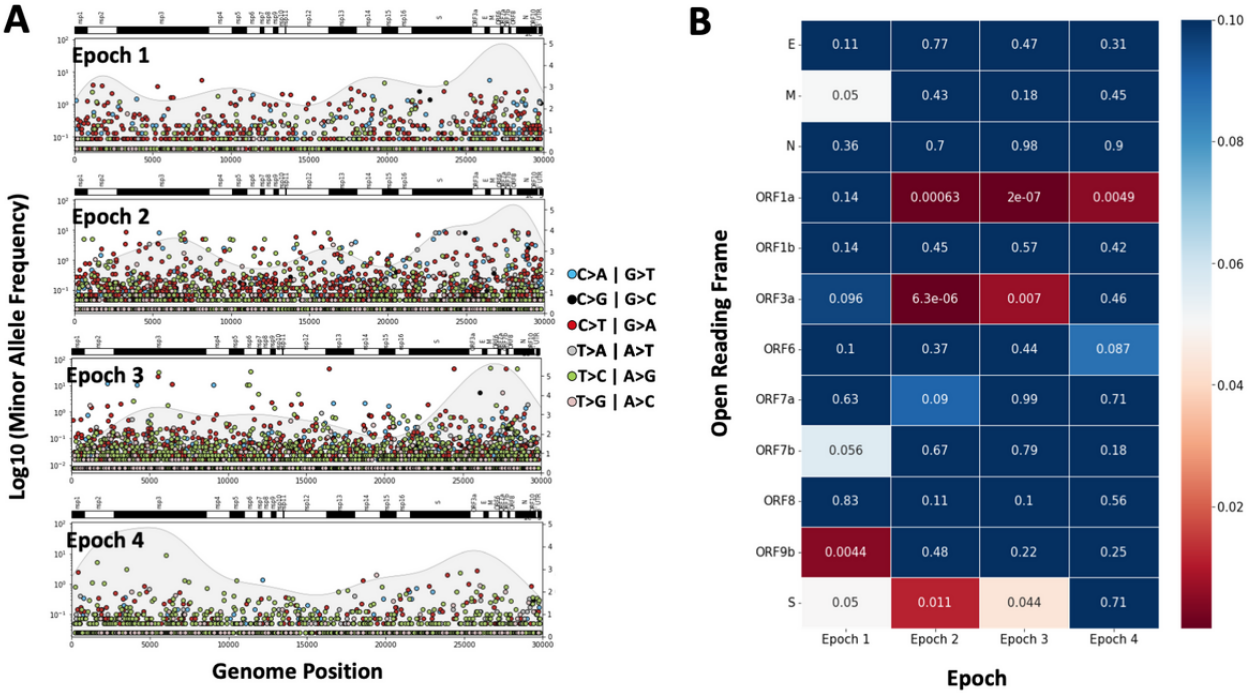
In addition, we observed the greatest maximum number of novel SBSs in Epoch 1 (maximum 26; median 2), followed by Epoch 2 (maximum 25; median 5), Epoch 3 (maximum 24; median 4), and Epoch 4 (maximum 8; median 1) (Figure 3C). The distribution of the mutational burden of novel SBSs in Epochs 2 and 3 was bimodal, with 2 different clade-specific populations of SARS-CoV-2 genomes with different mean counts of novel SBSs observed. In contrast, Epoch 1 and Epoch 4 showed a unimodal distribution of novel SBSs across all genomes sampled during the time period.

In Epoch 2, SARS-CoV-2 genomes from the GRY clade reported a higher burden of novel SBSs (median 19), while GH and GR clade genomes reported a lower burden of novel SBSs (GH median 5; GR median 5) (Figure 3D). In Epoch 3, SARS-CoV-2 genomes from the GR, GRY, and GH clades reported a lower burden of novel SBSs (GR median 1; GRY median 1; GH median 1), while GK clade genomes reported a higher burden of novel SBSs (median 10) (Figure 3D).

Clusters of Novel SBSs and Selection of Open Reading Frames may be Associated With Viral Fitness

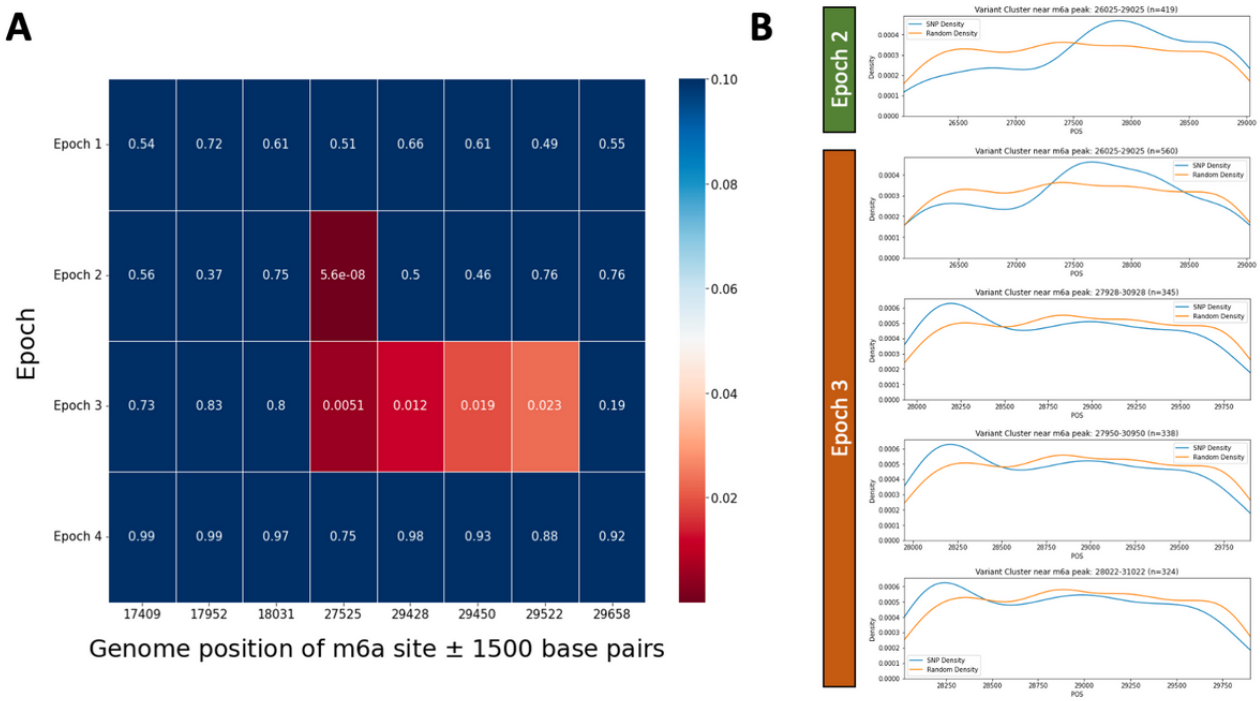
To identify open reading frames on SARS-CoV-2 genomes sampled in Ontario that may be associated with viral fitness, we observed the presence of clusters of novel SBSs in specific open reading frames compared between different epochs (Figure 4A). We consistently observed clusters of epoch-specific SBSs in the spike protein during Epochs 1, 2, and 3, but not during Epoch 4. Clusters of epoch-specific SBSs were observed in ORF1a during Epochs 2, 3, and 4. Likewise, clusters of epoch-specific SBSs were observed in ORF3a during Epochs 2 and 3, but not Epoch 4.

