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 5 (2024) ISSN 2563-3570 Editor in Chief: Ece D. Uzun, MS, PhD, FAMIA

Contents

Reviews

Viewpoints

Corrigenda and Addenda

Editorial

Original Papers

Review

Internet-Based Abnormal Chromosomal Diagnosis During Pregnancy Using a Noninvasive Innovative Approach to Detecting Chromosomal Abnormalities in the Fetus: Scoping Review

Mega Obukohwo Sr Oyovwi[1*], PhD; Ejiro Peggy Ohwin[2*], PhD; Rume Arientare Rotu[3*], MSc; Temitope Gideon Olowe[4], PhD

[1]Department of Physiology, Adeleke University, Ede, Nigeria

[2]Department of Human Physiology, Faculty of Basic Medical Science, Delta State University, Abraka, Nigeria

[3]Department of Physiology, University of Ibadan, Ibadan, Nigeria

[4]Department of Obstetrics & Gynaecology, University of Medical Sciences, Ondo, Nigeria

[*]these authors contributed equally

Corresponding Author:
Mega Obukohwo Sr Oyovwi, PhD
Department of Physiology
Adeleke University
Ede, Osun State
Ede, 33105
Nigeria
Phone: 234 8066096369
Email: megalect@gmail.com

Abstract

Background: Chromosomal abnormalities are genetic disorders caused by chromosome errors, leading to developmental delays, birth defects, and miscarriages. Currently, invasive procedures such as amniocentesis or chorionic villus sampling are mostly used, which carry a risk of miscarriage. This has led to the need for a noninvasive and innovative approach to detect and prevent chromosomal abnormalities during pregnancy.

Objective: This review aims to describe and appraise the potential of internet-based abnormal chromosomal preventive measures as a noninvasive approach to detecting and preventing chromosomal abnormalities during pregnancy.

Methods: A thorough review of existing literature and research on chromosomal abnormalities and noninvasive approaches to prenatal diagnosis and therapy was conducted. Electronic databases such as PubMed, Google Scholar, ScienceDirect, CENTRAL, CINAHL, Embase, OVID MEDLINE, OVID PsycINFO, Scopus, ACM, and IEEE Xplore were searched for relevant studies and articles published in the last 5 years. The keywords used included *chromosomal abnormalities*, *prenatal diagnosis*, *noninvasive*, and *internet-based*, and *diagnosis*.

Results: The review of literature revealed that internet-based abnormal chromosomal diagnosis is a potential noninvasive approach to detecting and preventing chromosomal abnormalities during pregnancy. This innovative approach involves the use of advanced technology, including high-resolution ultrasound, cell-free DNA testing, and bioinformatics, to analyze fetal DNA from maternal blood samples. It allows early detection of chromosomal abnormalities, enabling timely interventions and treatment to prevent adverse outcomes. Furthermore, with the advancement of technology, internet-based abnormal chromosomal diagnosis has emerged as a safe alternative with benefits including its cost-effectiveness, increased accessibility and convenience, potential for earlier detection and intervention, and ethical considerations.

Conclusions: Internet-based abnormal chromosomal diagnosis has the potential to revolutionize prenatal care by offering a safe and noninvasive alternative to invasive procedures. It has the potential to improve the detection of chromosomal abnormalities, leading to better pregnancy outcomes and reduced risk of miscarriage. Further research and development in this field is needed to make this approach more accessible and affordable for pregnant women.

(JMIR Bioinform Biotech 2024;5:e58439) doi:10.2196/58439

KEYWORDS

internet-based; abnormal chromosomal diagnosis; pregnancy; noninvasive; innovative approach; detecting; preventing; chromosomal abnormalities; fetus

Introduction

Background

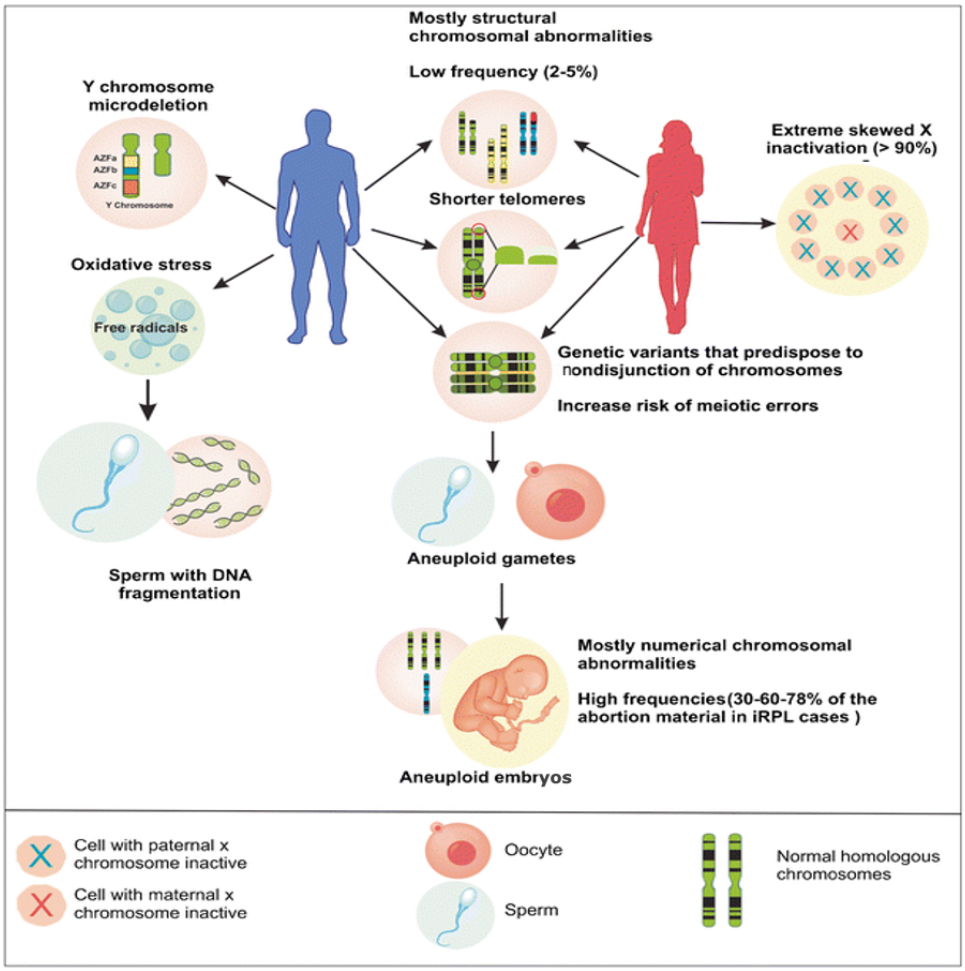
Prenatal diagnosis of chromosomal abnormalities is an important part of prenatal care. Chromosomal abnormalities are the major cause of pregnancy complications, including miscarriage, stillbirth, and birth defects [1]. Understanding the prevalence and impact of commonly diagnosed chromosomal abnormalities in pregnancy is essential for providing accurate genetic counseling and appropriate prenatal care. Traditionally, prenatal diagnosis has been performed using invasive methods such as amniocentesis and chorionic villus sampling. However, these methods are associated with a small risk of miscarriage [2,3]. In recent years, noninvasive prenatal testing (NIPT) has emerged as a safe and effective alternative to invasive methods. NIPT is based on the analysis of cell-free DNA (cfDNA) in the maternal blood [3]. cfDNA is released into the maternal blood by the placenta and contains genetic material from both the mother and the fetus. This advent of NIPT has revolutionized prenatal diagnosis [3]. While NIPT has emerged as a powerful tool for detecting common chromosomal abnormalities such as Down syndrome, its accessibility and potential for broader application through internet-based platforms remain relatively unexplored. This review focuses on understanding the feasibility, benefits, and challenges of using internet-based technologies to deliver NIPT services effectively. Internet-based NIPT presents a compelling opportunity to overcome barriers associated with traditional prenatal diagnostics [4]. Web-based platforms can extend NIPT services to geographically remote areas and underserved populations, bridging health care disparities. Web-based platforms offer flexible scheduling and internet consultations, reducing the need for multiple clinic visits, especially beneficial for working mothers [5]. Internet-based platforms can potentially streamline administrative processes and reduce operational costs, making NIPT more affordable for a wider population [5]. This review aims to provide a comprehensive overview of the current state of internet-based NIPT, discussing its technical feasibility, ethical considerations, and potential impact on prenatal care. Notwithstanding, current prenatal chromosomal diagnosis methods have several limitations. They are invasive, expensive, and can cause anxiety in pregnant women. Therefore, there is a need for internet-based abnormal chromosomal diagnosis, a noninvasive, cost-effective, and anxiety-reducing method for chromosomal abnormality detection.

Internet-based methods for prenatal diagnosis of chromosomal abnormalities are becoming increasingly popular. These methods allow pregnant women to access information and support from health care professionals and other parents who have experienced similar challenges [6]. There are a number of different internet-based methods for prenatal diagnosis, including web-based genetic counseling, online support groups, and web-based prenatal testing [7-9]. Internet-based abnormal chromosomal diagnosis during pregnancy is a noninvasive and innovative approach to detecting chromosomal abnormalities in fetuses, offering several advantages over traditional invasive procedures [5]. This review aims to provide a comprehensive overview of this emerging technique, highlighting its benefits, limitations, and implications for prenatal care. Moreover, internet-based abnormal chromosomal diagnosis during pregnancy aims to address these limitations by using advanced computational techniques to analyze fetal genetic material obtained through noninvasive methods, such as maternal blood samples. This approach offers a safe and convenient alternative to traditional invasive procedures. This review aims to provide a comprehensive understanding of internet-based abnormal chromosomal diagnosis during pregnancy. By exploring this emerging technology, we can contribute to improving the safety, accessibility, and effectiveness of prenatal chromosomal abnormality detection.

Basics of Chromosomal Abnormalities

Chromosomal abnormalities involve changes in the number or structure of chromosomes, which contain genetic information determining physical traits [10]. These can lead to health issues such as developmental delays, birth defects, and genetic disorders (Figure 1 [10]). There are 2 main types of abnormalities: numerical and structural [11]. Numerical abnormalities involve whole chromosome loss, while structural abnormalities involve chromosome structure changes [12] (Textbox 1).

Figure 1. Chromosomal abnormalities linked to repeated miscarriages. There is evidence linking chromosomal anomalies to repeated miscarriages at the parent, gamete, and fetal levels. Abnormalities in numbers and structure provide the most compelling evidence of a connection to the illness. iRPL: idiopathic recurrent pregnancy loss.



Textbox 1. Numerical and structural abnormalities.

Numerical abnormalities

- Aneuploidy: aneuploidy is a condition where there is an abnormal number of chromosomes in the cells. The most common examples of aneuploidy include trisomy (an extra copy) and monosomy (a missing copy) of a chromosome. The most well-known example of aneuploidy is Down syndrome, which is caused by an extra copy of chromosome 21 [13].
- Polyploidy: polyploidy refers to the presence of >2 sets of chromosomes in a cell. It is relatively rare in humans, but it can lead to severe birth defects and developmental delays [14].

Structural abnormalities

- Deletion: a deletion occurs when a part of a chromosome is missing or deleted. This can result in the loss of essential genetic information and can lead to various health issues, including physical and cognitive disabilities [15].
- Duplication: duplication is when a section of a chromosome is duplicated, resulting in an extra copy of genetic material [16]. Duplication can lead to developmental delays, intellectual disabilities, and other health problems.
- Translocation: translocation occurs when a part of one chromosome breaks off and attaches to another chromosome [17]. This can result in a rearrangement of genetic material and can cause various health issues depending on the genes involved.

Causes and Risk Factors

Chromosomal abnormalities can occur due to various causes [18], including (1) genetic inheritance: some chromosomal abnormalities can be inherited from one or both parents, such as Down syndrome, which is caused by an extra copy of chromosome 21 inherited from the mother or father; (2) errors in cell division: chromosomal abnormalities can also occur during the process of cell division, for example, an error in the division of sex cells (eggs and sperm) can result in an embryo with an abnormal number of chromosomes; and (3) exposure to environmental factors: exposure to certain environmental factors, such as radiation, chemicals, and toxins, can increase the risk of chromosomal abnormalities in pregnancy.

Methods

Search Strategy

To comprehensively explore the landscape of chromosomal abnormalities and noninvasive prenatal diagnosis and therapy, a thorough literature review was undertaken. This review encompassed a wide range of electronic databases including PubMed, Google Scholar, ScienceDirect, CENTRAL, CINAHL, Embase, OVID MEDLINE, OVID PsycINFO, Scopus, ACM, and IEEE Xplore (Multimedia Appendix 1). The search focused on studies and articles published within the last 5 years, using keywords such as *chromosomal abnormalities*, *prenatal diagnosis*, *noninvasive*, and *internet-based approach*. This multifaceted search strategy aimed to capture the most relevant and current research on this topic. The search was further refined by applying filters for language (English), publication type (journal articles, systematic reviews, and meta-analyses), and time frame (from database inception to the present). In addition, reference lists of retrieved articles and relevant textbooks were manually inspected for additional pertinent studies. This comprehensive search strategy ensured the identification of a wide range of literature exploring the internet-based abnormal chromosomal diagnosis during pregnancy: a noninvasive innovative approach to detecting chromosomal abnormalities in the fetus, thus providing a robust foundation for this review.

Inclusion and Exclusion Criteria

The inclusion criteria for this review were studies that focused on chromosomal abnormalities and internet-based diagnosis. Studies that used an internet-based approach to detect and quantify chromosomal abnormalities in the fetus were also included. The exclusion criteria were studies that did not focus on chromosomal abnormalities or did not have a specific focus on internet-based approaches. Studies that were not published in the English language or were published before 2000 were also excluded.

Ethics Approval

This review was conducted in accordance with the guidelines and approval of the Research, Ethics, and Grants Committee of the Faculty of Basic Medical Sciences, Adeleke University, Ede, Nigeria.

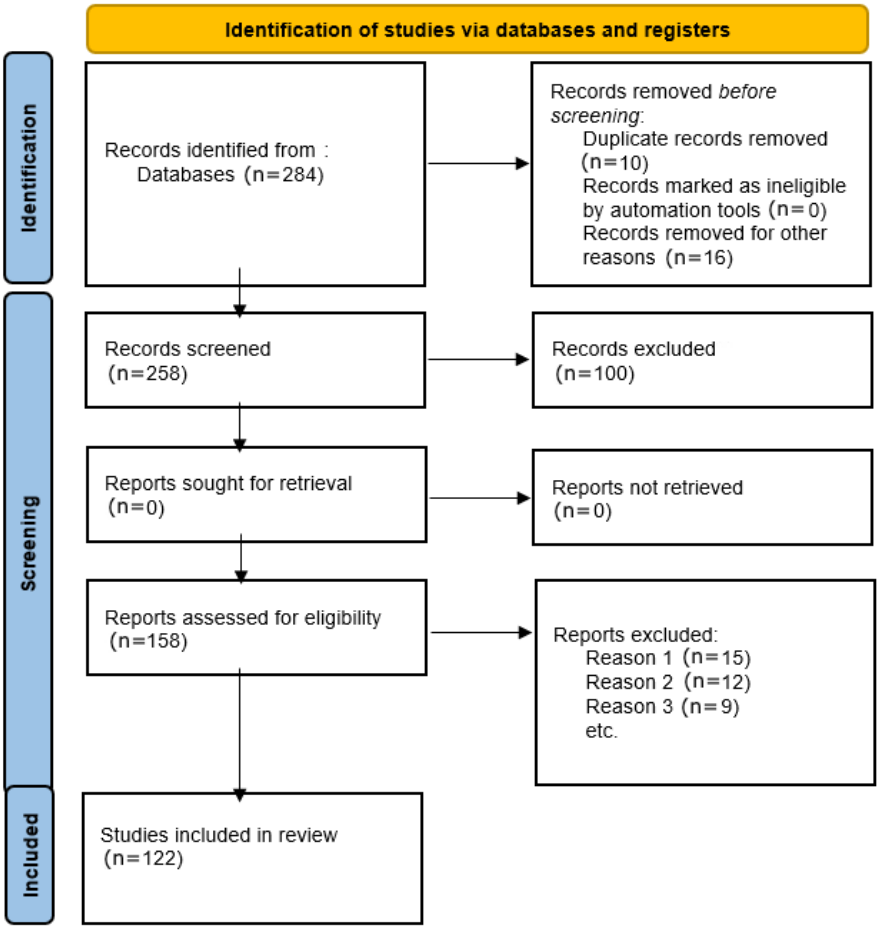
Results and Discussion

Internet-Based Abnormal Chromosomal Diagnosis

Overview

Figure 2 shows an overview of the included studies. The rapid advancements in technology have transformed the field of medicine, including the way we diagnose and treat diseases. One such groundbreaking approach is internet-based abnormal chromosomal diagnosis. This approach uses the internet to provide genetic counseling and testing for individuals with abnormal chromosomal conditions. Here, we discuss the definition and explanation of this approach as well as how it works through genetic counseling and testing via web-based platforms and kits.