Figure 4. Minor allele frequency and clustering of novel epoch-specific single base substitutions (SBSs). (A) Scatterplot of the base position and the log-transformed minor allele frequency of novel epoch-specific SBSs first observed in each of the 4 epochs. The annotation of the SARS-CoV-2 genomic elements is shown above each plot. (B) Heatmap of the Kolmogorov-Smirnov test *P* value comparing the base positions of novel epoch-specific SBSs observed in each open reading frame of the SARS-CoV-2 genome and 100 randomly sampled base positions in the open reading frame, colored by *P* value (*P*<.05 is shown in red). Clusters of epoch-specific SBSs were observed in the spike protein, ORF1a, and ORF3a open reading frames during multiple epochs.



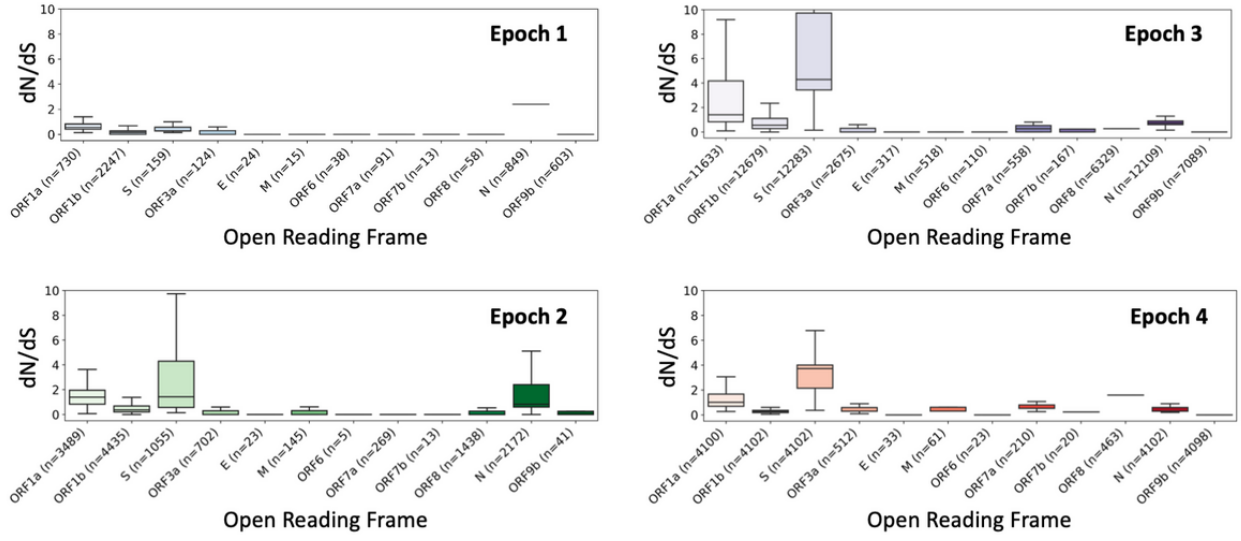
In addition, we tested for the existence of epoch-specific clusters of novel SBSs compared to a randomly sampled distribution of 100 substitution positions within a window of ±1500 base pairs of each of the 8 known m6A methylation sites using the 2-sample Kolmogorov-Smirnov test. In Epoch 2, there was 1 m6A methylation site at position 27525 in ORF6 of the SARS-CoV-2 reference genome with a significant cluster of novel SBSs (Figure 5A). In Epoch 3, there were 4 m6A methylation sites at position 27525 in ORF6 as well as positions 29428, 29450, and 29522 in ORF10 with a cluster of novel SBSs (Figure 5A). To confirm these findings at each of the 5 m6A methylation sites with potential clusters of novel SBSs within the window, we observed that the density of the positions of mutations did show nonrandom clustering with peaks of density that differed from the randomly sampled distribution (Figure 5B).

Figure 5. Clustering of novel epoch-specific single base substitutions (SBSs) within a ±1500 base-pair window around 8 N6-methyladenosine (m6a) methylation sites identified in the SARS-CoV-2 genome. (A) Heatmap of the Kolmogorov-Smirnov test *P* value comparing the base positions of the novel epoch-specific SBSs observed in each ±1500 base-pair window of 8 m6a methylation sites and 100 randomly sampled base positions in the same window, colored by *P* value (*P*<.05 is shown in red). (B) Density distribution of the novel epoch-specific SBSs (blue) and 100 randomly sampled base positions (orange) in the ±1500 base-pair window of 1 m6a methylation site in Epoch 2, and 4 m6a methylation sites in Epoch 3. The density distribution of SBSs is shown for 4 unique m6a methylation sites across 2 epochs since potential clusters of SBSs were detected within the ±1500 base-pair window around the m6a methylation sites.



To quantify the selection pressures on different open reading frames of the SARS-CoV-2 genome over time, we compared the dN/dS ratio for each open reading frame between different epochs. We observed that across successive epochs, there was an increase in the median dN/dS ratio from −0.5 in Epoch 1 to 0.5 in Epoch 4 for the spike protein open reading frame (Figure 6). This suggests an increase in positive selection for nonsynonymous mutations in the spike protein. Conversely, the median dN/dS metric of ORF1b was consistently below 0 across all epochs, indicating negative selection for nonsynonymous mutations in ORF1b (Figure 6).

Figure 6. Distribution of the dN/dS metric (ratio of nonsynonymous to synonymous mutations in a given open reading frame) calculated from the novel epoch-specific single base substitutions observed in SARS-CoV-2 genomic sequences sampled during the 4 different epochs, grouped by open reading frame. The numbers of SARS-CoV-2 genomic sequences where a nonnull dN/dS metric could be calculated for the open reading frame are shown.



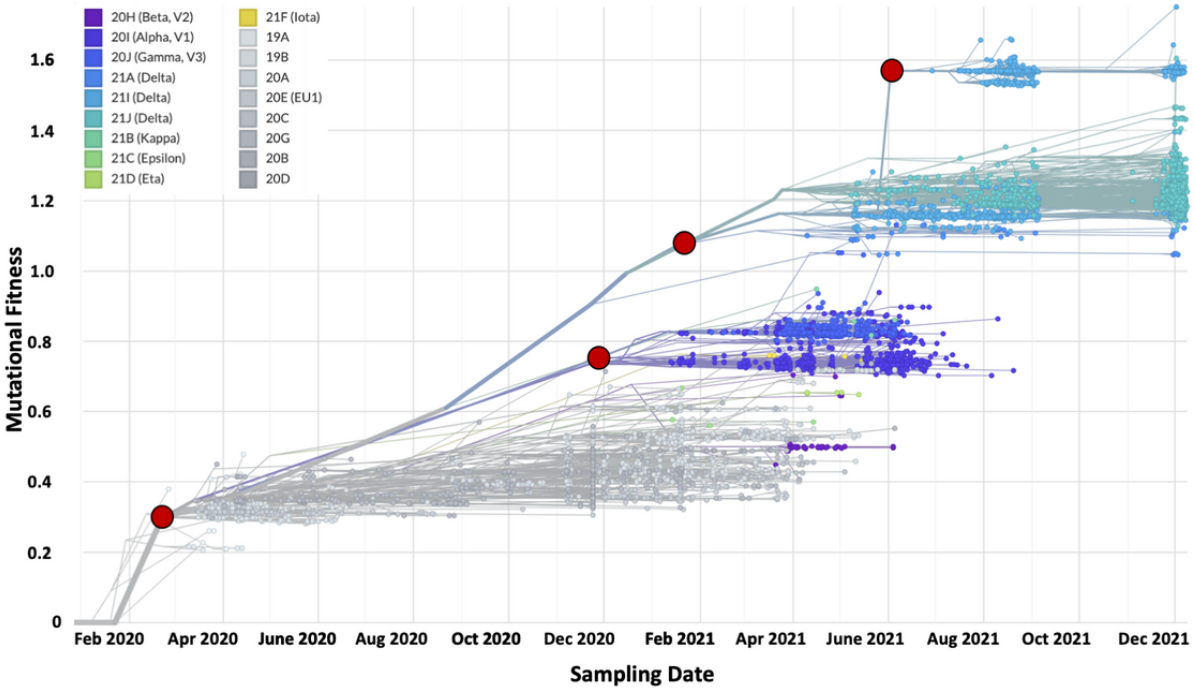
To characterize the diversity of different sites in the SARS-CoV-2 genome based on the different mutation types observed at the same site, we calculated the Shannon diversity index at each base position along the SARS-CoV-2 whole

genome (Multimedia Appendix 3). Then, we qualitatively compared the overall profile of Shannon diversity indices between different epochs to identify if diverse mutation types tended to be observed at specific open reading frames. We found that the most recent Epoch 4 showed a relatively broad diversity across the entire genome instead of peaks of high diversity near ORF6a, ORF7a, ORF7b, ORF8, and nucleocapsid protein observed in the previous 3 epochs. Moreover, Epoch 4 most closely resembled Epoch 1 in the shared peak of site-specific diversity in nonstructural protein (NSP) 2 of ORF1a.

Increases in SARS-CoV-2 Mutational Fitness Across Divergent Lineages Occur in Spurts

To estimate the rate of novel substitutions per year and potential divergence events indicated by changes in mutational fitness, we built a rooted maximum likelihood phylogenetic tree using the Nextstrain Augur pipeline with 7398 SARS-CoV-2 genomic sequences having complete sampling dates sampled in Ontario (Figure 7). Eighteen unique clades were represented in the tree, including several variants of concern, such as Delta (21A, 21I, 21J), Gamma (20J), Alpha (20I), and Beta (20H) (Figure 6). The tree shows that the COVID-19 pandemic in Ontario is caused by multiple different lineages of SARS-CoV-2 viruses, many of which are also concurrently observed in the same epoch. The positive linear rate of acquiring novel SBSs was observed to be around 23 mutations per year.

Figure 7. Mutational fitness of 7398 SARS-CoV-2 genomic sequences over time, colored by variant type identified using the Augur pipeline. Mutation fitness is a relative unitless metric that compares the estimated fitness of a given SARS-CoV-2 genomic sequence compared to the Wuhan reference genome, based on mutations annotated to be associated with a change in fitness. Higher mutational fitness suggests increased viral reproduction and transmissibility. Four high-level clusters of SARS-CoV-2 genomes were observed based on similarities in mutational fitness and SARS-CoV-2 variant composition. Clusters are denoted by a red marker at the earliest cluster-specific phylogenetic branch point.



The phylogenetic analysis revealed the timepoints of divergence of SARS-CoV-2 clades based on mutational fitness into 4 clusters. Cluster 1 was marked by an early divergence in mutational fitness in March 2020 of Epoch 1 based on branch length compared to the reference genome. Cluster 2 was predominantly composed of Alpha and Gamma variants of concern that diverged from Cluster 1 in December 2020 of Epoch 2. Following this, Cluster 3 was predominantly composed of Delta variants of concern that diverged from Cluster 2 in April 2021 of Epoch 3. Lastly, Cluster 4 was the most divergent cluster based on branch length and showed the highest mutational fitness, consisting of Delta variants of concern associated with the 20I clade observed from July 2021 in Epochs 3 and 4.

Discussion

In this study, we quantified the changes in the landscape of SBSs observed in SARS-CoV-2 genomic sequences sampled from 4 successive epochs in Ontario from March 2020 to December 2021 during the COVID-19 pandemic. Among sequences sampled during Epoch 1, we observed positive selection and a high median dN/dS ratio above 1 in the nucleocapsid protein–coding region. Moreover, we observed clustering of novel epoch-specific SBSs, positive selection, and higher base-specific Shannon diversity in the spike protein– and nucleocapsid protein–coding regions near the 3′ end of the SARS-CoV-2 genome. During Epoch 3, we observed that there was a bimodal distribution of the number of novel epoch-specific SBSs in the GK clade and a lower number of novel

epoch-specific SBSs in the GH, GV, and GRY clades. Similar to Epoch 2, there was a higher prevalence of novel SBSs in the spike protein and ORF3a-coding regions as well as positive selection of ORF1a and the spike protein–coding region. Finally, we uniquely noted that there was an increase in the mean proportion of the U>C substitution type up to 19.2% of all substitutions, especially at the AUG trinucleotide context, in Epoch 4 compared to previous epochs.

We observed that the open reading frames of SARS-CoV-2 viruses, including ORF1a and the spike protein, were increasingly impacted by positive selection over time, and this was in part consistent with mutations associated with host antiviral defenses. Future therapies can target these regions of positive selection since they affect the fitness of the SARS-CoV-2 virus [44]. Previous studies have similarly observed positive selection in the SARS-CoV-2 spike protein open reading frame [44,45] and ORF1a [46,47]. SARS-CoV-2 ORF1a and ORF1b genomic RNAs are translated and cleaved into 16 NSPs [48] associated with viral genomic replication as well as suppression of host immune response and gene expression [49]. For example, the NSP 7, 8, and 12 complex forms viral replicase machinery [50,51] and the NSP 10-16 complex is involved with capping viral mRNA transcripts and immune response evasion [52]. Due to the broad function of multiple NSPs translated from ORF1a for viral replication and the modulation of host immune response, future research to screen or design novel pharmaceutical drugs that target specific NSPs may show promise [53]. However, novel nonsynonymous mutations associated with positive selection affecting ORF1a and spike coding regions may not be conserved over time, so new iterations of pharmaceutical drugs, vaccines, or antigen tests targeting these regions may be required to maintain specificity and efficacy.

In Ontario, Epoch 1 marked the start of the pandemic from when the Canadian borders were shut down in late March 2020 to the return to school in October 2020. Similar to other regions around the world, such as the Baltimore-Washington area in the United States [54], the early measures designed to reduce local transmission of COVID-19 may have been compromised in part by the introduction of COVID-19 from national and international sources as well as the interconnectedness of the Ontario region. The high median dN/dS ratio above 1 in the nucleocapsid protein–coding region suggested that positive selection for nonsynonymous substitutions drove the early genomic evolution of the nucleocapsid protein and its function in viral genome packaging, despite being generally understood as a conserved region of the coronavirus genome [55]. Likewise, the high Shannon diversity index peaks near NSP1 and NSP2, which are involved in viral gene expression [56] and RNA binding [55], respectively, as well as the 3′ end of the SARS-CoV-2 genome, indicate a high prevalence of multiple different substitution types observed at a single genomic site. Since Epoch 1 is the first time period of the COVID-19 pandemic, the negative selection observed at all other open reading frames aside from the nucleocapsid protein–coding region may be attributed to the lag between the observation of a deleterious mutation and its subsequent selective removal from the gene pool [57]. Thus, SARS-CoV-2 genomic evolution associated with open reading frames involved in viral genome packaging and the resulting modulation of immune response were likely early events during Epoch 1 in the Ontario microcosm.

Epoch 2 was the time period from the September 2020 return to school and the introduction of the first COVID-19 vaccines in December 2020 to the expansion of COVID-19 vaccine eligibility to the general public in February 2021. Previous studies have suggested that COVID-19 vaccines likely play a role in initially reducing the genomic diversity of SARS-CoV-2 [58] and that early vaccine candidates targeting the spike protein involved in viral entry would likely be therapeutically effective against SARS-CoV-2 variants in Epoch 2 [59]. SARS-CoV-2 variants associated with clades GH and GR were the majority populations observed in both Epochs 1 and 2, but the introduction of variants associated with clade GRY (UK B.1.1.7 strain) was unique to Epoch 2 [60]. Similar to Epoch 1, clustering of novel epoch-specific SBSs and high base-specific Shannon diversity near the 3′ end of the SARS-CoV-2 genome was observed. Coupled with the positive selection observed in the spike protein– and nucleocapsid protein–coding regions near the 3′ end of the SARS-CoV-2 genome, variants observed in Epoch 2 showed increased genomic diversity in regions associated with viral attachment and entry as well as RNA genome packaging [61]. As time progressed in Epoch 2, selection pressure against SARS-CoV-2 variants with relatively decreased transmission, rate of replication, or immune defense evasion may have driven genomic evolution in favor of variants with novel mutations associated with increased fitness. Taken together, the introduction of vaccines targeting the spike protein is consistent with selection for novel mutations in the spike protein open reading frame introducing less virus susceptibility to vaccine-induced immune responses during Epoch 2 in the Ontario microcosm.