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart. Reason 1 (n=15): studies not providing data specifically related to noninvasive methods for detecting chromosomal abnormalities in the fetus. Reason 2 (n=12): reports that were not focused on internet-based or telehealth approaches in delivering prenatal diagnosis or genetic counseling. Reason 3 (n=9): articles lacking peer-reviewed status, including nonscientific articles, opinion pieces, or conference abstracts that did not meet rigorous research standards.



Definition and Explanation of the Approach

Internet-based abnormal chromosomal diagnosis is a novel approach in which individuals with abnormal chromosomal conditions can receive genetic counseling and testing through web-based platforms [19]. This approach uses the internet to provide individuals with access to genetic counseling and testing services without the need to physically visit a health care facility [19]. Genetic counselors, which are health care professionals trained in genetics, use web-based platforms to communicate with patients and provide them with information about their condition, potential risks, and available treatment options. This approach also offers genetic testing kits that can be used at home to collect samples, which are then sent to a laboratory for analysis.

Decoding DNA: A Guide to Web-Based Genetic Testing and Counseling

Genetic Counseling Through Web-Based Platforms

The advent of internet-based technologies has revolutionized the delivery of health care services, including genetic counseling. Scientific research has explored the effectiveness and benefits of genetic counseling through web-based platforms, offering valuable insights into the transformative potential of this approach [20]. Web-based genetic counseling involves using virtual communication tools, such as videoconferencing and secure messaging, to provide genetic risk assessment, education, and support to individuals and families seeking genetic information [21].

Studies have demonstrated the efficacy of web-based genetic counseling in providing accurate and comprehensive genetic information [22]. Through secure and Health Insurance Portability and Accountability Act–compliant platforms, genetic counselors effectively collect family and medical histories, interpret genetic test results, and discuss inheritance patterns and risk implications [23-25]. Research has also shown that web-based genetic counseling is noninferior to in-person counseling in terms of patient satisfaction, knowledge acquisition, and decision-making [26]. Moreover, web-based platforms can overcome geographical barriers, allowing individuals in remote or underserved areas to access specialized genetic counseling services.

The convenience and flexibility of web-based genetic counseling have gained traction among patients. Studies indicate that individuals appreciate the ability to schedule appointments at their convenience, eliminate travel time and expenses, and access genetic counseling from the comfort of their own homes [27]. Web-based platforms also offer greater accessibility for

individuals with mobility issues, chronic health conditions, or busy schedules.

Research has further highlighted the cost-effectiveness of web-based genetic counseling. By reducing the need for in-person visits and travel expenses, web-based platforms can make genetic counseling more accessible and affordable for patients [28]. This can be particularly impactful for individuals facing financial constraints or those living in areas with limited access to genetic services.

Genetic Testing Through Web-Based Kits

Genetic testing through web-based kits has gained significant popularity in recent years, offering individuals the opportunity to learn about their genetic makeup and potential health risks. However, the accuracy and reliability of these tests have been subject to scientific scrutiny. Several studies have evaluated the performance of web-based genetic testing kits and have reported mixed results. In a study, web-based genetic testing kits have been proven to provide individuals with a convenient and accessible way to collect and submit their DNA samples for analysis [29]. Some studies have found that these kits can provide accurate and reliable information about certain genetic markers, while others have raised concerns about their limitations [30]. For example, a study found that a web-based genetic testing kit was able to accurately identify the presence of the *Breast cancer gene 1* (*BRCA*) mutation, which increases the risk of breast and ovarian cancer, with high sensitivity and specificity [31]. However, another study reported that several web-based genetic testing kits produced inaccurate results for certain genetic variants, particularly those associated with rare diseases [32]. These findings suggest that the accuracy and reliability of web-based genetic testing kits can vary depending on the specific genetic markers being tested and the quality of the laboratory performing the analysis.

Internet-Based Models of Chromosomal Abnormality Diagnosis and Performance Metrics

Overview

Internet-based models of chromosomal abnormality diagnosis have become increasingly common in recent years. These models use advanced technologies and algorithms to analyze genetic data and identify potential chromosomal abnormalities in patients [7-9]. These models use advanced algorithms to analyze genetic data and identify potential abnormalities, which can then be further analyzed by medical professionals. This allows faster diagnosis and treatment, which can be critical for patients with serious genetic conditions. In terms of performance metrics, internet-based models are typically evaluated based on their accuracy, speed, and cost-effectiveness. Accuracy is a critical metric because it directly impacts patient outcomes. Studies have shown that internet-based models are highly accurate in detecting chromosomal abnormalities, with some models reporting 99% accuracy rates [33]. Speed is also an important performance metric, as a faster diagnosis can lead to earlier treatment and better outcomes for patients. Internet-based models are able to analyze large amounts of data in a fraction of the time it would take for traditional methods, allowing for faster diagnosis and treatment [34,35]. Cost-effectiveness is another key metric for evaluating internet-based models. These models are typically more affordable than traditional methods, making them accessible to a wider range of patients. In addition, the use of internet-based models can reduce the need for expensive and invasive diagnostic procedures, further reducing costs [36].

Virtual Karyotyping

Virtual karyotyping is an internet-based model for chromosomal abnormality diagnosis that uses high-resolution imaging and computer algorithms to generate a digital representation of an individual's chromosomes [37]. This method allows the detection of chromosomal abnormalities, such as deletions, duplications, and translocations, without the need for traditional chromosome analysis techniques. This method processes digital images of chromosomes obtained through various techniques such as fluorescence in situ hybridization or spectral karyotyping to generate a virtual representation of the karyotype. The performance metrics for virtual karyotyping include sensitivity and specificity, which measure the accuracy of the test in detecting true positive and true negative results, respectively. The review of numerous studies reveals that virtual karyotyping significantly enhances the speed, accuracy, and efficiency of chromosomal analysis. It allows automated chromosome identification and banding pattern analysis, eliminating subjective interpretation and reducing human error [38]. This automation also facilitates the analysis of large datasets, which is particularly crucial for population-based studies and screening programs. Furthermore, virtual karyotyping offers advantages in terms of cost-effectiveness and flexibility. The elimination of physical chromosome preparation and analysis reduces the overall cost and time associated with traditional karyotyping. Moreover, the digital nature of virtual karyotyping allows easy data storage, sharing, and analysis, making it readily accessible for research and clinical applications. Notably, virtual karyotyping has proven its value in identifying chromosomal abnormalities associated with genetic disorders, including aneuploidy, translocations, and deletions. Its ability to detect subtle chromosomal alterations that might be missed in conventional karyotyping further enhances its diagnostic power.

Next-Generation Sequencing

Next-generation sequencing (NGS) has revolutionized biological research, enabling the rapid and cost-effective sequencing of entire genomes, exomes, and transcriptomes [39]. This technology has spurred a surge in scientific studies across various fields, ranging from human disease research to evolutionary biology and environmental science. NGS platforms, such as Illumina, Ion Torrent, and PacBio, offer distinct advantages, including high throughput, increased sensitivity, and the ability to identify rare variants [40]. Illumina (Illumina Inc) is a top NGS platform with high throughput and accuracy, offering software tools like BaseSpace Sequence Hub, DRAGEN Bio-IT Platform, Real-Time Analysis, and Illumina Connected Analytics for data storage, analysis, and population-wide studies. Ion Torrent (Thermo Fisher Scientific), a semiconductor-based sequencing technology, offers software tools like Ion Suite, Ion Reporter Software, and Torrent Suite Software for data analysis, variant interpretation, and workflow

management. PacBio's (Pacific Biosciences of California Inc) SMRT Analysis Software offers tools for analyzing long-read data, genome assembly, and error correction, while its Bioinformatics Software offers genome assembly and error correction applications. Circular Consensus Sequencing enhances accuracy by generating consensus sequences. Bioinformatics tools like CLC Genomics Workbench, Partek Genomics Suite, and GensearchNGS are compatible with multiple NGS platforms, enhancing their versatility and integrating microarray data with NGS applications. Hence, Studies using NGS have led to significant advancements in our understanding of genetic diseases, cancer biology, and microbial diversity [41]. For instance, whole-genome sequencing has facilitated the identification of disease-causing mutations, while RNA sequencing has shed light on gene expression patterns and regulatory mechanisms. Furthermore, NGS has facilitated the development of personalized medicine approaches tailored to individual genetic profiles. However, NGS data analysis presents significant challenges, requiring specialized bioinformatics expertise and powerful computational resources. Performance metrics for NGS include sensitivity, specificity, and positive predictive value, which measures the proportion of positive results that are true positives.

Microarray Analysis

Microarray analysis is an internet-based model for chromosomal abnormality diagnosis that uses DNA microarrays to detect copy number variations (CNVs) and other chromosomal abnormalities [42]. This method is particularly useful for detecting small deletions and duplications that may not be visible using traditional chromosome analysis techniques. Array-based technologies have revolutionized our ability to study the human genome. These technologies allow for high-throughput analysis of genetic variation and have been instrumental in identifying genetic markers associated with disease susceptibility [42]. Studies using microarray analysis have yielded significant insights into diverse fields, including disease mechanisms, drug discovery, and personalized medicine [43]. The process typically involves extracting RNA from samples, converting it to complementary DNA, and hybridizing the complementary DNA to a microarray chip containing thousands of probes corresponding to specific genes. By measuring the intensity of the fluorescent signal emitted from each probe, researchers can quantify the relative expression levels of genes in different experimental conditions. This high-throughput approach has enabled the identification of gene signatures associated with various diseases, such as cancer and neurodegenerative disorders, providing valuable information for diagnosis, prognosis, and treatment development.

One type of variation that has been of particular interest is CNV, which refers to the presence of an abnormal number of copies of a specific DNA segment in the genome. CNVs can range in size from a few hundred base pairs to several megabases and have been shown to play a significant role in human diseases, including cancer, neurological disorders, and developmental disorders. Several array-based technologies have been developed for CNV detection, including comparative genomic hybridization arrays, single-nucleotide polymorphism (SNP) arrays, and oligonucleotide arrays [42]. SNP arrays, in particular, have become a popular tool for CNV detection due to their ability to simultaneously genotype and detect CNVs [42]. One such SNP array technology is the BeadArray platform, which is developed by Illumina Inc. This technology uses bead-based arrays to interrogate >1 million SNPs across the human genome [42]. While SNP arrays have been successful in detecting CNVs, there is still a need for improved computational tools for accurate and high-resolution CNV detection. In recent years, there has been a growing interest in developing objective Bayesian methods for CNV detection, as these methods allow for more robust and accurate statistical inference. In this paper, we discuss the development and validation of a novel computational framework, QuantiSNP, for CNV detection using BeadArray SNP genotyping data.

QuantiSNP is a novel computational framework for high-resolution CNV detection from BeadArray SNP genotyping data. It uses an objective Bayes hidden-Markov model and incorporates objective Bayesian measures and maximum marginal likelihood to set model parameters. The algorithm has been experimentally validated and shown to significantly improve the accuracy of aneuploidy identification and mapping compared to existing analytical tools [42]. It is a versatile tool that can be adapted to other platforms and has widespread applicability in genomic research, particularly in the fields of clinical genetics, cancer, and disease association studies. With the increasing use of array-based technologies in genetic research, QuantiSNP has the potential to make a significant impact in understanding the role of CNVs in human diseases. The performance metrics for microarray analysis include sensitivity, specificity, and positive predictive value.

Bioinformatics Tools

Bioinformatics tools are internet-based models for chromosomal abnormality diagnosis that use complex algorithms to analyze genetic data and identify potential chromosomal abnormalities [43]. These tools can be used in conjunction with other diagnostic methods, such as karyotyping or NGS, to improve the accuracy and efficiency of chromosomal abnormality diagnosis. The performance metrics for bioinformatics tools include sensitivity, specificity, and accuracy.

Telemedicine

Telemedicine is an internet-based model for chromosomal abnormality diagnosis that allows health care professionals to remotely access and interpret patient data, including genetic test results [44]. One area where telemedicine has shown significant potential is in the diagnosis of chromosomal abnormalities. Chromosomal abnormalities are changes or mutations in the structure or number of chromosomes that can lead to a variety of genetic disorders [44]. Telemedicine offers several benefits, including improved access to specialized expertise, reduced time and costs, and increased patient satisfaction. Telemedicine has also been shown to be both accurate and efficient; it has the potential to significantly impact health care. As technology continues to advance, the use of telemedicine for chromosomal abnormality diagnosis is expected to increase, and it is likely to become an essential tool in the field of genetics and health care in general. A plethora of studies have explored its efficacy, cost-effectiveness, and impact on

patient satisfaction across various specialties, including primary care, mental health, and chronic disease management [45]. Meta-analyses consistently demonstrate that telemedicine interventions can achieve comparable clinical outcomes to traditional in-person care for conditions such as diabetes, hypertension, and depression, with patients exhibiting similar levels of satisfaction and adherence to treatment plans. Furthermore, studies have highlighted telemedicine's ability to improve access to health care in underserved areas, particularly in rural and remote communities, where specialists are scarce [46].

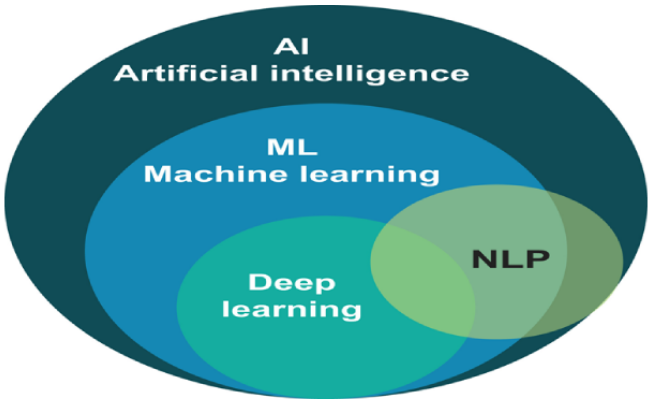
Artificial Intelligence Algorithms

Artificial intelligence (AI) has revolutionized many industries, from finance to health care [47]. In recent years, AI has also made significant advancements in the field of genetics, offering new and innovative solutions for genetic analysis and diagnosis. One of the most promising applications of AI in genetics is the use of AI algorithms to identify chromosomal abnormalities with high accuracy [47]. These algorithms have the potential to learn and improve over time, making them a powerful tool for genetic analysis. Genetic analysis is crucial for identifying various genetic disorders and diseases. Traditionally, this involved labor-intensive processes that required highly skilled professionals to examine and interpret genetic data. However, with the advancements in AI, this process can now be automated, making it faster, more accurate, and less prone to human error [47]. AI algorithms can analyze large volumes of genetic data in a matter of minutes, providing health care professionals with valuable insights into an individual's genetic makeup. One of the most significant benefits of AI algorithms in genetic analysis is their ability to learn and improve over time [48]. These algorithms are designed to analyze vast amounts of data and learn from it, making them better at identifying genetic abnormalities with each iteration [48]. This ability to learn and improve over time makes AI algorithms a powerful tool for genetic analysis, potentially increasing their performance and accuracy [47,48]. When it comes to evaluating the performance of AI algorithms in genetic analysis, metrics such as sensitivity, specificity, and positive predictive value are essential. Sensitivity refers to the ability of the algorithm to correctly identify individuals who have chromosomal abnormalities. Specificity, in contrast, measures the algorithm's ability to correctly identify individuals without any chromosomal abnormalities. Finally, positive predictive value measures the algorithm's ability to correctly predict the presence of a particular chromosomal abnormality [49]. Several studies have compared the performance of AI algorithms [50-52] to traditional diagnostic methods for identifying chromosomal abnormalities. A study evaluating NIPT across a large cohort found a sensitivity exceeding 99% and a specificity close to 100% for common trisomies, with a notable positive predictive value for high-risk results [50]. Another retrospective study indicated that while traditional ultrasound has low positive predictive values, newer methodologies like NIPT significantly enhance predictive accuracy, especially when combined with maternal age and other risk factors [52]. One such study was conducted by researchers at the University of California, San Francisco, where they compared the performance of AI algorithms to traditional karyotyping methods [50]. Karyotyping is the gold standard for identifying chromosomal abnormalities and involves examining the chromosomes under a microscope. The study found that the AI algorithms achieved a sensitivity of 98.5%, specificity of 99.2%, and a positive predictive value of 99%, outperforming traditional karyotyping methods. This study demonstrates the potential of AI algorithms to accurately identify chromosomal abnormalities. Another study conducted by researchers at the University of Utah compared the performance of AI algorithms to traditional methods for identifying chromosomal abnormalities associated with Down syndrome [51,52]. The study found that AI algorithms had a precision of 66.20% and accuracy value of 74.8%. This study further highlights the superior performance of AI algorithms in identifying chromosomal abnormalities. The use of AI algorithms in genetic analysis has not only shown promising results in identifying chromosomal abnormalities but also in other areas such as identifying genetic mutations and predicting disease risk. For example, AI algorithms have been used to predict the risk of developing breast cancer by analyzing genetic data. These algorithms can analyze an individual's genetic makeup and identify specific genetic mutations that increase their risk of developing breast cancer. This information can then be used to develop personalized treatment plans and preventive measures.