Epoch 3 spanned from March 2021 following the expansion of COVID-19 vaccine eligibility in Ontario to the September 2021 return to school. By August 5, 2021, 72% of Canadians had received 1 or more doses of a COVID-19 vaccine and 61% of Canadians were fully vaccinated with 2 doses, which were comparable to vaccination statistics observed in Ontario by this time [62]. Moreover, there was increasing demand on intensive care unit resources, and implementation of stay-at-home orders in Ontario as well as federal-mandated COVID-19 testing and 14-day quarantine of international air travelers to Canada were new measures to reduce COVID-19 case counts [62]. The emergence of GK clade SARS-CoV-2 variants with relatively high counts of novel epoch-specific SBSs and GH, GV, and GRY clade SARS-CoV-2 variants with relatively lower counts of novel SBSs comprised the bimodal distribution in the number of novel epoch-specific SBSs across all variants sampled in Epoch 3. Moreover, Epoch 3 variants were characterized by a relatively high prevalence of novel SBSs observed in the spike protein– and ORF3a-coding regions similar to Epoch 2. Likewise, both ORF1a and the spike protein–coding region showed a dN/dS ratio above 1, suggesting continued positive selection associated with nonsynonymous mutations in genomic regions involved with viral transmission and immune evasion. Moreover, the peak in the site-specific Shannon diversity index near the 3′ end of the SARS-CoV-2 genome indicates that the

nonsynonymous mutations in the spike protein–coding region are diverse in the observed alternate base at each mutation site. Population mixing among vaccinated and unvaccinated populations is consistent as a contributor to the increased infection rates among vaccinated individuals than expected in a fully vaccinated population [63]. Therefore, the observed positive selection associated with novel mutations in genomic regions with impact on SARS-CoV-2 transmission and immune evasion may have been driven in part by viral evolution in the human host population with variable immune responses due to the heterogeneity of individual vaccination statuses.

Epoch 4 was from the September 2021 return to school to the end of December 2021. This epoch followed the start of Step 3 of the Roadmap to Reopen, allowing for increased numbers of people at indoor and outdoor gatherings and increased capacity at nonessential venues with the requirement of face coverings in indoor settings [64]. The first case of the emergent SARS-CoV-2 Omicron variant was identified on November 22, 2021, during Epoch 4 [65]. Interestingly, we noted that there was an increase in the proportion of U>C substitutions in SARS-CoV-2 genomes sampled from Epoch 1 to Epoch 4. Our finding is consistent with a previous report of ADAR-induced editing of A>G and complementary T>C substitutions as mutations observed more commonly in genomes sampled from late 2020 onwards [66]. Moreover, the degree of RNA deamination has been reported as a potential determinant of viral immunogenicity and infectivity in emergent minor viral populations, warranting further investigation into RNA deamination as the main driver of SARS-CoV-2 genomic evolution [66]. Compared with Epoch 3, the lower median number of novel epoch-specific SBSs observed in Epoch 4 variants and the lower dN/dS ratio of the spike protein–coding region may be due to a combination of the shorter time period, a decrease in the mutation rate, and a reduction in positive selection pressure for nonsynonymous mutations. SBSs unique to Epoch 4 were clustered in ORF1a, namely NSP2- and NSP3-coding regions associated with viral replication, and were predominantly nonsynonymous mutations as evidenced by a dN/dS ratio above 1. These findings confirm a previous report of positive selection driving the genomic evolution of NSP2 and NSP3, and the high transmissibility of COVID-19 [67]. Compared to previous epochs, the marked increase in nonsynonymous mutations and relatively higher dN/dS ratio above 2 in the NSP8 region of variants sampled in Epoch 4 may be potential mechanisms for increasing stability of the SARS-CoV-2 viral replication and transcription complex [68]. Further research is required to determine the set of genomic mutations unique to Omicron variant genomes that may provide further insights into its mechanisms of increased transmissibility, immune evasion, and decreased pathogenicity [69]. The mutation fitness of SARS-CoV-2 genomic sequences sampled in Ontario was observed to increase in spurts over short time periods, likely coinciding with the introduction of novel SARS-CoV-2 variants with acquired genomic mutations that confer a fitness advantage [70]. Thus, future research predicting the functional impact of different sets of mutations on fitness could improve the surveillance of emergent SARS-CoV-2 variants for public health and inform the design of specific antiviral therapies [71].

This study used specific dates associated with the enactment of government public health policies to examine subsequent epoch-specific mutational patterns in SARS-CoV-2 genomic sequences. The nature of this study does not permit assessment of causation between government public health policies and mutational patterns due to the potential for other contributing and potentially confounding factors, including regional weather patterns, time lag between instantiation of public health policy and practical implementation, and the development of natural and vaccine immunity in the population. Future studies may identify specific time periods when these additional factors are impactful and assess their association with mutational patterns.

As the COVID-19 pandemic continues, there is a possibility of co-infection with other respiratory pathogens [72], as well as reinfection or co-infection with multiple different variants of SARS-CoV-2 [73]. Moreover, successive selective sweeps caused in part by both mutations that confer increased fitness [74,75] and homologous recombination may give rise to novel variants [76], such as the BA.2 Omicron variant. Thus, further research into the clinical impacts of co-infection and reinfection with different SARS-CoV-2 strains may highlight the interplay between genomic variation and COVID-19 symptoms and severity.

Another consideration is the impact of COVID-19 seasonality due to regional differences in environmental factors, including temperature and humidity, that can influence viral transmission, the diversity of SARS-CoV-2 variants selected based on tolerance to different environmental conditions, and the resulting case counts [77]. Interestingly, the annual winter influenza peaks in case counts reduced during 2020 and 2021, suggesting that COVID-19–related public health measures may impact the seasonal transmission of other respiratory viruses [78]. Thus, the development of public health policies should take into account the variation in the seasonality of COVID-19 and other respiratory viruses so that health care systems are prepared for fluctuations in case counts. Further surveillance of SARS-CoV-2 genomic variation and transmission patterns across Ontario can inform effective public health decision-making and serve as a microcosm of the COVID-19 pandemic as the case count of the novel Omicron variant increases. Future design of specific antiviral therapeutics should consider ongoing genomic surveillance as a tool to identify candidate targets [79,80].

In summary, we uniquely noted a bimodal distribution in epoch-specific counts of SBSs in sequences sampled during Epoch 3, where there was a high count observed in GK clade sequences and a lower count observed in GH, GV, and GRY clade sequences. Moreover, we uniquely observed an increase in the mean proportion of the U>C substitution type up to 19.2% of all substitutions, especially at the AUG trinucleotide context, in Epoch 4 compared to previous epochs. We confirmed previous reports of positive selection and clustering of SBSs near or within the ORF1a-, nucleocapsid protein–, and spike protein–coding regions.

We characterized the mutational profile of 24,244 SARS-CoV-2 genomic sequences sampled from January 1, 2020, to December 31, 2021, in Ontario, Canada. Our findings highlight how SARS-CoV-2 genomic sequences sampled from different epochs

harbor different patterns in mutational types, counts, and clusters that may be associated with differences in the transmissibility and virulence of SARS-CoV-2. Nonrandom biases in the abundance of different SBS types are consistent with the activity of host antiviral defense mechanisms and are in agreement with previous reports of the impact of host antiviral defense activity on the SARS-CoV-2 genome. Clusters of epoch-specific SBSs were observed in the spike protein, envelope protein, membrane protein, ORF3a, ORF6, and ORF7a open reading frames across all epochs, as well as near 4 unique m6A methylation sites during Epochs 2 and 3. Positive selection of the spike protein open reading frame, responsible for encoding the spike protein involved in viral entry, was observed. The estimated mutational fitness of SARS-CoV-2 genomic sequences increased in short-term spurts over time, suggesting that only a subset of somatic mutations confers a fitness advantage.

The microcosm of Ontario uniquely focuses on the evolution of the SARS-CoV-2 mutational profile associated with Ontario-specific public health events and policies. The mutational analysis of SARS-CoV-2 genomic sequences can in part reflect the impact of different public health policies during different epochs, such as the limiting of travel across Canadian borders in Epoch 1, and the impact on the genetic diversity of the SARS-CoV-2 viral population. This study of the mutational profile of SARS-CoV-2 in Ontario may serve as a model of the evolution of the SARS-CoV-2 mutational profile for comparison with other regions around the world that have implemented similar or different public health policies.

Further research of therapeutic agents designed to target conserved epitopes under negative selection, such as ORF1b, may shed light on how genomic surveillance can be a useful tool to inform the development of more effective antiviral therapies. Simulation tools used to project the evolution of SARS-CoV-2 genetic diversity due to somatic mutations or prediction models of COVID-19 waves may be parameterized using the mutational profiles and time points observed from SARS-CoV-2 sequences included in this study. To track the emergence of novel SARS-CoV-2 variants with reduced vaccine efficacy in the future, increased genomic surveillance is required, and it will inform public health policies associated with vaccine boosters as well as the implementation of nonpharmaceutical interventions such as wearing face masks and physical distancing.

Acknowledgments

We thank Connor Holmes, Joseph Butler, Amirhesam Afsharpour, Kate Deebrah, and Fatima Hassonali for reviewing the manuscript. We gratefully acknowledge all data contributors, that is, the authors and their originating laboratories responsible for obtaining the specimens as well as their submitting laboratories for generating the genetic sequences and metadata, and sharing via the Global Initiative on Sharing All Influenza Data, on which this research is based. This work was supported by Natural Science and Engineering Research Council of Canada Grants R3511A12 to KAH, R2824A01 to LK, and R2258A01 to SMS. This research was enabled in part by support provided by Compute Canada. The funders had no role in the preparation of the manuscript.

Data Availability

The open-source analysis code, sequence data sets, and output data sets are available at http://github.com/HillLab/SARS_CoV_2_Ontario_Genomic_Surveillance under the terms of the Creative Commons Attribution 4.0 International License.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The number of unique and shared single base substitutions observed in SARS-CoV-2 genomic sequences sampled from each of the 4 epochs.
[PNG File , 106 KB - bioinform_v3i1e42243_app1.png]

Multimedia Appendix 2
Three-dimensional uniform manifold approximation and projection plot of the single base substitution-96 mutation types and counts observed in 24,244 SARS-CoV-2 genomic sequences sampled in Ontario, colored by the epoch when they were sampled. Each SARS-CoV-2 genomic sequence is represented as 1 circle marker. Lighter shades of each marker color indicate that the SARS-CoV-2 genomic sequence was sampled at an earlier timepoint during the same epoch compared to sequences colored in darker shades of the same color. For example, a SARS-CoV-2 genomic sequence sampled during April 2020 in Epoch 1 would be shown in lighter red, while a SARS-CoV-2 genomic sequence sampled during August 2020 in Epoch 1 would be shown in darker red.
[PNG File , 183 KB - bioinform_v3i1e42243_app2.png]

Multimedia Appendix 3
Mean site-specific Shannon diversity index at each base position of the SARS-CoV-2 genome for SARS-CoV-2 genomic sequences sampled from each of the 4 epochs. Greater Shannon diversity index values at 1 base position suggest that the types of nucleotide bases are more diverse and evenly distributed between the different types compared to lower Shannon diversity values.
[PNG File , 60 KB - bioinform_v3i1e42243_app3.png]

References

1. COVID-19 epidemiology update: Key updates. Government of Canada. URL: https://health-infobase.canada.ca/covid-19/ [accessed 2022-07-21]
2. Ontario COVID-19 Data Tool. Public Health Ontario. URL: https://www.publichealthontario.ca/en/data-and-analysis/infectious-disease/covid-19-data-surveillance/covid-19-data-tool?tab=summary [accessed 2022-08-19]
3. Population estimates, quarterly. Statistics Canada. URL: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710000901 [accessed 2022-08-19]
4. Rochman ND, Wolf YI, Faure G, Mutz P, Zhang F, Koonin EV. Ongoing global and regional adaptive evolution of SARS-CoV-2. Proc Natl Acad Sci U S A 2021 Jul 20;118(29):e2104241118 [FREE Full text] [doi: 10.1073/pnas.2104241118] [Medline: 34292871]
5. Azgari C, Kilinc Z, Turhan B, Circi D, Adebali O. The Mutation Profile of SARS-CoV-2 Is Primarily Shaped by the Host Antiviral Defense. Viruses 2021 Mar 02;13(3):394 [FREE Full text] [doi: 10.3390/v13030394] [Medline: 33801257]
6. Graudenzi A, Maspero D, Angaroni F, Piazza R, Ramazzotti D. Mutational signatures and heterogeneous host response revealed via large-scale characterization of SARS-CoV-2 genomic diversity. iScience 2021 Feb 19;24(2):102116 [FREE Full text] [doi: 10.1016/j.isci.2021.102116] [Medline: 33532709]
7. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, PCAWG Mutational Signatures Working Group, PCAWG Consortium. The repertoire of mutational signatures in human cancer. Nature 2020 Feb 05;578(7793):94-101 [FREE Full text] [doi: 10.1038/s41586-020-1943-3] [Medline: 32025018]
8. Mourier T, Sadykov M, Carr MJ, Gonzalez G, Hall WW, Pain A. Host-directed editing of the SARS-CoV-2 genome. Biochem Biophys Res Commun 2021 Jan 29;538:35-39 [FREE Full text] [doi: 10.1016/j.bbrc.2020.10.092] [Medline: 33234239]
9. Khateeb J, Li Y, Zhang H. Emerging SARS-CoV-2 variants of concern and potential intervention approaches. Crit Care 2021 Jul 12;25(1):244 [FREE Full text] [doi: 10.1186/s13054-021-03662-x] [Medline: 34253247]
10. Liu J, Xu Y, Li K, Ye Q, Zhou H, Sun H, et al. The mA methylome of SARS-CoV-2 in host cells. Cell Res 2021 Apr 28;31(4):404-414 [FREE Full text] [doi: 10.1038/s41422-020-00465-7] [Medline: 33510385]
11. Zhang J, Litvinova M, Wang W, Wang Y, Deng X, Chen X, et al. Evolving epidemiology and transmission dynamics of coronavirus disease 2019 outside Hubei province, China: a descriptive and modelling study. Lancet Infect Dis 2020 Jul;20(7):793-802 [FREE Full text] [doi: 10.1016/S1473-3099(20)30230-9] [Medline: 32247326]
12. Yang J, Li J, Lai S, Ruktanonchai C, Xing W, Carioli A, et al. Uncovering two phases of early intercontinental COVID-19 transmission dynamics. J Travel Med 2020 Dec 23;27(8):taaa200 [FREE Full text] [doi: 10.1093/jtm/taaa200] [Medline: 33094347]
13. Han E, Tan MMJ, Turk E, Sridhar D, Leung GM, Shibuya K, et al. Lessons learnt from easing COVID-19 restrictions: an analysis of countries and regions in Asia Pacific and Europe. The Lancet 2020 Nov;396(10261):1525-1534. [doi: 10.1016/s0140-6736(20)32007-9]
14. Wang R, Chen J, Gao K, Hozumi Y, Yin C, Wei G. Analysis of SARS-CoV-2 mutations in the United States suggests presence of four substrains and novel variants. Commun Biol 2021 Feb 15;4(1):228 [FREE Full text] [doi: 10.1038/s42003-021-01754-6] [Medline: 33589648]
15. Benslimane FM, Al Khatib HA, Al-Jamal O, Albatesh D, Boughattas S, Ahmed AA, et al. One Year of SARS-CoV-2: Genomic Characterization of COVID-19 Outbreak in Qatar. Front Cell Infect Microbiol 2021 Nov 17;11:768883 [FREE Full text] [doi: 10.3389/fcimb.2021.768883] [Medline: 34869069]
16. Lythgoe KA, Hall M, Ferretti L, de Cesare M, MacIntyre-Cockett G, Trebes A, Oxford Virus Sequencing Analysis Group (OVSG), COVID-19 Genomics UK (COG-UK) Consortium, et al. SARS-CoV-2 within-host diversity and transmission. Science 2021 Apr 16;372(6539):eabg0821 [FREE Full text] [doi: 10.1126/science.abg0821] [Medline: 33688063]
17. Elizondo V, Harkins GW, Mabvakure B, Smidt S, Zappile P, Marier C, et al. SARS-CoV-2 genomic characterization and clinical manifestation of the COVID-19 outbreak in Uruguay. Emerg Microbes Infect 2021 Dec 15;10(1):51-65 [FREE Full text] [doi: 10.1080/22221751.2020.1863747] [Medline: 33306459]
18. Paré B, Rozendaal M, Morin S, Kaufmann L, Simpson SM, Poujol R, et al. Patient health records and whole viral genomes from an early SARS-CoV-2 outbreak in a Quebec hospital reveal features associated with favorable outcomes. PLoS One 2021 Dec 2;16(12):e0260714 [FREE Full text] [doi: 10.1371/journal.pone.0260714] [Medline: 34855869]
19. Grubaugh ND, Hodcroft EB, Fauver JR, Phelan AL, Cevik M. Public health actions to control new SARS-CoV-2 variants. Cell 2021 Mar 04;184(5):1127-1132 [FREE Full text] [doi: 10.1016/j.cell.2021.01.044] [Medline: 33581746]