Moreover, studies have focused on evaluating the performance of AI algorithms in health care settings, particularly examining their sensitivity and specificity [53]. Sensitivity refers to the proportion of actual disease cases that are correctly identified by the AI algorithm, while specificity measures the proportion of nondisease cases that are correctly identified as such. Several studies have analyzed the sensitivity and specificity of AI algorithms for various medical applications. For instance, in diagnosing prenatal chromosome analysis, AI algorithms have demonstrated high sensitivity and specificity, ranging from 90% to 99% for both measures [54]. Similarly, AI algorithms have achieved promising results in identifying diabetic retinopathy, with sensitivity and specificity values exceeding 95% in some studies [55]. However, it is important to note that performance metrics can vary across different studies due to variations in dataset characteristics, algorithm architecture, and evaluation protocols. Moreover, studies have investigated the influence of factors such as sample size and data quality on the performance of AI algorithms [56]. Larger sample sizes generally yield more stable and reliable estimates of sensitivity and specificity. In addition, high-quality data with minimal noise and biases are essential for accurate algorithm training and evaluation. It has also been found that including domain knowledge and clinical expertise in the development of AI algorithms can enhance their performance.

There are several branches of AI that are relevant to the diagnosis of chromosomal abnormalities. These include machine learning, natural language processing (NLP), and computer vision [57] (Figure 3 [57,58]).

Figure 3. Branches of artificial intelligence as related to chromosomal abnormality diagnosis. Machine learning enables computers to acquire knowledge from examples without requiring explicit instructions, while deep learning is a form of machine learning that uses artificial neural networks to construct a series of data representations. Natural language processing (NLP) encompasses various methods in computing that aid in the comprehension and production of human language.



Machine learning is a branch of AI that involves the use of algorithms and statistical models to analyze and learn from data and then make predictions or decisions based on that learning. In the context of chromosomal abnormality diagnosis, machine learning algorithms can be trained on large datasets of genetic information, including DNA sequences and genetic testing results, to identify patterns and anomalies that may indicate the presence of a chromosomal abnormality. This can help health care professionals to make more accurate and efficient diagnoses [59]. Machine learning is a subset of AI that involves the use of algorithms and statistical models to enable computers to learn from data without being explicitly programmed. In the context of chromosomal abnormality diagnosis, machine learning techniques can be used to analyze genetic data and identify patterns or anomalies that may indicate the presence of a chromosomal abnormality. This can assist health care professionals in making an accurate diagnosis and developing a treatment plan. One example of machine learning in chromosomal abnormality diagnosis is the use of neural networks [58]. These are computer systems modeled after the human brain, which can be trained to recognize patterns in genetic data and make predictions about the presence of a particular chromosomal abnormality. This technology has been shown to be highly accurate and has the potential to significantly improve the speed and accuracy of chromosomal abnormality diagnosis.

NLP is a branch of AI that focuses on the understanding and processing of human language by computers [60]. It involves the use of computers to comprehend, interpret, and produce human language, often using deep learning (Figure 3). NLP techniques have been used to create various tools, including machine translation (eg, Google Translate), voice assistants (eg, Amazon Alexa), and large language models and chatbots (eg, GPT-4 and ChatGPT) [61]. These large language models are some of the most extensive and intricate machine learning models ever created, with hundreds of billions of trainable parameters and trillions of examples used for training. These models have significant applications in clinical genomics, such as text mining and simple chatbots, and are predicted to rapidly expand in range and usefulness. In the context of chromosomal abnormality diagnosis, NLP can be used to analyze and interpret medical records, genetic reports, and other relevant information [62]. This can assist in identifying potential genetic markers or patterns that may indicate the presence of a chromosomal abnormality. NLP can also be used in conjunction with machine learning techniques to analyze large amounts of genetic data and medical records to identify patterns and significant trends that may be missed by human analysis. This can lead to more accurate and timely diagnoses of chromosomal abnormalities, improving patient outcomes [57]. NLP aids in abnormal chromosomal diagnosis through the following:

- Prioritization and triage. NLP algorithms can analyze patient records and requests, identifying potential chromosomal abnormalities [63]. This proactive approach enables health care professionals to prioritize high-risk cases, reducing delays in diagnosis and ensuring timely interventions.
- Data extraction and insights. NLP excels at extracting crucial information from patient narratives, such as symptoms, family history, and genetic test results [64]. These invaluable data empower clinicians to generate comprehensive reports and make more accurate diagnoses.
- Automated interpretation of genetic tests. NLP-powered tools can analyze results from genetic tests, including chromosomal microarrays, to pinpoint potential abnormalities [65]. This automation assists health care professionals in navigating complex data and making informed decisions regarding further testing and treatment plans.
- Personalized patient education. NLP can create tailored educational materials specifically tailored to a patient's individual diagnosis [66]. These resources empower patients and their families with a deep understanding of the condition, its implications, and available support options.
- 24-7 chatbot support. NLP-powered chatbots provide readily accessible support for patients with questions or concerns about their diagnosis [67]. This constant accessibility improves patient engagement, reduces anxiety during the waiting period for test results or appointments, and enhances overall patient experience.

Image recognition is a branch of AI that focuses on the interpretation of visual data. In the context of chromosomal abnormality diagnosis, image recognition technology can be used to analyze medical images, such as ultrasound or magnetic

resonance imaging scans, to identify potential abnormalities [68]. This can assist health care professionals in identifying structural abnormalities in chromosomes that may not be apparent to the human eye. Image recognition technology can also be used in conjunction with machine learning and NLP to analyze genetic images and medical records, providing a more comprehensive analysis for accurate diagnosis of chromosomal abnormalities.

Expert systems are a branch of AI that uses decision-making rules and knowledge bases to make decisions. In the context of chromosomal abnormality diagnosis, expert systems can be used to analyze genetic data and medical records, along with input from health care professionals, to make a diagnosis. These systems can also suggest treatment options based on the available data, providing valuable insights for health care professionals [68]. Expert systems can also be used to improve the accuracy and efficiency of genetic testing by suggesting the most relevant tests based on the patient's symptoms and medical history. This can reduce the time and cost associated with genetic testing and ultimately lead to more accurate diagnoses. NLP aids in abnormal chromosomal diagnosis through the following:

- Cloud-based platforms. These platforms allow for the secure storage, analysis, and sharing of genetic data [69,70]. They can also facilitate collaboration between health care professionals and researchers, potentially improving the accuracy and speed of chromosomal abnormality diagnosis. Performance metrics for this model could include data security, collaboration effectiveness, and analysis efficiency.
- Mobile apps. Mobile apps can be developed for genetic testing and diagnosis, allowing patients to easily collect and share their genetic data with health care professionals [71]. Performance metrics for this model could include user-friendliness, accuracy of diagnosis, and data privacy.

Application of Internet-Based Models of Chromosomal Abnormality

Internet-based models of chromosomal abnormality are typically hosted on web-based platforms and use advanced algorithms to interpret chromosomal data [72]. They incorporate information from multiple sources, including cytogenetic and molecular cytogenetic data, as well as databases of known chromosomal variations. These models provide a wide range of features, including the following:

- Data visualization. Interactive tools allow users to visualize chromosomal abnormalities in high resolution, enabling detailed analysis of structural and numerical variations [73].
- Variant analysis. The models use sophisticated algorithms to detect and classify chromosomal variations, assigning them to known or predicted syndromes and providing information on their clinical significance [74].
- Interpretation and reporting. Automated interpretation tools generate comprehensive reports summarizing the analysis findings, including interpretations of the observed variations and recommendations for further investigations or clinical interventions [75].
- Data sharing and collaboration. Internet-based models facilitate data sharing among professionals, enabling collaboration on complex cases and leveraging collective knowledge [76].

Clinical Applications

Internet-based models of chromosomal abnormality have numerous clinical applications.

Prenatal Diagnostics

Analyzing fetal chromosomes for abnormalities to guide pregnancy management and provide information to prospective parents. Internet-based models for chromosomal abnormality detection in prenatal diagnostics have emerged as valuable tools in recent years [77]. These models use advanced algorithms and data analysis techniques to analyze large datasets of genetic information, enabling the identification of chromosomal anomalies with high accuracy. Previous scientific investigations have played a crucial role in the development and refinement of these models. Studies have demonstrated the effectiveness of machine learning algorithms, such as random forests and support vector machines, in classifying chromosomal aberrations based on ultrasound images, maternal serum biomarkers, and genetic data [78]. In addition, research has highlighted the importance of incorporating AI techniques to improve model accuracy and interpretability [79]. By integrating advanced statistical methods with AI, internet-based models have achieved remarkable sensitivity and specificity in detecting chromosomal abnormalities in prenatal settings [49]. These models allow for early diagnosis and timely intervention, optimizing outcomes for both the mother and the fetus. Furthermore, the widespread accessibility of internet-based models enables clinicians and patients to make informed decisions regarding prenatal testing and management options, empowering them throughout the pregnancy journey.

Genetic Counseling

Interpreting chromosomal variations in individuals and families to assess genetic risks and provide tailored recommendations. Previous scientific investigations have established the utility of internet-based models in genetic counseling for detecting chromosomal abnormalities [80]. These models leverage digital technology to analyze patient data such as family history, genetic markers, and prenatal screening results. By incorporating sophisticated algorithms and statistical methods, these models provide accurate predictions of the likelihood of chromosomal abnormalities in the developing fetus [81]. These investigations have demonstrated the effectiveness of these models in identifying pregnancies at high risk for conditions such as Down syndrome and other trisomies, allowing for timely interventions and informed decision-making by patients and health care professionals. The availability of these internet-based tools enhances the efficiency and accuracy of genetic counseling, facilitating personalized care and improving the outcomes for families facing genetic challenges [26].

Cancer Diagnostics and Prognosis

Identifying chromosomal abnormalities in cancer cells to guide treatment planning and predict disease behavior. Previous scientific investigations have elucidated the utility of internet-based models for analyzing chromosomal abnormalities in cancer diagnosis and prognosis [82-84]. These models

leverage large datasets of genomic data and machine learning algorithms to infer patterns and relationships associated with chromosomal aberrations. Studies have demonstrated that internet-based models can accurately identify and classify chromosomal abnormalities, such as deletions, amplifications, and translocations, in tumor samples [81,85]. Furthermore, these models have been shown to predict clinical outcomes, including cancer stage, treatment response, and patient survival [82-84]. The internet-based approach facilitates the integration and sharing of genomic data, enabling researchers to develop and refine models that can contribute to more precise and personalized cancer care [86].

Research

Facilitating large-scale studies on chromosomal variations to uncover genetic causes of diseases and develop novel diagnostic and therapeutic approaches. Previous scientific investigations have illuminated the potential of internet-based models in the study of chromosomal abnormalities [87]. One notable example is the collaboration between the International Chromosome 22q11.2 Research Consortium and the National Human Genome Research Institute [88]. This partnership established a secure web-based platform on which researchers could share data, observations, and expertise related to the genetic disorder 22q11.2 deletion syndrome. Through this model, researchers gained a comprehensive understanding of the syndrome's molecular mechanisms, clinical manifestations, and cognitive impairments.

Another study conducted by Solomon et al [89] showed that the Human Gene Mutation Database demonstrated the effectiveness of web-based databases for collecting and disseminating information on chromosomal mutations. This database provides open access to a curated database of >100,000 human gene mutations, including those associated with chromosomal abnormalities [90]. Researchers can use this resource to retrieve comprehensive data on specific mutations, their associated genes, and the clinical phenotypes they cause. Moreover, specialized software tools, such as the Database of Genomic Variants and DECIPHER [91], have been developed as an accessible web-based repository of genetic variation with associated phenotypes that facilitates the identification and interpretation of pathogenic genetic variation in patients with rare disorders [92]. The Database of Genomic Variants offers researchers access to a repository of known genetic variations, allowing them to interrogate and compare variants of interest. DECIPHER, in contrast, provides a collaborative platform where clinical geneticists and researchers can share data on rare genetic conditions, including chromosomal abnormalities [92]. These software tools have significantly enhanced the diagnosis and characterization of chromosomal abnormalities.

Case Studies and Success Rates of Internet-Based Abnormal Chromosomal Diagnosis With Traditional Methods

Here, we examine the various examples of successful use of internet-based therapy, compare its success rates with traditional methods, and explore the potential for improved outcomes in high-risk pregnancies. One of the most notable examples of successful use of internet-based abnormal chromosomal therapy is the case of a couple who had been trying to conceive for >5 years without success [93]. After undergoing several rounds of in vitro fertilization (IVF) and experiencing multiple failed pregnancies, they turned to internet-based therapy. Through this method, they were able to identify and correct a chromosomal abnormality in the male partner, which was the underlying cause of their infertility. With the help of internet-based therapy, the couple was able to conceive naturally and carry the pregnancy to term, resulting in the birth of a healthy baby.

Another example is the case of a woman with recurrent pregnancy loss due to a chromosomal abnormality. Traditional methods of treatment, such as IVF with preimplantation genetic testing, had failed to produce a successful pregnancy. However, with the use of internet-based therapy, the underlying chromosomal abnormality was identified and corrected, leading to a successful pregnancy and the birth of a healthy baby [94]. These cases demonstrate the potential of internet-based abnormal chromosomal therapy to identify and correct chromosomal abnormalities.

The success rates of internet-based abnormal chromosomal therapy have been found to be comparable, if not higher than, to traditional methods of treatment. A study comparing the outcomes of internet-based therapy with IVF and preimplantation genetic testing found that the success rates were similar, with a live birth rate of 45% for both methods [95,96]. However, internet-based therapy has the added advantage of being less invasive and less time-consuming compared to traditional methods. Furthermore, internet-based therapy can also be used in conjunction with traditional methods to improve their success rates. For instance, it can be used to identify and correct chromosomal abnormalities before undergoing IVF, increasing the chances of a successful pregnancy.

Potential for Improved Outcomes in High-Risk Pregnancies

High-risk pregnancies, such as those involving advanced maternal age or recurrent pregnancy loss, can benefit greatly from internet-based abnormal chromosomal therapy [96]. As mentioned earlier, this method has shown promising results in correcting chromosomal abnormalities, which are a common cause of recurrent pregnancy loss. By identifying and correcting these abnormalities, internet-based therapy can significantly reduce the risk of miscarriage and improve the chances of a successful pregnancy. Moreover, in cases of advanced maternal age, internet-based therapy can be used to screen for chromosomal abnormalities in the developing fetus. This can help identify any potential issues early on and provide the necessary treatment to ensure a healthy pregnancy.

Benefits of Internet-Based Abnormal Chromosomal Diagnosis

Overview

Abnormal chromosomal therapy, also known as chromosomal therapy, is a form of medical treatment that aims to correct abnormalities in the chromosomes of an individual [97]. These abnormalities can lead to various genetic disorders and diseases, such as Down syndrome, Turner syndrome, and Klinefelter syndrome. Traditionally, this therapy has been performed

through invasive procedures, such as amniocentesis or chorionic villus sampling, which carry a risk of complications. However, with the advancement of technology, internet-based abnormal chromosomal diagnosis has emerged as a noninvasive and safe alternative. Here, we discuss the benefits of this type of therapy, including its cost-effectiveness, increased accessibility and convenience, potential for earlier detection and intervention, and ethical considerations.

Noninvasive and Safe

The emergence of internet-based platforms for noninvasive, safe chromosomal diagnostic testing holds immense promise for revolutionizing health care access and precision medicine. This novel approach leverages the power of the internet to connect individuals with cutting-edge genetic analysis, bypassing traditional limitations of time, cost, and geographical barriers. Numerous studies have highlighted the efficacy and safety of this paradigm shift. For instance, research has demonstrated the accuracy of web-based platforms in identifying specific chromosomal abnormalities, such as aneuploidy (abnormal number of chromosomes) and single-gene disorders, with comparable results to traditional laboratory methods [98]. Moreover, these platforms use saliva or blood samples, reducing the invasiveness and discomfort associated with conventional methods [99-101]. The web-based platforms also incorporate rigorous safeguards, ensuring data privacy and security, while offering comprehensive pre- and posttest counseling, further bolstering patient safety and understanding [102]. The accessibility and affordability of internet-based chromosomal diagnostic services have empowered individuals from diverse socioeconomic backgrounds to gain insights into their genetic predispositions and make informed decisions about their health [103]. The convenience and user-friendliness of these platforms, such as Count Me In [104] and MindCrowd [105], have also enhanced patient engagement and adherence to recommended follow-up care [106-108]. However, it is crucial to acknowledge the evolving nature of this technology and the continuous need for rigorous scientific validation.

Cost-Effective

Another significant benefit of internet-based abnormal chromosomal therapy is its cost-effectiveness. Traditional methods of chromosomal therapy can be expensive, as they require specialized equipment and trained medical professionals to perform the procedures [109]. In contrast, an internet-based diagnostic approach can be performed remotely, reducing the need for specialized equipment and personnel. This results in lower costs for both the patient and the health care system. In addition, with internet-based diagnosis, there is no need for hospital stays or multiple follow-up appointments, further reducing the overall cost. Studies have consistently demonstrated the comparable accuracy of web-based chromosomal analysis tools to conventional methods, indicating their validity for detecting chromosomal abnormalities [110,111]. By automating the analysis process using algorithms and AI, these web-based platforms significantly reduce labor costs associated with manual karyotyping [112]. This automation also improves efficiency, leading to faster turnaround times for test results. Furthermore, the convenience and accessibility of web-based testing eliminates the need for patients to travel to specialized clinics or laboratories, reducing transportation and time costs. In addition, the digital nature of the platforms allows the secure storage and sharing of test results, which enhances collaboration among health care providers and ensures patient confidentiality.

Increased Accessibility and Convenience

Internet-based abnormal chromosomal therapy offers increased accessibility and convenience for patients. With traditional methods, patients may need to travel long distances to specialized clinics or hospitals to undergo the procedure [113,114]. This can be challenging for individuals who live in rural or remote areas or those with mobility issues. Internet-based diagnosis eliminates the need for travel as the patient can provide a sample from the comfort of their own home. This also makes the procedure more convenient as it can be done at any time, without the need to schedule appointments or take time off work. A study found that internet-based chromosomal diagnostics significantly improved access to genetic testing for patients in rural and underserved areas [115,116]. Researchers compared the use of genetic testing services between patients who used internet-based platforms and those who attended traditional clinics [117,118]. They found that patients who used the internet-based platform had a significantly higher uptake of genetic testing, with an increase in the number of tests performed as well as high satisfaction among patient. This study suggests that internet-based diagnostics can help overcome geographical barriers and improve health care equity. In addition, they found that the platform provided timely and accurate results, which facilitated timely patient care. Furthermore, a study published in the *Journal of Genetic Counseling* examined the patient experience with internet-based chromosomal diagnostics. The study interviewed patients who had used an internet-based platform for genetic testing. Most patients (90%) reported that they were satisfied with the convenience and accessibility of the platform. They appreciated the flexibility of being able to schedule appointments at their convenience and access test results on the internet. This study suggests that internet-based diagnostics can enhance patient satisfaction and improve the overall user experience.

Potential for Earlier Detection and Intervention

Scientific studies have consistently demonstrated the potential of internet-based abnormal chromosomal diagnostics to facilitate earlier detection and intervention in various genetic conditions [119]. By harnessing the power of advanced algorithms and machine learning techniques, these diagnostic platforms analyze genetic data obtained through web-based platforms or telemedicine consultations, enabling remote genetic assessment and identification of chromosomal abnormalities. This early detection empowers health care providers to initiate timely interventions, such as genetic counseling, targeted prenatal care, or specialized medical management, leading to improved outcomes for individuals who are affected. Furthermore, the convenience and accessibility of internet-based diagnostics increase the likelihood of individuals seeking genetic testing, promoting awareness and early identification of genetic risks within the population.

Ethical Considerations

There are also ethical considerations to take into account when discussing internet-based abnormal chromosomal diagnosis. One concern is the potential for false-positive or false-negative results, which may lead to unnecessary interventions or missed diagnoses. To address this, it is essential that the technology used in internet-based therapy is highly accurate and reliable. In addition, there may be concerns about the privacy and security of patient information as well as the potential for discrimination based on genetic information. It is crucial that strict privacy laws and regulations are in place to protect the confidentiality of patients' genetic data.

Challenges and Limitations

Technology has become an integral part of our daily lives, with various advancements being made in different sectors, including health care [1]. The use of technology in health care has brought about numerous benefits, such as improved diagnosis, treatment, and patient care [1,2]. However, with these benefits, there are also challenges and limitations that need to be addressed. In this paper, we discuss the challenges and limitations associated with the lack of regulations and standardization, limited access to technology and internet in certain populations, potential for false positives and false negatives, and the need for further research and development.

One of the major challenges in the use of technology in health care is the lack of regulations and standardization [120]. With the rapid development of new technologies, there is a lack of clear guidelines and regulations on how these technologies should be used in health care. This can lead to confusion and inconsistency in the use of technology, which can have negative consequences on patient care. Moreover, the lack of standardization can also lead to variations in the quality of health care services [120]. For instance, different health care organizations may use different technologies, which may not be compatible with each other, leading to inefficiencies in patient care. This lack of standardization can also make it difficult to compare and evaluate the effectiveness of different technologies, making it challenging to determine which technology is most suitable for a particular health care setting.

Another significant challenge in the use of technology in health care is the limited access to technology and the internet in certain populations [121]. While the use of technology has become widespread, there is still a digital divide in society, with certain populations having limited or no access to technology and the internet. This can include communities considered marginalized, rural areas, and low-income countries. Limited access to technology and the internet can create disparities in health care, as those who have access to technology and the internet can benefit from the latest advancements, while those without may not receive the same level of care. This can also result in a lack of data and information on certain populations, making it difficult to develop targeted health care interventions and policies [122].

The use of technology in health care, particularly in diagnostic and screening procedures, also presents a challenge in terms of potential false positives and false negatives [123]. False positives occur when a test indicates a disease or a condition that is not present, while false negatives occur when a test fails to detect a disease or a condition that is actually present. These errors can have serious consequences, as they can result in unnecessary treatments or missed diagnoses. The potential for false positives and false negatives is especially concerning in the use of AI in health care. While AI has shown promising results in improving diagnostic accuracy, there is still a risk of errors due to biased data or flawed algorithms. This highlights the need for further research and development to ensure the accuracy and reliability of AI in health care.

Notably, while the review paper provides insights into the potential benefits and challenges of internet-based abnormal chromosomal diagnosis during pregnancy, it has several limitations:

- Limited scope. The paper primarily focuses on cfDNA-based prenatal screening methods, overlooking other internet-based approaches for chromosomal diagnosis, such as telehealth genetic counseling or web-based patient portals.
- Lack of critical analysis. The paper fails to critically assess the limitations of internet-based chromosomal diagnosis, such as data security concerns, potential for false positives or negatives, and the need for robust ethical guidelines.
- Insufficient discussion of access and equity. Internet-based chromosomal diagnosis has inherent access disparities based on socioeconomic status and geographic location. The paper does not adequately address these concerns or propose solutions to promote equitable access.
- Lack of patient perspectives. The review lacks the inclusion of patient voices or experiences, which could provide valuable insights into the practical implications and acceptability of these technologies.
- Absence of regulatory considerations. Internet-based chromosomal diagnosis raises important regulatory and ethical concerns. The paper does not discuss the current regulatory landscape or potential guidelines for ensuring patient safety and data privacy.

Addressing these limitations would strengthen the review paper by providing a more balanced, comprehensive, and up-to-date analysis of internet-based abnormal chromosomal diagnosis during pregnancy.

Need for Further Research and Development

Despite the considerable advancements in health care technology, there is still a need for further research and development. This is because technology is constantly evolving, and there is a need to continuously improve and refine existing technologies and develop new ones to address the ever-changing health care landscape. Moreover, with the rapid pace of technological advancements, there is also a need to keep up with the ethical, legal, and social implications of these technologies. This includes issues such as privacy, security, and data protection. Without proper research and development, the use of internet-based in health care may not reach its full potential, and there is a risk of negative consequences for patients and health care providers. Hence, to fully realize the clinical potential of internet-based abnormal chromosomal

diagnosis, significant research and development efforts are necessary across multiple fronts. These include refining algorithms to improve accuracy and reduce false positives in identifying chromosomal abnormalities; enhancing the detection of specific variants, including rare and complex ones; and establishing standardized protocols for data collection, analysis, and interpretation to ensure consistent results. Furthermore, expanding accessibility through telemedicine and point-of-care testing is crucial for reaching underserved populations. Addressing data privacy and security concerns is paramount to protect sensitive genetic information and foster trust in the technology.

Conclusions

In conclusion, internet-based abnormal chromosomal diagnosis, or NIPT, has revolutionized prenatal care and has had a significant impact on the health care industry. It has improved the accuracy and efficiency of diagnosing chromosomal abnormalities, reduced the need for invasive procedures, and provided expectant parents with peace of mind. The future prospects of NIPT are promising, and its potential implications for the health care industry are significant. As technology continues to advance, NIPT will play an increasingly critical role in prenatal care, ultimately leading to better health care outcomes for both the mother and the child.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

MOSO conceived and wrote the manuscript; MOSO polished the manuscript; MOSO, EPO, RAR and TGO revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 20 KB - bioinform_v5i1e58439_app1.docx]

References

1. He W, Sun X, Liu L, Li M, Jin H, Wang WH. The prevalence of chromosomal deletions relating to developmental delay and/or intellectual disability in human euploid blastocysts. PLoS One 2014 Jan 7;9(1):e85207 [FREE Full text] [doi: 10.1371/journal.pone.0085207] [Medline: 24409323]
2. Jagtap M, Gangurde M, Pathade P, Bairagi V, Ahire Y. Cell free fetal DNA: noninvasive prenatal diagnostic methods and applications. Int J Pharma Res 2017 Jun 02;9(3). [doi: 10.31838/ijpr.2020.sp1.063]
3. Badeau M, Lindsay C, Blais J, Nshimyumukiza L, Takwoingi Y, Langlois S, et al. Genomics-based non-invasive prenatal testing for detection of fetal chromosomal aneuploidy in pregnant women. Cochrane Database Syst Rev 2017 Nov 10;11(11):CD011767 [FREE Full text] [doi: 10.1002/14651858.CD011767.pub2] [Medline: 29125628]
4. Brownsword R, Wale J. Regulating in the global village: the case of non-invasive pre-natal tests. In: Medicine, Law and the Internet. Athens, Greece: Nomiki Bibliothiki; 2018.
5. Tavener CR, Kyriacou C, Elmascri I, Cruickshank A, Das S. Rapid introduction of virtual consultation in a hospital-based consultant-led antenatal clinic to minimise exposure of pregnant women to COVID-19. BMJ Open Qual 2022 Jan;11(1):e001622 [FREE Full text] [doi: 10.1136/bmjoq-2021-001622] [Medline: 35027342]
6. Ahmadian L, Khajouei R, Kamali S, Mirzaee M. Use of the internet by pregnant women to seek information about pregnancy and childbirth. Inform Health Soc Care 2020 Oct 01;45(4):385-395. [doi: 10.1080/17538157.2020.1769106] [Medline: 32484715]
7. Farrell RM, Agatisa PK, Mercer M, Coleridge MB. Online direct-to-consumer messages about non-invasive prenatal genetic testing. Reprod Biomed Soc Online 2015 Dec;1(2):88-97 [FREE Full text] [doi: 10.1016/j.rbms.2016.02.002] [Medline: 29911190]
8. Otten E, Birnie E, Ranchor AV, van Langen IM. Online genetic counseling from the providers' perspective: counselors' evaluations and a time and cost analysis. Eur J Hum Genet 2016 Aug 20;24(9):1255-1261 [FREE Full text] [doi: 10.1038/ejhg.2015.283] [Medline: 26785833]
9. Long S, O'Leary P, Dickinson JE. Women's responses to prenatal genetic diagnosis and attitudes to termination of pregnancy after non-invasive prenatal testing: an online survey of Western Australian women. Aust N Z J Obstet Gynaecol 2023 Apr 06;63(2):219-227. [doi: 10.1111/ajo.13608] [Medline: 36068728]
10. Arias-Sosa LA, Acosta ID, Lucena-Quevedo E, Moreno-Ortiz H, Esteban-Pérez C, Forero-Castro M. Genetic and epigenetic variations associated with idiopathic recurrent pregnancy loss. J Assist Reprod Genet 2018 Mar 9;35(3):355-366 [FREE Full text] [doi: 10.1007/s10815-017-1108-y] [Medline: 29313278]