20. Sjaarda CP, Guthrie JL, Mubareka S, Simpson JT, Hamelin B, Wong H, et al. Temporal Dynamics and Evolution of SARS-CoV-2 Demonstrate the Necessity of Ongoing Viral Genome Sequencing in Ontario, Canada. mSphere 2021 Jun 30;6(3):00011-21. [doi: 10.1128/msphere.00011-21]
21. Gao J, Zhang P. China's Public Health Policies in Response to COVID-19: From an "Authoritarian" Perspective. Front Public Health 2021 Dec 15;9:756677 [FREE Full text] [doi: 10.3389/fpubh.2021.756677] [Medline: 34976920]
22. David G, Rafael H, Ayelén RB, Inmaculada L, Amparo L, Marina P, et al. Perimeter confinements of basic health zones and COVID-19 incidence in Madrid, Spain. BMC Public Health 2022 Feb 03;22(1):216 [FREE Full text] [doi: 10.1186/s12889-022-12626-x] [Medline: 35109838]
23. Cyr A, Mondal P, Hansen G. An Inconsistent Canadian Provincial and Territorial Response During the Early COVID-19 Pandemic. Front Public Health 2021 Sep 27;9:708903 [FREE Full text] [doi: 10.3389/fpubh.2021.708903] [Medline: 34646800]
24. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill 2017 Mar 30;22(13):30494 [FREE Full text] [doi: 10.2807/1560-7917.ES.2017.22.13.30494] [Medline: 28382917]
25. Zhang M, Li L, Luo M, Liang B. Genomic characterization and evolution of SARS-CoV-2 of a Canadian population. PLoS One 2021 Mar 4;16(3):e0247799 [FREE Full text] [doi: 10.1371/journal.pone.0247799] [Medline: 33662015]
26. Romano CM, Melo FL. Genomic surveillance of SARS-CoV-2: A race against time. Lancet Reg Health Am 2021 Sep;1:100029 [FREE Full text] [doi: 10.1016/j.lana.2021.100029] [Medline: 34386792]
27. Kames J, Holcomb DD, Kimchi O, DiCuccio M, Hamasaki-Katagiri N, Wang T, et al. Sequence analysis of SARS-CoV-2 genome reveals features important for vaccine design. Sci Rep 2020 Sep 24;10(1):15643 [FREE Full text] [doi: 10.1038/s41598-020-72533-2] [Medline: 32973171]
28. Bosch BJ, van der Zee R, de Haan CAM, Rottier PJM. The coronavirus spike protein is a class I virus fusion protein: structural and functional characterization of the fusion core complex. J Virol 2003 Aug 15;77(16):8801-8811 [FREE Full text] [doi: 10.1128/jvi.77.16.8801-8811.2003] [Medline: 12885899]
29. Duan L, Zheng Q, Zhang H, Niu Y, Lou Y, Wang H. The SARS-CoV-2 Spike Glycoprotein Biosynthesis, Structure, Function, and Antigenicity: Implications for the Design of Spike-Based Vaccine Immunogens. Front Immunol 2020 Oct 7;11:576622 [FREE Full text] [doi: 10.3389/fimmu.2020.576622] [Medline: 33117378]
30. Davidson AM, Wysocki J, Batlle D. Interaction of SARS-CoV-2 and Other Coronavirus With ACE (Angiotensin-Converting Enzyme)-2 as Their Main Receptor. Hypertension 2020 Nov;76(5):1339-1349. [doi: 10.1161/hypertensionaha.120.15256]
31. Traggiai E, Becker S, Subbarao K, Kolesnikova L, Uematsu Y, Gismondo MR, et al. An efficient method to make human monoclonal antibodies from memory B cells: potent neutralization of SARS coronavirus. Nat Med 2004 Aug 11;10(8):871-875 [FREE Full text] [doi: 10.1038/nm1080] [Medline: 15247913]
32. Yan W, Zheng Y, Zeng X, He B, Cheng W. Structural biology of SARS-CoV-2: open the door for novel therapies. Signal Transduct Target Ther 2022 Jan 27;7(1):26 [FREE Full text] [doi: 10.1038/s41392-022-00884-5] [Medline: 35087058]
33. Walls AC, Park Y, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell 2020 Apr 16;181(2):281-292.e6 [FREE Full text] [doi: 10.1016/j.cell.2020.02.058] [Medline: 32155444]
34. Mannar D, Saville J, Sun Z, Zhu X, Marti M, Srivastava S, et al. SARS-CoV-2 variants of concern: spike protein mutational analysis and epitope for broad neutralization. Nat Commun 2022 Aug 18;13(1):4696 [FREE Full text] [doi: 10.1038/s41467-022-32262-8] [Medline: 35982054]
35. Robishaw JD, Alter SM, Solano JJ, Shih RD, DeMets DL, Maki DG, et al. Genomic surveillance to combat COVID-19: challenges and opportunities. The Lancet Microbe 2021 Sep;2(9):e481-e484. [doi: 10.1016/s2666-5247(21)00121-x]
36. Martin D, Weaver S, Tegally H, San E, Shank S, Wilkinson E, NGS-SA, COVID-19 Genomics UK (COG-UK), et al. The emergence and ongoing convergent evolution of the N501Y lineages coincides with a major global shift in the SARS-CoV-2 selective landscape. medRxiv 2021 Jul 25:2021 [FREE Full text] [doi: 10.1101/2021.02.23.21252268] [Medline: 33688681]
37. Clade and lineage nomenclature, March 2, 2021. GISAID. URL: https://www.gisaid.org/resources/statements-clarifications/clade-and-lineage-nomenclature-aids-in-genomic-epidemiology-of-active-hcov-19-viruses/ [accessed 2022-06-01]
38. Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 1986 Sep;3(5):418-426. [doi: 10.1093/oxfordjournals.molbev.a040410] [Medline: 3444411]
39. Maiti AK. Evolutionary shift from purifying selection towards divergent selection of SARS-CoV2 favors its invasion into multiple human organs. Virus Res 2022 May;313:198712 [FREE Full text] [doi: 10.1016/j.virusres.2022.198712] [Medline: 35176330]
40. Ghanchi NK, Nasir A, Masood KI, Abidi SH, Mahmood SF, Kanji A, et al. Higher entropy observed in SARS-CoV-2 genomes from the first COVID-19 wave in Pakistan. PLoS One 2021 Aug 31;16(8):e0256451 [FREE Full text] [doi: 10.1371/journal.pone.0256451] [Medline: 34464419]
41. Nextstrain. GitHub. URL: https://github.com/nextstrain [accessed 2022-12-10]
42. Huddleston J, Hadfield J, Sibley T, Lee J, Fay K, Ilcisin M, et al. Augur: a bioinformatics toolkit for phylogenetic analyses of human pathogens. J Open Source Softw 2021 Jan;6(57):2906 [FREE Full text] [doi: 10.21105/joss.02906] [Medline: 34189396]