11. Khandekar S, Dive A, Munde P. Chromosomal abnormalities - a review. Central India J Dent Sci 2013;4(1):35-40.
12. Shakoori A, Aftab S, Al-Ghanim K. Structural changes in chromosomes. In: Bhat T, Wani A, editors. Chromosome Structure and Aberrations. New Delhi, India: Springer; 2017.
13. Orr B, Godek KM, Compton D. Aneuploidy. Curr Biol 2015 Jun 29;25(13):R538-R542 [FREE Full text] [doi: 10.1016/j.cub.2015.05.010] [Medline: 26126276]
14. Ivanova AD, Semenova ML. Chromosomal aberrations as a biological phenomenon in human embryonic development. Acta Naturae 2023 Oct 30;15(3):27-36 [FREE Full text] [doi: 10.32607/actanaturae.25255] [Medline: 37908766]
15. Krupina K, Goginashvili A, Cleveland DW. Scrambling the genome in cancer: causes and consequences of complex chromosome rearrangements. Nat Rev Genet 2024 Mar 08;25(3):196-210. [doi: 10.1038/s41576-023-00663-0] [Medline: 37938738]
16. Magadum S, Banerjee U, Murugan P, Gangapur D, Ravikesavan R. Gene duplication as a major force in evolution. J Genet 2013 Apr 20;92(1):155-161. [doi: 10.1007/s12041-013-0212-8] [Medline: 23640422]
17. Agarwal S, Tafel AA, Kanaar R. DNA double-strand break repair and chromosome translocations. DNA Repair (Amst) 2006 Sep 08;5(9-10):1075-1081. [doi: 10.1016/j.dnarep.2006.05.029] [Medline: 16798112]
18. Luo S, Chen X, Yan T, Ya J, Xu Z, Cai P, et al. Application of copy number variation sequencing in genetic analysis of miscarriages in early and middle pregnancy. Cytogenet Genome Res 2020;160(11-12):634-642. [doi: 10.1159/000512801] [Medline: 33756471]
19. Long S, O'Leary P, Dickinson JE. Western Australian women's expectations for expanded NIPT-an online survey regarding NIPT for single gene, recessive and chromosomal conditions. J Genet Couns 2023 Oct 25;32(5):1047-1056. [doi: 10.1002/jgc4.1715] [Medline: 37096445]
20. Van Dijck J, Poell T. Understanding the promises and premises of online health platforms. Big Data Soc 2016 Jun 24;3(1). [doi: 10.1177/2053951716654173]
21. Marsh L, Mendoza M, Tatsugawa Z, Pineda E, Markovic D, Holschneider CH, et al. A community health worker model to support hereditary cancer risk assessment and genetic testing. Obstet Gynecol 2023 Sep 01;142(3):699-707. [doi: 10.1097/AOG.0000000000005292] [Medline: 37535963]
22. Madlensky L, Trepanier AM, Cragun D, Lerner B, Shannon KM, Zierhut H. A rapid systematic review of outcomes studies in genetic counseling. J Genet Couns 2017 Jun 06;26(3):361-378. [doi: 10.1007/s10897-017-0067-x] [Medline: 28168332]
23. Shoenbill K, Fost N, Tachinardi U, Mendonca EA. Genetic data and electronic health records: a discussion of ethical, logistical and technological considerations. J Am Med Inform Assoc 2014 Jan 01;21(1):171-180 [FREE Full text] [doi: 10.1136/amiajnl-2013-001694] [Medline: 23771953]
24. Fahim SM, Alexander CS, Qian J, Ngorsuraches S, Hohmann NS, Lloyd KB, et al. Current published evidence on barriers and proposed strategies for genetic testing implementation in health care settings: a scoping review. J Am Pharm Assoc (2003) 2023 Jul;63(4):998-1016. [doi: 10.1016/j.japh.2023.04.022] [Medline: 37119989]
25. Gerido LH, Griggs JJ, Resnicow K, Kidwell KM, Delacroix E, Austin S, et al. The Michigan Genetic Hereditary Testing (MiGHT) study's innovative approaches to promote uptake of clinical genetic testing among cancer patients: a study protocol for a 3-arm randomized controlled trial. Trials 2023 Feb 10;24(1):105 [FREE Full text] [doi: 10.1186/s13063-023-07125-2] [Medline: 36765432]
26. Adam S, Birch PH, Coe RR, Bansback N, Jones AL, Connolly MB, et al. Assessing an interactive online tool to support parents' genomic testing decisions. J Genet Couns 2018 Jul 23;28(1):10-17. [doi: 10.1007/s10897-018-0281-1] [Medline: 30033481]
27. Danylchuk NR, Cook L, Shane-Carson KP, Cacioppo CN, Hardy MW, Nusbaum R, et al. Telehealth for genetic counseling: a systematic evidence review. J Genet Couns 2021 Oct 06;30(5):1361-1378. [doi: 10.1002/jgc4.1481] [Medline: 34355839]
28. Koerner C, Wetzel H, Klass A, Doyle LE, Mills R. Something to chat about: an analysis of genetic counseling via asynchronous messaging following direct-to-consumer genetic testing. J Genet Couns 2023 Jun 02;32(3):717-727. [doi: 10.1002/jgc4.1683] [Medline: 36732942]
29. Hall JA, Gertz R, Amato J, Pagliari C. Transparency of genetic testing services for 'health, wellness and lifestyle': analysis of online prepurchase information for UK consumers. Eur J Hum Genet 2017 Aug;25(8):908-917 [FREE Full text] [doi: 10.1038/ejhg.2017.75] [Medline: 28703222]
30. Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 2011 Jun 17;12(7):499-510. [doi: 10.1038/nrg3012] [Medline: 21681211]
31. Kemp Z, Turnbull A, Yost S, Seal S, Mahamdallie S, Poyastro-Pearson E, et al. Evaluation of cancer-based criteria for use in mainstream BRCA1 and BRCA2 genetic testing in patients with breast cancer. JAMA Netw Open 2019 May 03;2(5):e194428 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.4428] [Medline: 31125106]
32. Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nat Rev Genet 2013 Oct 3;14(10):681-691. [doi: 10.1038/nrg3555] [Medline: 23999272]
33. Zimmermann B, Hill M, Gemelos G, Demko Z, Banjevic M, Baner J, et al. Noninvasive prenatal aneuploidy testing of chromosomes 13, 18, 21, X, and Y, using targeted sequencing of polymorphic loci. Prenat Diagn 2012 Dec 30;32(13):1233-1241 [FREE Full text] [doi: 10.1002/pd.3993] [Medline: 23108718]

34. Ed-daoudy A, Maalmi K. A new internet of things architecture for real-time prediction of various diseases using machine learning on big data environment. J Big Data 2019 Nov 27;6:104. [doi: 10.1186/s40537-019-0271-7]
35. Steele AJ, Denaxas SC, Shah AD, Hemingway H, Luscombe NM. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. PLoS One 2018 Aug 31;13(8):e0202344 [FREE Full text] [doi: 10.1371/journal.pone.0202344] [Medline: 30169498]
36. Yadav AK, Verma D, Kumar A, Kumar P, Solanki PR. The perspectives of biomarker-based electrochemical immunosensors, artificial intelligence and the internet of medical things toward COVID-19 diagnosis and management. Mater Today Chem 2021 Jun;20:100443 [FREE Full text] [doi: 10.1016/j.mtchem.2021.100443] [Medline: 33615086]
37. Wee HY, Woo BH, Yeo GS. Isolated oral clefts (OC) detected by routine ultrasonography – is it associated with chromosomal abnormalities? Ultrasound Obstet Gynecol 2002 Apr 30;18(s1):F59. [doi: 10.1046/j.1469-0705.2001.abs23-5.x]
38. Wang X, Zheng B, Wood M, Li S, Chen W, Liu H. Development and evaluation of automated systems for detection and classification of banded chromosomes: current status and future perspectives. J Phys D Appl Phys 2005 Jul 22;38(15):2536-2542. [doi: 10.1088/0022-3727/38/15/003]
39. Satam H, Joshi K, Mangrolia U, Waghoo S, Zaidi G, Rawool S, et al. Next-generation sequencing technology: current trends and advancements. Biology (Basel) 2023 Jul 13;12(7):997 [FREE Full text] [doi: 10.3390/biology12070997] [Medline: 37508427]
40. Kumar KR, Cowley MJ, Davis RL. Next-generation sequencing and emerging technologies. Semin Thromb Hemost 2024 Oct 01;50(7):1026-1038. [doi: 10.1055/s-0044-1786397] [Medline: 38692283]
41. Dicks LM. Gut bacteria provide genetic and molecular reporter systems to identify specific diseases. Int J Mol Sci 2024 Apr 17;25(8):4431 [FREE Full text] [doi: 10.3390/ijms25084431] [Medline: 38674014]
42. Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, et al. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res 2007;35(6):2013-2025 [FREE Full text] [doi: 10.1093/nar/gkm076] [Medline: 17341461]
43. Fernandes TG, Diogo MM, Clark DS, Dordick JS, Cabral JM. High-throughput cellular microarray platforms: applications in drug discovery, toxicology and stem cell research. Trends Biotechnol 2009 Jun;27(6):342-349 [FREE Full text] [doi: 10.1016/j.tibtech.2009.02.009] [Medline: 19398140]
44. Odibo IN, Wendel PJ, Magann EF. Telemedicine in obstetrics. Clin Obstet Gynecol 2013 Sep;56(3):422-433. [doi: 10.1097/GRF.0b013e318290fef0] [Medline: 23903374]
45. Rene R, Cherson M, Rannazzisi A, Felter J, Silverio A, Cunningham AT. Transitioning from in-person to telemedicine within primary care behavioral health during COVID-19. Popul Health Manag 2022 Aug 01;25(4):455-461. [doi: 10.1089/pop.2021.0292] [Medline: 35196130]
46. Butzner M, Cuffee Y. Telehealth interventions and outcomes across rural communities in the United States: narrative review. J Med Internet Res 2021 Aug 26;23(8):e29575 [FREE Full text] [doi: 10.2196/29575] [Medline: 34435965]
47. Kumar Y, Koul A, Singla R, Ijaz MF. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. J Ambient Intell Humaniz Comput 2023 Jan 13;14(7):8459-8486 [FREE Full text] [doi: 10.1007/s12652-021-03612-z] [Medline: 35039756]
48. De Jong KA, Spears WM, Gordon DF. Using genetic algorithms for concept learning. Mach Learn 1993 Nov;13(2-3):161-188. [doi: 10.1007/bf00993042]
49. Pipal D, Pipal V, Pipal R, Yadav S. Application of artificial intelligence and digitalization in medical education and various surgical specialities: concerns and prospects. In: Ordóñez de Pablos P, editor. Digital Healthcare in Asia and Gulf Region for Healthy Aging and More Inclusive Societies. Cambridge, MA: Academic Press; 2024.
50. Qin S, Zhao Y, Deng F, Yan M, Xi N, Wei P, et al. Performance evaluation of noninvasive prenatal testing on 24 chromosomes in a cohort of 118,969 pregnant women in Sichuan, China. J Int Med Res 2024 Sep;52(9):03000605241274584. [doi: 10.1002/ajmg.c.32057] [Medline: 37507620]
51. Paredes N, Caicedo-Bravo EF, Bacca B, Olmedo G. Emotion recognition of down syndrome people based on the evaluation of artificial intelligence and statistical analysis methods. Symmetry 2022 Nov 24;14(12):2492. [doi: 10.3390/sym14122492]
52. Wang Y, Shao Y, Yu J. The clinical application and accuracy evaluation of noninvasive prenatal testing for common trisomy and sex chromosome aneuploidy. Discov Med 2023 Jun;35(176):353-360. [doi: 10.24976/Discov.Med.202335176.36] [Medline: 37272102]
53. Jones OT, Matin RN, van der Schaar M, Prathivadi Bhayankaram K, Ranmuthu CK, Islam MS, et al. Artificial intelligence and machine learning algorithms for early detection of skin cancer in community and primary care settings: a systematic review. The Lancet Digital Health 2022 Jun;4(6):e466-e476. [doi: 10.1016/s2589-7500(22)00023-1]
54. Boddupally K, Rani Thuraka E. Artificial intelligence for prenatal chromosome analysis. Clin Chim Acta 2024 Jan 01;552:117669. [doi: 10.1016/j.cca.2023.117669] [Medline: 38007058]
55. Huang X, Wang H, She C, Feng J, Liu X, Hu X, et al. Artificial intelligence promotes the diagnosis and screening of diabetic retinopathy. Front Endocrinol (Lausanne) 2022 Sep 29;13:946915 [FREE Full text] [doi: 10.3389/fendo.2022.946915] [Medline: 36246896]
56. Budach L, Feuerpfeil M, Ihde N, Nathansen A, Noack N, Patzlaff H, et al. The effects of data quality on machine learning performance. arXiv. Preprint posted online on July 29, 2022 [FREE Full text]

57. Aradhya S, Facio FM, Metz H, Manders T, Colavin A, Kobayashi Y, et al. Applications of artificial intelligence in clinical laboratory genomics. American J of Med Genetics Pt C 2023 Jul 28;193(3). [doi: 10.1002/ajmg.c.32057]
58. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to machine learning, neural networks, and deep learning. Transl Vis Sci Technol 2020 Feb 27;9(2):14 [FREE Full text] [doi: 10.1167/tvst.9.2.14] [Medline: 32704420]
59. Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism 2017 Apr;69S:S36-S40. [doi: 10.1016/j.metabol.2017.01.011] [Medline: 28126242]
60. Fanni SC, Febi M, Aghakhanyan G, Neri E. Natural language processing. In: Klontzas ME, Fanni SC, Neri E, editors. Introduction to Artificial Intelligence. Cham, Switzerland: Springer; 2023.
61. Kondurkar I, Raj A, Lakshmi D. Modern applications with a focus on training ChatGPT and GPT models: exploring generative AI and NLP. In: Obaid AJ, Bhushan B, S M, Rajest SS, editors. Advanced Applications of Generative AI and Natural Language Processing Models. Hershey, PA: IGI Global; 2023:186-227.
62. Kaswan KS, Gaur L, Dhatterwal JS, Kumar R. AI-based natural language processing for the generation of meaningful information electronic health record (EHR) data. In: Advanced AI Techniques and Applications in Bioinformatics. Boca Raton, FL: CRC Press; 2021.
63. Lo Barco T, Kuchenbuch M, Garcelon N, Neuraz A, Nabbout R. Improving early diagnosis of rare diseases using Natural Language Processing in unstructured medical records: an illustration from Dravet syndrome. Orphanet J Rare Dis 2021 Jul 13;16(1):309 [FREE Full text] [doi: 10.1186/s13023-021-01936-9] [Medline: 34256808]
64. Rybinski M, Dai X, Singh S, Karimi S, Nguyen A. Extracting family history information from electronic health records: natural language processing analysis. JMIR Med Inform 2021 Apr 30;9(4):e24020 [FREE Full text] [doi: 10.2196/24020] [Medline: 33664015]
65. 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]
66. Koleck TA, Dreisbach C, Bourne PE, Bakken S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. J Am Med Inform Assoc 2019 Apr 01;26(4):364-379 [FREE Full text] [doi: 10.1093/jamia/ocy173] [Medline: 30726935]
67. Mary SA, Srividya K. NLP-driven chatbots: applications and implications in conversational AI. In: Rawat R, Chakrawarti RK, Sarangi SK, Vyas P, Alamanda MS, Srividya K, et al, editors. Conversational Artificial Intelligence. Beverly, MA: Scrivener Publishing LLC; 2024.
68. Iscen A, Fathi A, Schmid C. Improving image recognition by retrieving from web-scale image-text data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023 Presented at: CVPR 2023; June 17-24, 2023; Vancouver, BC.
69. Langmead B, Nellore A. Cloud computing for genomic data analysis and collaboration. Nat Rev Genet 2018 Apr 30;19(4):208-219 [FREE Full text] [doi: 10.1038/nrg.2017.113] [Medline: 29379135]
70. Prasanna A, Pooja R, Suchithra V, Ravikumar A, Niranjan V. Cloud based solutions for genome informatics: challenges and applications. Mater Today Proc 2018;5(4):10652-10659. [doi: 10.1016/j.matpr.2017.12.340]
71. Senbekov M, Saliev T, Bukeyeva Z, Almabayeva A, Zhanaliyeva M, Aitenova N, et al. The recent progress and applications of digital technologies in healthcare: a review. Int J Telemed Appl 2020;2020:8830200 [FREE Full text] [doi: 10.1155/2020/8830200] [Medline: 33343657]
72. Becich M. Advancing pathology informatics, imaging, and the internet (APIII 2002): scientific session and e-poster session abstracts. Arch Pathol Lab Med 2003;127(7):789-813. [doi: 10.5858/2003-127-789-APIIAT]
73. Markham JF, Yerneni S, Ryland GL, Leong HS, Fellowes A, Thompson ER, et al. CNspector: a web-based tool for visualisation and clinical diagnosis of copy number variation from next generation sequencing. Sci Rep 2019 Apr 23;9(1):6426 [FREE Full text] [doi: 10.1038/s41598-019-42858-8] [Medline: 31015508]
74. Ghaheri A, Shoar S, Naderan M, Hoseini SS. The applications of genetic algorithms in medicine. Oman Med J 2015 Nov 10;30(6):406-416 [FREE Full text] [doi: 10.5001/omj.2015.82] [Medline: 26676060]
75. 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 [FREE Full text] [doi: 10.1126/scitranslmed.aat6177] [Medline: 31019026]
76. landoli L. Internet-based decision support systems: leveraging mass collaboration to address complex problems. J Inf Technol Case Appl Res 2014 Sep 12;11(4):1-10. [doi: 10.1080/15228053.2009.10856168]
77. Mercer MB, Agatisa PK, Farrell RM. What patients are reading about noninvasive prenatal testing: an evaluation of internet content and implications for patient-centered care. Prenat Diagn 2014 Oct 11;34(10):986-993. [doi: 10.1002/pd.4410] [Medline: 24823864]
78. Davidson L, Boland MR. Towards deep phenotyping pregnancy: a systematic review on artificial intelligence and machine learning methods to improve pregnancy outcomes. Brief Bioinform 2021 Sep 02;22(5):bbaa369 [FREE Full text] [doi: 10.1093/bib/bbaa369] [Medline: 33406530]