43. Obermeyer F, Jankowiak M, Barkas N, Schaffner S, Pyle J, Yurkovetskiy L, et al. Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. medRxiv 2022 Feb 16:1208 [FREE Full text] [doi: 10.1101/2021.09.07.21263228] [Medline: 35194619]
44. Emam M, Oweda M, Antunes A, El-Hadidi M. Positive selection as a key player for SARS-CoV-2 pathogenicity: Insights into ORF1ab, S and E genes. Virus Res 2021 Sep;302:198472 [FREE Full text] [doi: 10.1016/j.virusres.2021.198472] [Medline: 34118359]
45. Yang H, Chen C, Wang J, Liao H, Yang C, Chen C, et al. Analysis of genomic distributions of SARS-CoV-2 reveals a dominant strain type with strong allelic associations. Proc Natl Acad Sci U S A 2020 Dec 01;117(48):30679-30686 [FREE Full text] [doi: 10.1073/pnas.2007840117] [Medline: 33184173]
46. Velazquez-Salinas L, Zarate S, Eberl S, Gladue DP, Novella I, Borca MV. Positive Selection of ORF1ab, ORF3a, and ORF8 Genes Drives the Early Evolutionary Trends of SARS-CoV-2 During the 2020 COVID-19 Pandemic. Front Microbiol 2020 Oct 23;11:550674 [FREE Full text] [doi: 10.3389/fmicb.2020.550674] [Medline: 33193132]
47. Yépez Y, Marcano-Ruiz M, Bezerra RS, Fam B, Ximenez JP, Silva Jr WA, et al. Evolutionary history of the SARS-CoV-2 Gamma variant of concern (P.1): a perfect storm. Genet. Mol. Biol 2022;45(1):e20210309. [doi: 10.1590/1678-4685-gmb-2021-0309]
48. Finkel Y, Mizrahi O, Nachshon A, Weingarten-Gabbay S, Morgenstern D, Yahalom-Ronen Y, et al. The coding capacity of SARS-CoV-2. Nature 2021 Jan 09;589(7840):125-130. [doi: 10.1038/s41586-020-2739-1] [Medline: 32906143]
49. Tsai P, Wang M, Yang D, Liang K, Chou S, Chiou S, et al. Genomic variance of Open Reading Frames (ORFs) and Spike protein in severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). J Chin Med Assoc 2020 Aug;83(8):725-732 [FREE Full text] [doi: 10.1097/JCMA.0000000000000387] [Medline: 32773643]
50. te Velthuis AJW, van den Worm SHE, Snijder E. The SARS-coronavirus nsp7+nsp8 complex is a unique multimeric RNA polymerase capable of both de novo initiation and primer extension. Nucleic Acids Res 2012 Feb;40(4):1737-1747 [FREE Full text] [doi: 10.1093/nar/gkr893] [Medline: 22039154]
51. Kirchdoerfer RN, Ward AB. Structure of the SARS-CoV nsp12 polymerase bound to nsp7 and nsp8 co-factors. Nat Commun 2019 May 28;10(1):2342 [FREE Full text] [doi: 10.1038/s41467-019-10280-3] [Medline: 31138817]
52. Rosas-Lemus M, Minasov G, Shuvalova L, Inniss N, Kiryukhina O, Wiersum G, et al. The crystal structure of nsp10-nsp16 heterodimer from SARS-CoV-2 in complex with S-adenosylmethionine. bioRxiv 2020 Apr 26:2020 [FREE Full text] [doi: 10.1101/2020.04.17.047498] [Medline: 32511376]
53. Raj R. Analysis of non-structural proteins, NSPs of SARS-CoV-2 as targets for computational drug designing. Biochem Biophys Rep 2021 Mar;25:100847 [FREE Full text] [doi: 10.1016/j.bbrep.2020.100847] [Medline: 33364445]
54. Thielen P, Wohl S, Mehoke T, Ramakrishnan S, Kirsche M, Falade-Nwulia O, et al. Genomic diversity of SARS-CoV-2 during early introduction into the Baltimore-Washington metropolitan area. JCI Insight 2021 Mar 22;6(6):e144350 [FREE Full text] [doi: 10.1172/jci.insight.144350] [Medline: 33749660]
55. Dutta NK, Mazumdar K, Gordy JT. The Nucleocapsid Protein of SARS–CoV-2: a Target for Vaccine Development. J Virol 2020 Jun 16;94(13):e00647-20. [doi: 10.1128/jvi.00647-20]
56. Mendez AS, Ly M, González-Sánchez AM, Hartenian E, Ingolia NT, Cate JH, et al. The N-terminal domain of SARS-CoV-2 nsp1 plays key roles in suppression of cellular gene expression and preservation of viral gene expression. Cell Rep 2021 Oct 19;37(3):109841 [FREE Full text] [doi: 10.1016/j.celrep.2021.109841] [Medline: 34624207]
57. Morales A, Rice A, Ho A, Mordstein C, Mühlhausen S, Watson S, et al. Causes and Consequences of Purifying Selection on SARS-CoV-2. Genome Biol Evol 2021 Oct 01;13(10):196 [FREE Full text] [doi: 10.1093/gbe/evab196] [Medline: 34427640]
58. Neisen M, Anand P, Silvert E, Suratekar R, Pawlowski C, Ghosh P, et al. COVID-19 vaccines dampen genomic diversity of SARS-CoV-2: Unvaccinated patients exhibit more antigenic mutational variance. MedRxiv. URL: https://www.medrxiv.org/content/10.1101/2021.07.01.21259833v1 [accessed 2022-12-09]
59. Dearlove B, Lewitus E, Bai H, Li Y, Reeves DB, Joyce MG, et al. A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. Proc Natl Acad Sci U S A 2020 Sep 22;117(38):23652-23662 [FREE Full text] [doi: 10.1073/pnas.2008281117] [Medline: 32868447]
60. Bano I, Sharif M, Alam S. Genetic drift in the genome of SARS COV-2 and its global health concern. J Med Virol 2022 Jan 23;94(1):88-98 [FREE Full text] [doi: 10.1002/jmv.27337] [Medline: 34524697]
61. Jungreis I, Sealfon R, Kellis M. SARS-CoV-2 gene content and COVID-19 mutation impact by comparing 44 Sarbecovirus genomes. Nat Commun 2021 May 11;12(1):2642 [FREE Full text] [doi: 10.1038/s41467-021-22905-7] [Medline: 33976134]
62. Detsky AS, Bogoch II. COVID-19 in Canada: Experience and Response to Waves 2 and 3. JAMA 2021 Sep 28;326(12):1145-1146. [doi: 10.1001/jama.2021.14797] [Medline: 34424275]
63. Fisman DN, Amoako A, Tuite AR. Impact of population mixing between vaccinated and unvaccinated subpopulations on infectious disease dynamics: implications for SARS-CoV-2 transmission. CMAJ 2022 Apr 25;194(16):E573-E580 [FREE Full text] [doi: 10.1503/cmaj.212105] [Medline: 35470204]
64. Ontario Moving to Step Three of Roadmap to Reopen on July 16. Government of Ontario. 2021. URL: https://news.ontario.ca/en/release/1000501/ontario-moving-to-step-three-of-roadmap-to-reopen-on-july-16 [accessed 2022-08-19]