79. Shah V, Konda SR. Neural networks and explainable AI: bridging the gap between models and interpretability. Int J Comput Sci Technol 2021;5(2):163-176. [doi: 10.5281/zenodo.10779335]
80. Hennen L, Sauter A, Van Den Cruyce E. Direct to consumer genetic testing: insights from an internet scan. New Genet Soc 2010 Jun 21;29(2):167-186. [doi: 10.1080/14636778.2010.484232]
81. Catic A, Gurbeta L, Kurtovic-Kozaric A, Mehmedbasic S, Badnjevic A. Application of neural networks for classification of Patau, Edwards, Down, Turner and Klinefelter Syndrome based on first trimester maternal serum screening data, ultrasonographic findings and patient demographics. BMC Med Genomics 2018 Feb 13;11(1):19 [FREE Full text] [doi: 10.1186/s12920-018-0333-2] [Medline: 29439729]
82. Piazza R, Ramazzotti D, Spinelli R, Pirola A, De Sano L, Ferrari P, et al. OncoScore: a novel, internet-based tool to assess the oncogenic potential of genes. Sci Rep 2017 Apr 07;7:46290 [FREE Full text] [doi: 10.1038/srep46290] [Medline: 28387367]
83. Girard N, Lou E, Azzoli CG, Reddy R, Robson M, Harlan M, et al. Analysis of genetic variants in never-smokers with lung cancer facilitated by an internet-based blood collection protocol: a preliminary report. Clin Cancer Res 2010 Jan 15;16(2):755-763 [FREE Full text] [doi: 10.1158/1078-0432.CCR-09-2437] [Medline: 20068085]
84. Kute SS, Tyagi AK, Malik S, Deshmukh A. Internet-based healthcare things driven deep learning algorithm for detection and classification of cervical cells. In: Proceedings of the Congress on Intelligent Systems. 2022 Presented at: CIS 2022; September 4-5, 2021; Bengaluru, India. [doi: 10.1007/978-981-16-9416-5_19]
85. Shin SJ, You SC, Park YR, Roh J, Kim JH, Haam S, et al. Genomic common data model for seamless interoperation of biomedical data in clinical practice: retrospective study. J Med Internet Res 2019 Mar 26;21(3):e13249 [FREE Full text] [doi: 10.2196/13249] [Medline: 30912749]
86. Kondylakis H, Axenie C, Kiran Bastola D, Katehakis DG, Kouroubali A, Kurz D, et al. Status and recommendations of technological and data-driven innovations in cancer care: focus group study. J Med Internet Res 2020 Dec 15;22(12):e22034 [FREE Full text] [doi: 10.2196/22034] [Medline: 33320099]
87. Covolo L, Rubinelli S, Ceretti E, Gelatti U. Internet-based direct-to-consumer genetic testing: a systematic review. J Med Internet Res 2015 Dec 14;17(12):e279 [FREE Full text] [doi: 10.2196/jmir.4378] [Medline: 26677835]
88. Gur RE, Bassett AS, McDonald-McGinn DM, Bearden CE, Chow E, Emanuel BS, et al. A neurogenetic model for the study of schizophrenia spectrum disorders: the International 22q11.2 Deletion Syndrome Brain Behavior Consortium. Mol Psychiatry 2017 Dec 01;22(12):1664-1672 [FREE Full text] [doi: 10.1038/mp.2017.161] [Medline: 28761081]
89. Solomon BD, Nguyen AD, Bear KA, Wolfsberg TG. Clinical genomic database. Proc Natl Acad Sci U S A 2013 Jun 11;110(24):9851-9855 [FREE Full text] [doi: 10.1073/pnas.1302575110] [Medline: 23696674]
90. Trujillano D, Oprea GE, Schmitz Y, Bertoli-Avella AM, Abou Jamra R, Rolfs A. A comprehensive global genotype-phenotype database for rare diseases. Mol Genet Genomic Med 2017 Jan 23;5(1):66-75 [FREE Full text] [doi: 10.1002/mgg3.262] [Medline: 28116331]
91. DECIPHER homepage. DECIPHER. URL: https://decipher.sanger.ac.uk/ [accessed 2024-10-02]
92. Zelenova MA, Iourov IY. Possibilities and limitations of CNV interpretation software and algorithms in homo sapiens. Curr Bioinform 2022 Dec;17(10):883-887. [doi: 10.2174/1574893617666220907121155]
93. Hodes-Wertz B, Druckenmiller S, Smith M, Noyes N. What do reproductive-age women who undergo oocyte cryopreservation think about the process as a means to preserve fertility? Fertil Steril 2013 Nov;100(5):1343-1349 [FREE Full text] [doi: 10.1016/j.fertnstert.2013.07.201] [Medline: 23953326]
94. Luke B. Pregnancy and birth outcomes in couples with infertility with and without assisted reproductive technology: with an emphasis on US population-based studies. Am J Obstet Gynecol 2017 Sep;217(3):270-281 [FREE Full text] [doi: 10.1016/j.ajog.2017.03.012] [Medline: 28322775]
95. Abdelhamied MA, Fathy Elbeltagy ES, Osman NE, Ahmed RT, Saber EH. Effect of social media-based intervention on adjustment, anxiety and pregnancy rate among infertile women undergoing assisted reproductive technology. Assiut Sci Nurs J 2023 Jan 01;11(34):129-143. [doi: 10.21608/ASNJ.2023.187708.1490]
96. Bornstein E, Berger S, Cheung SW, Maliszewski KT, Patel A, Pursley AN, et al. Universal prenatal chromosomal microarray analysis: additive value and clinical dilemmas in fetuses with a normal karyotype. Am J Perinatol 2017 Mar 17;34(4):340-348. [doi: 10.1055/s-0036-1586501] [Medline: 27533100]
97. Jost E, Kosian P, Jimenez Cruz J, Albarqouni S, Gembruch U, Strizek B, et al. Evolving the era of 5D ultrasound? A systematic literature review on the applications for artificial intelligence ultrasound imaging in obstetrics and gynecology. J Clin Med 2023 Oct 29;12(21):6833 [FREE Full text] [doi: 10.3390/jcm12216833] [Medline: 37959298]
98. Skrzypek H, Hui L. Noninvasive prenatal testing for fetal aneuploidy and single gene disorders. Best Pract Res Clin Obstet Gynaecol 2017 Jul;42:26-38. [doi: 10.1016/j.bpobgyn.2017.02.007] [Medline: 28342726]
99. Ates HC, Brunauer A, von Stetten F, Urban GA, Güder F, Merkoçi A, et al. Integrated devices for non‐invasive diagnostics. Adv Funct Materials 2021 Jan 04;31(15):2010388. [doi: 10.1002/adfm.202010388]
100. Ferrara F, Zoupanou S, Primiceri E, Ali Z, Chiriacò MS. Beyond liquid biopsy: toward non-invasive assays for distanced cancer diagnostics in pandemics. Biosens Bioelectron 2022 Jan 15;196:113698 [FREE Full text] [doi: 10.1016/j.bios.2021.113698] [Medline: 34688113]

101. Liao C, Chen X, Fu Y. Salivary analysis: an emerging paradigm for non‐invasive healthcare diagnosis and monitoring. Interdiscip Med 2023 Jun 02;1(3):e20230009. [doi: 10.1002/inmd.20230009]
102. Wen KY, Dayaratna S, Slamon R, Granda-Cameron C, Tagai EK, Kohler RE, et al. Chatbot-interfaced and cognitive-affective barrier-driven messages to improve colposcopy adherence after abnormal Pap test results in underserved urban women: a feasibility pilot study. Transl Behav Med 2024 Jan 11;14(1):1-12. [doi: 10.1093/tbm/ibad064] [Medline: 38014626]
103. Rebbeck TR, Bridges JF, Mack JW, Gray SW, Trent JM, George S, et al. A framework for promoting diversity, equity, and inclusion in genetics and genomics research. JAMA Health Forum 2022 Apr 15;3(4):e220603 [FREE Full text] [doi: 10.1001/jamahealthforum.2022.0603] [Medline: 35755401]
104. Painter CA, Jain E, Tomson BN, Dunphy M, Stoddard RE, Thomas BS, et al. The Angiosarcoma Project: enabling genomic and clinical discoveries in a rare cancer through patient-partnered research. Nat Med 2020 Feb 10;26(2):181-187. [doi: 10.1038/s41591-019-0749-z] [Medline: 32042194]
105. Talboom JS, De Both MD, Naymik MA, Schmidt AM, Lewis CR, Jepsen WM, et al. Two separate, large cohorts reveal potential modifiers of age-associated variation in visual reaction time performance. NPJ Aging Mech Dis 2021 Jul 01;7(1):14 [FREE Full text] [doi: 10.1038/s41514-021-00067-6] [Medline: 34210964]
106. Kitsios F, Stefanakakis S, Kamariotou M, Dermentzoglou L. Digital service platform and innovation in healthcare: measuring users' satisfaction and implications. Electronics 2023 Jan 28;12(3):662. [doi: 10.3390/electronics12030662]
107. Zan S, Agboola S, Moore SA, Parks KA, Kvedar JC, Jethwani K. Patient engagement with a mobile web-based telemonitoring system for heart failure self-management: a pilot study. JMIR Mhealth Uhealth 2015 Apr 01;3(2):e33 [FREE Full text] [doi: 10.2196/mhealth.3789] [Medline: 25842282]
108. Peñate GP, Parra NO, Morera JA, Meñaca AM, Ramón ML, Menéndez SC, et al. Assessing the clinical benefit, safety, and patient-reported outcomes with the use of the PAHcare™ digital platform in pulmonary arterial hypertension: a pilot study. Front Public Health 2024 Feb 16;12:1335072 [FREE Full text] [doi: 10.3389/fpubh.2024.1335072] [Medline: 38435295]
109. Rodriguez RA, Sparks MA, Conway PT, Gavhane A, Reddy S, Awdishu L, et al. American board of internal medicine nephrology procedure requirements for initial certification: time for a change and pursuing consensus in the nephrology community. Am J Kidney Dis 2024 Oct;84(4):507-515. [doi: 10.1053/j.ajkd.2024.03.014] [Medline: 38640993]
110. Dhillon RK, Hillman SC, Morris RK, McMullan D, Williams D, Coomarasamy A, et al. Additional information from chromosomal microarray analysis (CMA) over conventional karyotyping when diagnosing chromosomal abnormalities in miscarriage: a systematic review and meta-analysis. BJOG 2014 Jan;121(1):11-21. [doi: 10.1111/1471-0528.12382] [Medline: 23859082]
111. Chai H, DiAdamo A, Grommisch B, Xu F, Zhou Q, Wen J, et al. A retrospective analysis of 10-year data assessed the diagnostic accuracy and efficacy of cytogenomic abnormalities in current prenatal and pediatric settings. Front Genet 2019 Nov 20;10:1162 [FREE Full text] [doi: 10.3389/fgene.2019.01162] [Medline: 31850057]
112. Choi J, Jeon S, Kim D, Chua M, Do S. A scalable artificial intelligence platform that automatically finds copy number variations (CNVs) in journal articles and transforms them into a database: CNV extraction, transformation, and loading AI (CNV-ETLAI). Comput Biol Med 2022 May;144:105332. [doi: 10.1016/j.compbiomed.2022.105332] [Medline: 35240378]
113. Wadon ME, Winter M, Peall MJ. Internet-based cognitive behavioural therapy programme as an intervention for people diagnosed with adult-onset, focal, isolated, idiopathic cervical dystonia: a feasibility study protocol. Pilot Feasibility Stud 2020;6:100 [FREE Full text] [doi: 10.1186/s40814-020-00641-x] [Medline: 32685184]
114. Genoff Garzon MC, Rubin LR, Lobel M, Stelling J, Pastore LM. Review of patient decision-making factors and attitudes regarding preimplantation genetic diagnosis. Clin Genet 2018 Jul 11;94(1):22-42. [doi: 10.1111/cge.13174] [Medline: 29120067]
115. Cheng HH, Sokolova AO, Gulati R, Bowen D, Knerr SA, Klemfuss N, et al. Internet-based germline genetic testing for men with metastatic prostate cancer. JCO Precis Oncol 2023 Jan;7:e2200104. [doi: 10.1200/PO.22.00104] [Medline: 36623239]
116. Tiller J, Lacaze P. Regulation of internet-based genetic testing: challenges for Australia and other jurisdictions. Front Public Health 2018;6:24 [FREE Full text] [doi: 10.3389/fpubh.2018.00024] [Medline: 29497607]
117. Gonzalez T, Tucker K, Wakefield CE, Geelan-Small P, Macmillan S, Taylor N, et al. Comparing cancer genetic counselling using telegenetics with in-person and telephone appointments: results of a partially randomised patient-preference pilot study. J Telemed Telecare 2022 Jul 14;30(6):949-960. [doi: 10.1177/1357633x221112556]
118. Otten E, Birnie E, Ranchor AV, van Langen IM. Telegenetics use in presymptomatic genetic counselling: patient evaluations on satisfaction and quality of care. Eur J Hum Genet 2016 Apr 15;24(4):513-520 [FREE Full text] [doi: 10.1038/ejhg.2015.164] [Medline: 26173963]
119. Goddard KA, Robitaille J, Dowling NF, Parrado AR, Fishman J, Bradley LA, et al. Health-related direct-to-consumer genetic tests: a public health assessment and analysis of practices related to internet-based tests for risk of thrombosis. Public Health Genomics 2009 Nov 26;12(2):92-104 [FREE Full text] [doi: 10.1159/000176794] [Medline: 19039253]
120. Dhingra S, Raut R, Gunasekaran A, Rao Naik BK, Masuna V. Analysis of the challenges for blockchain technology adoption in the Indian health-care sector. J Model Manag 2023 Jul 12;19(2):375-406. [doi: 10.1108/JM2-09-2022-0229]
121. Masum H, Lackman R, Bartleson K. Developing global health technology standards: what can other industries teach us? Global Health 2013 Oct 17;9(1):49 [FREE Full text] [doi: 10.1186/1744-8603-9-49] [Medline: 24134866]

122. Baker SB, Xiang W, Atkinson I. Internet of things for smart healthcare: technologies, challenges, and opportunities. IEEE Access 2017;5:26521-26544. [doi: 10.1109/access.2017.2775180]
123. Maxim LD, Niebo R, Utell MJ. Screening tests: a review with examples. Inhal Toxicol 2014 Nov 29;26(13):811-828 [FREE Full text] [doi: 10.3109/08958378.2014.955932] [Medline: 25264934]

Abbreviations

AI: artificial intelligence
BRCA: Breast cancer gene 1
cfDNA: cell-free DNA
CNV: copy number variation
IVF: in vitro fertilization
NGS: next-generation sequencing
NIPT: noninvasive prenatal testing
NLP: natural language processing
SNP: single-nucleotide polymorphism

Edited by S Hacking; submitted 15.03.24; peer-reviewed by L Guo, A Glotov; comments to author 11.06.24; revised version received 13.06.24; accepted 18.08.24; published 16.10.24.

Please cite as:
Oyovwi MOS, Ohwin EP, Rotu RA, Olowe TG
Internet-Based Abnormal Chromosomal Diagnosis During Pregnancy Using a Noninvasive Innovative Approach to Detecting Chromosomal Abnormalities in the Fetus: Scoping Review
JMIR Bioinform Biotech 2024;5:e58439
URL: https://bioinform.jmir.org/2024/1/e58439
doi:10.2196/58439
PMID:39412876

Viewpoint

It Is in Our DNA: Bringing Electronic Health Records and Genomic Data Together for Precision Medicine

Alan J Robertson[1,2,3,4], BSc, MPhil; Andrew J Mallett[5,6,7], MBBS, MMed, PhD; Zornitza Stark[8,9,10], BMBS, MBS, MA, DM; Clair Sullivan[3,11,12], MBBS, MD

[1]Faculty of Medicine, University of Queensland, Hertson, Australia
[2]Medical Genomics Group, QIMR Berghofer Medical Research Institute, Brisbane, Australia
[3]Queensland Digital Health Centre, University of Queensland, Brisbane, Australia
[4]The Genomic Institute, Department of Health, Queensland Government, Brisbane, Australia
[5]Department of Renal Medicine, Townsville University Hospital, Townsville, Australia
[6]College of Medicine and Dentistry, James Cook University, Townsville, Australia
[7]Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia
[8]Victorian Clinical Genetics Services, Murdoch Children's Research Institute, Melbourne, Australia
[9]Australian Genomics, Melbourne, Australia
[10]University of Melbourne, Melbourne, Australia
[11]Centre for Health Services Research, Faculty of Medicine, University of Queensland, Woolloongabba, Australia
[12]Metro North Hospital and Health Service, Department of Health, Queensland Government, Brisbane, Australia

Corresponding Author:
Clair Sullivan, MBBS, MD
Queensland Digital Health Centre
University of Queensland
Health Sciences Building, Herston Campus
Royal Brisbane and Women's Hospital
Brisbane, 4029
Australia
Phone: 61 733465343
Email: c.sullivan1@uq.edu.au

Abstract

Health care is at a turning point. We are shifting from protocolized medicine to precision medicine, and digital health systems are facilitating this shift. By providing clinicians with detailed information for each patient and analytic support for decision-making at the point of care, digital health technologies are enabling a new era of precision medicine. Genomic data also provide clinicians with information that can improve the accuracy and timeliness of diagnosis, optimize prescribing, and target risk reduction strategies, all of which are key elements for precision medicine. However, genomic data are predominantly seen as diagnostic information and are not routinely integrated into the clinical workflows of electronic medical records. The use of genomic data holds significant potential for precision medicine; however, as genomic data are fundamentally different from the information collected during routine practice, special considerations are needed to use this information in a digital health setting. This paper outlines the potential of genomic data integration with electronic records, and how these data can enable precision medicine.

(JMIR Bioinform Biotech 2024;5:e55632) doi:10.2196/55632

KEYWORDS

genomics; digital health; genetics; precision medicine; genomic; genomic data; electronic health records; DNA; supports; decision-making; timeliness; diagnosis; risk reduction; electronic medical records

Introduction

Digital Health Care Systems Are Transforming Health Care

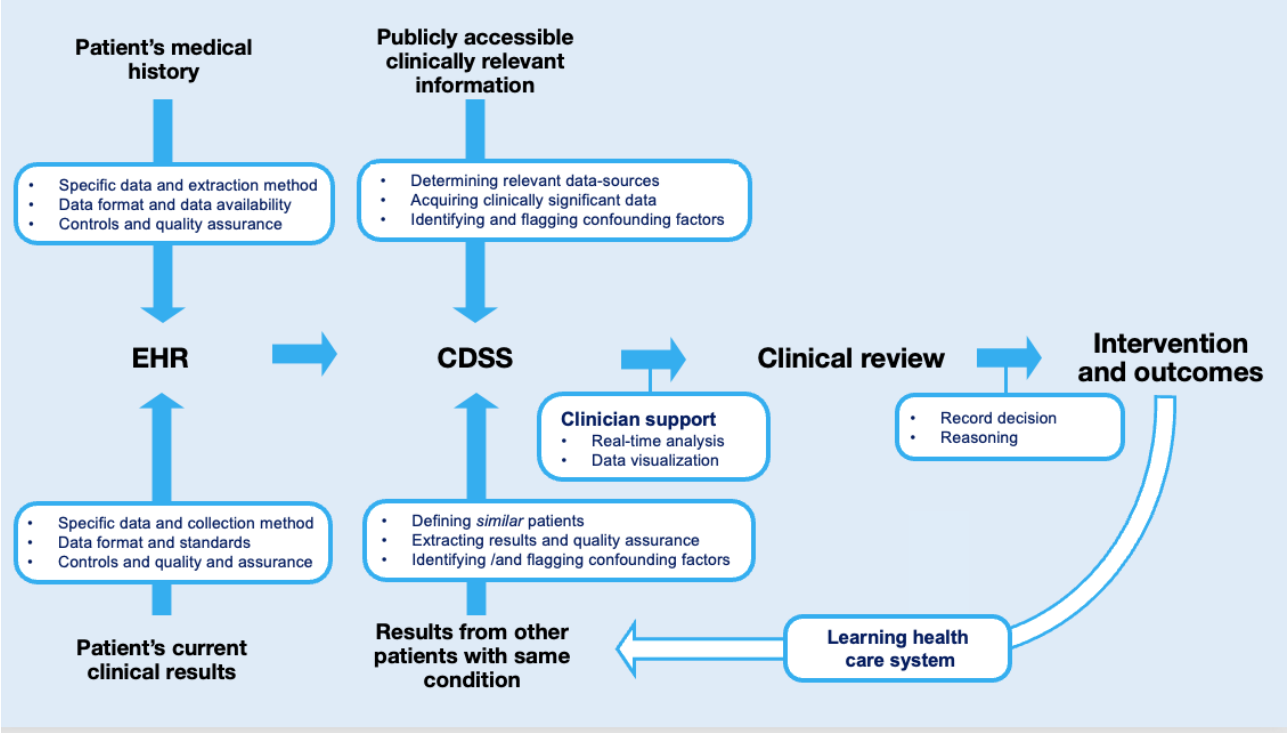
The adoption of electronic health records (EHRs) is transforming health care [1-4]. This digital infrastructure allows health services to store a patient's complete medical history and collect additional observations and results in real time. Having this information in a standardized, readily accessible format provides a foundation for clinical tools to analyze these data and provide clinicians with the information to make evidence-based decisions at the point of care [1,2,4].

EHRs are enabling health care to move from protocol-based medicine to precision medicine [5,6] and helping bring about the next generation of evidence-based practice. Critical to this transformation are the clinical decision support systems (CDSSs). CDSSs are electronic systems that use the information in an EHR to support the treatment of a specific disease or group of related diseases [7]. Using a patient's data in the EHR, a CDSS processes this information in real time and presents the results to clinicians, often with the context provided by the relevant clinical guidelines [7]. The clinician is then able to filter these outputs through the lens of their clinical experience, and the nuance of the scenario, to provide an individual with a precise intervention based on their unique physiology, medical history, and current situation (Figure 1).

CDSSs are usually carefully designed by groups of experts, undergo rigorous testing, and operate within strict governance structures. As a result, CDSSs have been shown to reduce medication errors and adverse clinical events [8]. By using the information in EHRs, CDSSs allow health care systems to move past models of practice designed for paper-based systems and enable new models of care that are better able to meet the quadruple aim of health care [9,10].

One exciting model of care, enabled by EHRs and CDSSs, is learning health care systems (LHSs). An LHS uses the data collected in routine clinical practice as evidence to determine the efficacy of an intervention. These learnings can then be used to inform clinicians treating patients with the same condition. An LHS shows how using the data routinely captured by an EHR in routine practice can be used to provide value to patients, clinicians, and the broader health care system [1,2,4]; however, for many health care systems, it is an aspirational goal (Figure 1).

Figure 1. A simplified overview of a patient's journey through a modern digitally enabled health care system, with an emphasis on the role of the EHR and CDSS. Each of the dot points linked to a solid blue arrow represents some of the specific decisions that must be made in order to integrate, analyze, and report information to clinicians. A single CDSS is not required to interact with every one of the data sources to provide clinical value but instead provide an example of some of the processes likely to occur. The white arrow represents the learning health care system, an aspirational goal for a digitally enabled health care system that uses the data collected in clinical practice as evidence for the treatment of patients afflicted with the same condition. CDSS: clinical decision support system; EHR: electronic health record.



Digital Health Systems Will Be Essential to Precision Medicine

Outside of LHSs, EHRs and CDSSs have the potential to facilitate a new paradigm in care—precision medicine [11,12]. Precision medicine refers to a tailored approach to care, guided by an individual's medical history, environment, and genetic makeup [13,14]. The structured information in an EHR and the tools to contextualize and present this information to clinicians at the point of care have been used to benefit patients across a range of different areas of health [15,16]. While the capacity for digital health systems to capture and return information

surrounding the patient's medical history is well established, genomic data are not routinely incorporated into CDSSs alongside traditional clinical data sources.

Genomic Data Are an Important Element of Precision Medicine

Genomic data are widely accepted to be a foundational component of precision medicine [13,14]. Identifying the molecular cause of a patient's condition can lead to tailored interventions [17], a better understanding of a patient's prognosis [18], and can help individuals make informed decisions in family planning [19]. The information in an individual's DNA is routinely being used to provide precision clinical care across a range of different areas (Table 1). A prime example of the potential of genomic information is oncology, where genomic testing is used to identify the range of mutations acquired by an individual's tumor, leading to tailored therapeutic interventions [20]. The management of infectious disease is another area that shows the potential of genomics in personalized medicine, as genome sequencing can be used to diagnose specific pathogen as well as determine the strain of the infectious agent as well as its antibiotic-resistance profile [21]. The information in an individual's DNA can have tremendous potential for many different areas of precision health care. However, for many clinicians in different areas of medicine, this information is only accessible by ordering a genomic test.

Table 1. Clinical applications of genomics.

Application	Description	References
Diagnosis of genetic disease	• While genetic testing has existed for decades, the use of next-generation sequencing has made it possible for clinicians to examine the entire genome, enabling faster and more accurate diagnosis for a broad range of rare disorders [22].	[19,22]
Disease screening and early detection	• Genomic technologies can be used to identify individuals who are at a higher risk for developing certain conditions. This information can be used to manage risk and, in some cases, make interventions before the disease begins to impact the individual's quality of life.	[19,23]
Family planning	• The identification of genetic variants that place an individual at a higher risk of developing a specific condition information can be used to make informed decisions in family planning and access to reproductive technologies.	[19,23]
Cancer diagnosis, treatment, and monitoring	• Cancer is a disease of the genome arising from mutations that have been acquired by an individual's DNA [24]. By comparing the DNA from a patient's tumor to their normal "germline" DNA, it is possible to identify the full spectrum of mutations in a tumor, including those driving disease progression. • While the SHIVA study highlighted the overenthusiasm many had for this approach [25], determining the specific mutations, driving a patient's disease, and selecting a treatment based on this information have proven to be an effective form of treatment for a range of different tumors. • Moreover, monitoring a patient's blood for the unique mutations associated with their tumor after treatment is a powerful way to monitor the progression of the disease, the effectiveness of an intervention, and if the disease is likely to reoccur [26].	[24-27]
Infectious disease diagnosis characterization	• Nucleic acids are used by all living organisms. By examining patient samples, for specific nucleic acid sequences that are not from the human genome, it is possible to find sequences that are indicative of certain pathogens. The application of genome sequencing methods here provides an accurate method to detect pathogens, and in some scenarios, this approach can be used to determine the strain and specific antibiotic resistance profile of an infectious agent. • As the genomes of many pathogens are significantly smaller than the human genome, it is possible to sequence large volumes of samples and screen them for pathogen DNA. The scalability of genomics in the monitoring of infectious diseases has been highlighted by the COVID-19 pandemic. Here, genomics was not only used to diagnose infection at a population scale but also to identify and track novel variants.	[21,28]
Precision treatment and pharmacogenomics	• Specific genetic variants can produce molecules that behave in different ways. Some variants can completely disrupt the function of a gene, while others can change how efficiently it performs its role. As a result, certain variants can impact the way certain individuals metabolize drugs. The identification of these variants and the use of information to guide treatment can ensure that each individual receives the best intervention for their unique physiology. • While only a small number of drugs are prescribed using this information, some have suggested that the metabolism of one-third of all drugs may be impacted by genetic variants.	[17,20,29-31]

Access to the Right Genomic Data Will Enable the Realization of Precision Medicine

Population studies have revealed that each individual's genome contains millions of different genetic variants [32]. The sheer number of variants means that it is unrealistic for a single specialist to keep track of the clinical significance of each of these variants across the range of diseases they examine. While genomic analyses would appear to be a prime candidate for the development of specialized CDSSs to support the use of genomic practice across a range of different areas of health (Table 1), CDSSs that routinely incorporate genetic information are rare [33,34]. There are likely many causes to this deficit; however, a significant factor to this can be attributed to the availability of interoperable genomic data within EHR. As a result, when many clinicians order genomic tests, the data are analyzed once, and the results are stored as a static PDF, locking the information away from future analyses.

Significant progress has been made in the development of systems to facilitate the use of genomic data in EHRs, such as clinical-grade genomic standards, file formats, and terminologies like Logical Observation Identifiers Names and Codes and Systematized Nomenclature of Medicine—Clinical Terms [35-38]. However, the adoption of these advances by EHR providers has been sluggish. As a result, EHRs are still struggling to store genomic data in a way that allows this information to be used by CDSSs. Without the capacity to access genomic data, clinicians are removed from an essential data source and will struggle to realize the full potential of precision medicine [12].

The reluctance to integrate genomic data into EHRs is likely due to a number of reasons. Some may suggest that the cause of this hesitation reflects the sheer volume and complexity of genomic data as well as the substantial amount of computer processing power and expertise required for genome analysis [39]. However, given the capacity of a VCF (variant call format) or VRS (variation representation) file to summarize the variants in a patient's genome in a relatively potable format, the hesitancy to adopt these standards could be attributed to the complex ethical or social or legal questions surrounding genomics [12,40].

Despite these challenges, there are 2 questions that must be addressed to build a foundation to integrate genomic data into an EHR and enable genomics-empowered precision medicine: determining the right data to store and determining the right structure of these data. These questions are unlikely to have simple answers, as the answers will reflect the specific clinical questions being asked. While it is tempting to compare the virtues of exome and genome sequencing, discuss the impact of emerging technologies, or highlight the potential to bring other types of "omics" data into the EHR, these conversations are out of scope for this viewpoint. To us, it is clear that clinicians, scientists, and administrators must answer these questions together to ensure that genomic data can provide value across a range of different areas of precision medicine in their unique health service.

Genomic Data Are New, Complex, and Different From Other Types of Health Data but Offer the Potential for New Models of Care

When determining *how* genomic data will be stored in an EHR, these conversations must address a unique attribute of genomic data—its (largely) static and unchanging nature. This attribute is typically brought up in discussions of secondary uses of genome data within the health care system [41]. However, a separate area of tremendous importance surrounds our evolving understanding of the clinical significance of a patient's genomic data [42], as our changing understanding of the clinical relevance of a patient's genetic data opens up new potential models of care.

The unchanging nature of a patient's DNA and a rapidly changing understanding of the importance of that data mean that if a patient did not receive a molecular diagnosis after genomic testing, reanalyzing the same information at a later date with the context of new discoveries and new techniques can produce new molecular diagnoses [43-45]. While discovery and changing understandings are not unique to genomics, in contrast to other fields, the *rate* and *volume* at which new genomic information is accumulating is so extraordinary that reinterpreting existing genomic data with the context provided by new discoveries is known to increase diagnostic yields [42].

Special considerations will be needed to harness the levels of change associated with genomic data when designing genomics-enabled EHRs and CDSSs. Moreover, they highlight the need for these digital solutions to alert laboratories and clinicians when clinically important information has changed and robust systems in place for clinicians and laboratories to be empowered to use this information (Textbox 1).

Textbox 1. A clinical vignette.

To contextualize the static nature of genome data and our changing understanding of that data, a patient aged 9 years may present to the clinic with the hallmark signs of a metabolic disorder. However, genomic testing might not confidently identify a causative pathogenic variant. Suppose the patient's existing genomic data are routinely reanalyzed when the patient reaches the age of 14 years. In that case, clinicians are able to take advantage of all the genes found to be associated with metabolism that have occurred in the last 5 years. This information could be used to inform the patient's treatment or potentially slow their decline. This example also highlights the potential for a "push" style approach, in which the clinician is alerted each time a gene associated with metabolism is discovered—ensuring that the patient can benefit from this new information as soon as it occurs.