65. Ulloa AC, Buchan SA, Daneman N, Brown KA. Estimates of SARS-CoV-2 Omicron Variant Severity in Ontario, Canada. JAMA 2022 Apr 05;327(13):1286-1288 [FREE Full text] [doi: 10.1001/jama.2022.2274] [Medline: 35175280]
66. Ringlander J, Fingal J, Kann H, Prakash K, Rydell G, Andersson M, et al. Impact of ADAR-induced editing of minor viral RNA populations on replication and transmission of SARS-CoV-2. Proc Natl Acad Sci U S A 2022 Feb 08;119(6):e2112663119 [FREE Full text] [doi: 10.1073/pnas.2112663119] [Medline: 35064076]
67. Angeletti S, Benvenuto D, Bianchi M, Giovanetti M, Pascarella S, Ciccozzi M. COVID-2019: The role of the nsp2 and nsp3 in its pathogenesis. J Med Virol 2020 Jun 28;92(6):584-588 [FREE Full text] [doi: 10.1002/jmv.25719] [Medline: 32083328]
68. Reshamwala SMS, Likhite V, Degani MS, Deb SS, Noronha SB. Mutations in SARS-CoV-2 nsp7 and nsp8 proteins and their predicted impact on replication/transcription complex structure. J Med Virol 2021 Jul 14;93(7):4616-4619 [FREE Full text] [doi: 10.1002/jmv.26791] [Medline: 33433004]
69. Jung C, Kmiec D, Koepke L, Zech F, Jacob T, Sparrer KMJ, et al. Omicron: What Makes the Latest SARS-CoV-2 Variant of Concern So Concerning? J Virol 2022 Mar 23;96(6):e0207721 [FREE Full text] [doi: 10.1128/jvi.02077-21] [Medline: 35225672]
70. Wei C, Shan K, Wang W, Zhang S, Huan Q, Qian W. Evidence for a mouse origin of the SARS-CoV-2 Omicron variant. J Genet Genomics 2021 Dec;48(12):1111-1121 [FREE Full text] [doi: 10.1016/j.jgg.2021.12.003] [Medline: 34954396]
71. Kumar A, Asghar A, Dwivedi P, Kumar G, Narayan RK, Jha RK, et al. A Bioinformatics Tool for Predicting Future COVID-19 Waves Based on a Retrospective Analysis of the Second Wave in India: Model Development Study. JMIR Bioinform Biotech 2022 Sep 22;3(1):e36860 [FREE Full text] [doi: 10.2196/36860] [Medline: 36193192]
72. Lin D, Liu L, Zhang M, Hu Y, Yang Q, Guo J, et al. Co-infections of SARS-CoV-2 with multiple common respiratory pathogens in infected patients. Sci China Life Sci 2020 Apr 05;63(4):606-609 [FREE Full text] [doi: 10.1007/s11427-020-1668-5] [Medline: 32170625]
73. Tillett RL, Sevinsky JR, Hartley PD, Kerwin H, Crawford N, Gorzalski A, et al. Genomic evidence for reinfection with SARS-CoV-2: a case study. The Lancet Infectious Diseases 2021 Jan;21(1):52-58. [doi: 10.1016/s1473-3099(20)30764-7]
74. Wang X, Hu M, Jin Y, Wang B, Zhao Y, Liang L, et al. Global Mutational Sweep of SARS-CoV-2: From Chaos to Order. Front Microbiol 2022 Feb 8;13:820919 [FREE Full text] [doi: 10.3389/fmicb.2022.820919] [Medline: 35211106]
75. Mastriani E, Rakov AV, Liu S. Isolating SARS-CoV-2 Strains From Countries in the Same Meridian: Genome Evolutionary Analysis. JMIR Bioinform Biotech 2021 Jan 22;2(1):e25995 [FREE Full text] [doi: 10.2196/25995] [Medline: 33497425]
76. Kozlakidis Z. Evidence for Recombination as an Evolutionary Mechanism in Coronaviruses: Is SARS-CoV-2 an Exception? Front Public Health 2022 Mar 17;10:859900 [FREE Full text] [doi: 10.3389/fpubh.2022.859900] [Medline: 35372203]
77. Liu X, Huang J, Li C, Zhao Y, Wang D, Huang Z, et al. The role of seasonality in the spread of COVID-19 pandemic. Environ Res 2021 Apr;195:110874 [FREE Full text] [doi: 10.1016/j.envres.2021.110874] [Medline: 33610582]
78. Maharaj AS, Parker J, Hopkins JP, Gournis E, Bogoch II, Rader B, et al. The effect of seasonal respiratory virus transmission on syndromic surveillance for COVID-19 in Ontario, Canada. The Lancet Infectious Diseases 2021 May;21(5):593-594. [doi: 10.1016/s1473-3099(21)00151-1]
79. Gupta E, Mishra RK, Kumar Niraj RR. Identification of Potential Vaccine Candidates Against SARS-CoV-2 to Fight COVID-19: Reverse Vaccinology Approach. JMIR Bioinform Biotech 2022 Apr 26;3(1):e32401 [FREE Full text] [doi: 10.2196/32401] [Medline: 35506029]
80. Math RK, Mudennavar N, Javaregowda PK, Savanur A. In Silico Comparative Analysis of the Functional, Structural, and Evolutionary Properties of SARS-CoV-2 Variant Spike Proteins. JMIR Bioinform Biotech 2022 May 30;3(1):e37391 [FREE Full text] [doi: 10.2196/37391] [Medline: 35669291]

Abbreviations

ADAR: adenosine deaminase acting on RNA
APOBEC: apolipoprotein B mRNA editing enzyme catalytic polypeptide-like
GISAID: Global Initiative on Sharing All Influenza Data
m6A: N6-methyladenosine
NSP: nonstructural protein
ROS: reactive oxygen species
SBS: single base substitution

Edited by A Mavragani; submitted 02.09.22; peer-reviewed by A Banerjee, J Yang, J Walsh; comments to author 07.10.22; revised version received 29.11.22; accepted 05.12.22; published 22.12.22.

Please cite as:
Chen D, Randhawa GS, Soltysiak MPM, de Souza CPE, Kari L, Singh SM, Hill KA
Mutational Patterns Observed in SARS-CoV-2 Genomes Sampled From Successive Epochs Delimited by Major Public Health Events in Ontario, Canada: Genomic Surveillance Study
JMIR Bioinform Biotech 2022;3(1):e42243
URL: https://bioinform.jmir.org/2022/1/e42243
doi:10.2196/42243
PMID:

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