Moving From Prescriptive to Precision Medicine

While there is still work to be done, the eventual widespread adoption of genomic-enabled EHRs will facilitate the move from a traditional, prescriptive approach to medicine to personalized models of care. However, this will require a change in the way we approach genomic testing.

Currently, genomic tests resemble a "pull-based" approach. In this approach, only the genes of interest are analyzed, and the additional information needed to contextualize a patient's

genetic variants is "pulled" from the literature or analysis resources once. While there is a movement away from this philosophy, the singular, request nature of this approach prevents patients and clinicians from benefiting from our rapidly evolving understanding of genetic variants.

An alternative approach would be to perform genome sequencing once and store this information with the view that it will be used across the range of interactions an individual would have with the health system throughout their lifetime (Table 2). This will be facilitated by storing the data in structured, secure, interoperable formats, with the assumption that these data will be aligned to newer reference genomes, analyzed with different variant callers, and compared to constantly evolving virtual gene panels. While the raw genomic data might not need to be directly accessible in the EHR, reliable access to genome data will support every future interaction with a precision medicine–enabled health care system.

In this model, a CDSS could be designed around a "push" model. In the event of an inconclusive test, changes in the amount of information associated with the condition can be automatically monitored, and when it passes a threshold, the EHR can alert both the patient and the clinician to the potential for reanalysis. Patients who receive a molecular diagnosis from genomic testing could still benefit from continued monitoring by a CDSS. For example, the CDSS could highlight novel treatment interventions based on new information, such as new, targeted pharmacogenomic recommendations and potential clinical trial opportunities.

Key to this approach is the accessibility of genomic data for CDSSs. To give CDSSs access in a safe and transparent manner, there are significant challenges to overcome. Some of these challenges will be addressed from a bioinformatics perspective; however, others will require a clinical or health informatics solution, and some others still will require a policy or multidisciplinary approach.

Table 2. Moving to a model of genomics-enabled precision medicine.

Activity	Genetic+genomic testing	
	Traditional practice	A potential model of genomics-enabled care
Generation of sequence data	• DNA from the genes associated with the condition is sequenced when a test is ordered	• Individual's whole genome sequence is available from a prior interaction with the health care system. • A CDSS[a] recommends if there is a benefit to generate complementary sequence data (eg, long read, transcriptomic, cell-free).
Analysis and interpretation of genetic data	• Variants within the sequenced DNA are determined • The clinical significance of the variants is accessed	• A CDSS accesses the specific genes currently associated with condition from multiple high-quality, peer-reviewed resources. • A CDSS recommends if genome data should be aligned to a new reference genome or use updated variant detection methods. • Variants within the selected genes are determined. • The clinical significance of the variants is accessed.
Clinical decisions and reporting	• Clinician synthesizes genetic results, patient's history, and clinical experience to make decision • A clinical report is generated • Report is uploaded to the EHR[b] as a PDF	• Clinician synthesizes genetic results, patient's history, and CDSS recommendations through the lens of their clinical experience to make decision. • The CDSS interacts with LIMs[c] and identifies any potential pharmacogenomic interventions or potential interactions. • A clinical report is generated. • Findings reported to patient and other clinicians (secure portal+PDF). • Report findings to EHR. • Flag that the test was successful or inconclusive. • If successful, share causative variants with public repositories and related individuals. Make results accessible to other clinicians treating the individual (where appropriate). • If inconclusive, flag candidate variants of uncertain significance for automatic monitoring, monitor information associated with disease, and determine when the individual should be reanalyzed.
Data storage	• Raw sequence data and results stored in the laboratory system • Note: external collaborators do not always provide raw-sequence data	• Store raw sequencing data, processed results, and variant interpretations in laboratory LIMs. • Store all clinically significant (and potentially significant) variants in EHR. • Ensure all information is in a standardized interoperable and time-stamped format (ie, GA4GH or eMerge).

[a]CDSS: clinical decision support system.

[b]EHR: electronic health record.

[c]LIM: Laboratory Information Management System.

Conclusions

The clinical potential of integrating genomics information with the range of clinically relevant data collected by an EHR has been long recognized as an important element for precision medicine [46]. However, the slow adoption of the standards needed to capture and use genomic data alongside the other information in the EHR is preventing the realization of this potential. Moreover, as genomic data associated with unique attributes are so different from other health care data, special considerations are needed to harness this potential when designing the systems. As many health care systems are revising their digital health strategies, there is an opportunity to address this oversight and guide the development of EHRs that are committed to determining and incorporating the right kinds of genomic data for their unique needs.

EHRs that have been designed to accommodate the unique attributes of genomic information will benefit patients, clinicians, and health services. These EHRs will enable the production of disease-specific, genomic-enabled CDSS applications, allow more clinicians to use genomic data in practice, and collect information that can be used to better characterize relationships between genotype and phenotype. Together these systems will support precision medicine, and also provide a framework to capture the efficacy of genomically informed treatments, for a next-generation, genomics-empowered LHS.

Authors' Contributions

AJR contributed to initial concept. All authors were involved in writing and editing the manuscript.

Conflicts of Interest

AJR is the founder and director of ClearSKY Genomics.

References

1. Poissant L, Pereira J, Tamblyn R, Kawasumi Y. The impact of electronic health records on time efficiency of physicians and nurses: a systematic review. J Am Med Inform Assoc 2005;12(5):505-516 [FREE Full text] [doi: 10.1197/jamia.M1700] [Medline: 15905487]
2. Jha AK, DesRoches CM, Campbell EG, Donelan K, Rao SR, Ferris TG, et al. Use of electronic health records in US hospitals. N Engl J Med 2009;360(16):1628-1638 [FREE Full text] [doi: 10.1056/NEJMsa0900592] [Medline: 19321858]
3. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary use of EHR: data quality issues and informatics opportunities. Summit Transl Bioinform 2010;2010:1-5. [Medline: 21347133]
4. Lim HC, Austin JA, van der Vegt AH, Rahimi AK, Canfell OJ, Mifsud J, et al. Toward a learning health care system: a systematic review and evidence-based conceptual framework for implementation of clinical analytics in a digital hospital. Appl Clin Inform 2022;13(2):339-354 [FREE Full text] [doi: 10.1055/s-0042-1743243] [Medline: 35388447]
5. Wei WQ, Denny JC. Extracting research-quality phenotypes from electronic health records to support precision medicine. Genome Med 2015;7(1):41. [doi: 10.1186/s13073-015-0166-y] [Medline: 25937834]
6. Abul-Husn NS, Kenny EE. Personalized medicine and the power of electronic health records. Cell 2019;177(1):58-69. [doi: 10.1016/j.cell.2019.02.039] [Medline: 30901549]
7. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3(1):17 [FREE Full text] [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
8. Aggarwal A, Aeran H, Rathee M. Quality management in healthcare: the pivotal desideratum. J Oral Biol Craniofac Res 2019;9(2):180-182. [doi: 10.1016/j.jobcr.2018.06.006] [Medline: 31211031]
9. Miles P, Hugman A, Ryan A, Landgren F, Liong G. Towards routine use of national electronic health records in Australian emergency departments. Med J Aust 2019;210(Suppl 6):S7-S9 [FREE Full text] [doi: 10.5694/mja2.50033] [Medline: 30927465]
10. Sullivan C, Staib A, Ayre S, Daly M, Collins R, Draheim M, et al. Pioneering digital disruption: Australia's first integrated digital tertiary hospital. Med J Aust 2016;205(9):386-389 [FREE Full text] [doi: 10.5694/mja16.00476] [Medline: 27809727]
11. Akhoon N. Precision medicine: a new paradigm in therapeutics. Int J Prev Med 2021;12(1):12. [doi: 10.4103/ijpvm.IJPVM_375_19] [Medline: 34084309]
12. Aronson SJ, Rehm HL. Building the foundation for genomics in precision medicine. Nature 2015;526(7573):336-342 [FREE Full text] [doi: 10.1038/nature15816] [Medline: 26469044]
13. Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med 2015;372(9):793-795 [FREE Full text] [doi: 10.1056/NEJMp1500523] [Medline: 25635347]
14. Ashley EA. Towards precision medicine. Nat Rev Genet 2016;17(9):507-522 [FREE Full text] [doi: 10.1038/nrg.2016.86] [Medline: 27528417]
15. Chang E, Mostafa J. The use of SNOMED CT, 2013-2020: a literature review. J Am Med Inform Assoc 2021;28(9):2017-2026 [FREE Full text] [doi: 10.1093/jamia/ocab084] [Medline: 34151978]

16. Reinecke I, Zoch M, Reich C, Sedlmayr M, Bathelt F. The usage of OHDSI OMOP—a scoping review. Stud Health Technol Inform 2021;283:95-103 [FREE Full text] [doi: 10.3233/SHTI210546] [Medline: 34545824]
17. Relling MV, Evans WE. Pharmacogenomics in the clinic. Nature 2015;526(7573):343-350 [FREE Full text] [doi: 10.1038/nature15817] [Medline: 26469045]
18. Reitz C. Genetic diagnosis and prognosis of Alzheimer's disease: challenges and opportunities. Expert Rev Mol Diagn 2015;15(3):339-348 [FREE Full text] [doi: 10.1586/14737159.2015.1002469] [Medline: 25634383]
19. Stark Z, Schofield D, Alam K, Wilson W, Mupfeki N, Macciocca I, et al. Prospective comparison of the cost-effectiveness of clinical whole-exome sequencing with that of usual care overwhelmingly supports early use and reimbursement. Genet Med 2017;19(8):867-874 [FREE Full text] [doi: 10.1038/gim.2016.221] [Medline: 28125081]
20. Morganti S, Tarantino P, Ferraro E, D'Amico P, Duso BA, Curigliano G. Next generation sequencing (NGS): a revolutionary technology in pharmacogenomics and personalized medicine in cancer. Adv Exp Med Biol 2019;1168:9-30 [FREE Full text] [doi: 10.1007/978-3-030-24100-1_2] [Medline: 31713162]
21. Cao MD, Ganesamoorthy D, Elliott AG, Zhang H, Cooper MA, Coin LJM. Streaming algorithms for identification of pathogens and antibiotic resistance potential from real-time MinION(TM) sequencing. Gigascience 2016;5(1):32 [FREE Full text] [doi: 10.1186/s13742-016-0137-2] [Medline: 27457073]
22. Grody WW. The transformation of medical genetics by clinical genomics: hubris meets humility. Genet Med 2019;21(9):1916-1926 [FREE Full text] [doi: 10.1038/s41436-019-0450-0] [Medline: 30842646]
23. Primiero CA, Finnane A, Yanes T, Peach B, Soyer HP, McInerney-Leo AM. Protocol to evaluate a pilot program to upskill clinicians in providing genetic testing for familial melanoma. PLoS One 2022;17(12):e0275926 [FREE Full text] [doi: 10.1371/journal.pone.0275926] [Medline: 36477719]
24. Hanahan D. Hallmarks of Cancer: new dimensions. Cancer Discov 2022;12(1):31-46 [FREE Full text] [doi: 10.1158/2159-8290.CD-21-1059] [Medline: 35022204]
25. Le Tourneau C, Kamal M, Trédan O, Delord JP, Campone M, Goncalves A, et al. Designs and challenges for personalized medicine studies in oncology: focus on the SHIVA trial. Target Oncol 2012;7(4):253-265 [FREE Full text] [doi: 10.1007/s11523-012-0237-6] [Medline: 23161020]
26. Dawson SJ, Tsui DWY, Murtaza M, Biggs H, Rueda OM, Chin SF, et al. Analysis of circulating tumor DNA to monitor metastatic breast cancer. N Engl J Med 2013;368(13):1199-1209 [FREE Full text] [doi: 10.1056/NEJMoa1213261] [Medline: 23484797]
27. Nakagawa H, Fujita M. Whole genome sequencing analysis for cancer genomics and precision medicine. Cancer Sci 2018;109(3):513-522 [FREE Full text] [doi: 10.1111/cas.13505] [Medline: 29345757]
28. Seemann T, Lane CR, Sherry NL, Duchene S, da Silva AG, Caly L, et al. Tracking the COVID-19 pandemic in Australia using genomics. Nat Commun 2020;11(1):4376 [FREE Full text] [doi: 10.1038/s41467-020-18314-x] [Medline: 32873808]
29. Weinshilboum RM, Wang L. Pharmacogenomics: precision medicine and drug response. Mayo Clin Proc 2017;92(11):1711-1722 [FREE Full text] [doi: 10.1016/j.mayocp.2017.09.001] [Medline: 29101939]
30. Vadlamudi L, Bennett CM, Tom M, Abdulrasool G, Brion K, Lundie B, et al. A multi-disciplinary team approach to genomic testing for drug-resistant epilepsy patients—the GENIE study. J Clin Med 2022;11(14):4238 [FREE Full text] [doi: 10.3390/jcm11144238] [Medline: 35888005]
31. Bielinski SJ, Olson JE, Pathak J, Weinshilboum RM, Wang L, Lyke KJ, et al. Preemptive genotyping for personalized medicine: design of the right drug, right dose, right time-using genomic data to individualize treatment protocol. Mayo Clin Proc 2014;89(1):25-33 [FREE Full text] [doi: 10.1016/j.mayocp.2013.10.021] [Medline: 24388019]
32. Ng PC, Levy S, Huang J, Stockwell TB, Walenz BP, Li K, et al. Genetic variation in an individual human exome. PLoS Genet 2008;4(8):e1000160 [FREE Full text] [doi: 10.1371/journal.pgen.1000160] [Medline: 18704161]
33. Freimuth RR, Formea CM, Hoffman JM, Matey E, Peterson JF, Boyce RD. Implementing genomic clinical decision support for drug-based precision medicine. CPT Pharmacometrics Syst Pharmacol 2017;6(3):153-155 [FREE Full text] [doi: 10.1002/psp4.12173] [Medline: 28109071]
34. Mattick JS, Dziadek MA, Terrill BN, Kaplan W, Spigelman AD, Bowling FG, et al. The impact of genomics on the future of medicine and health. Med J Aust 2014;201(1):17-20 [FREE Full text] [doi: 10.5694/mja13.10920] [Medline: 24999876]
35. Rehm HL, Page AJH, Smith L, Adams JB, Alterovitz G, Babb LJ, et al. GA4GH: international policies and standards for data sharing across genomic research and healthcare. Cell Genom 2021;1(2):100029 [FREE Full text] [doi: 10.1016/j.xgen.2021.100029] [Medline: 35072136]
36. Rasmussen-Torvik LJ, Stallings SC, Gordon AS, Almoguera B, Basford MA, Bielinski SJ, et al. Design and anticipated outcomes of the eMERGE-PGx project: a multicenter pilot for preemptive pharmacogenomics in electronic health record systems. Clin Pharmacol Ther 2014;96(4):482-489 [FREE Full text] [doi: 10.1038/clpt.2014.137] [Medline: 24960519]
37. Forrey AW, McDonald CJ, DeMoor G, Huff SM, Leavelle D, Leland D, et al. Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results. Clin Chem 1996;42(1):81-90. [Medline: 8565239]
38. El-Sappagh S, Franda F, Ali F, Kwak K. SNOMED CT standard ontology based on the ontology for general medical science. BMC Med Inform Decis Mak 2018;18(1):76 [FREE Full text] [doi: 10.1186/s12911-018-0651-5] [Medline: 30170591]

39. Krumm N, Hoffman N. Practical estimation of cloud storage costs for clinical genomic data. Pract Lab Med 2020;21:e00168 [FREE Full text] [doi: 10.1016/j.plabm.2020.e00168] [Medline: 32529017]
40. Hazin R, Brothers KB, Malin BA, Koenig BA, Sanderson SC, Rothstein MA, et al. Ethical, legal, and social implications of incorporating genomic information into electronic health records. Genet Med 2013;15(10):810-816 [FREE Full text] [doi: 10.1038/gim.2013.117] [Medline: 24030434]
41. Kahn SD. On the future of genomic data. Science 2011;331(6018):728-729 [FREE Full text] [doi: 10.1126/science.1197891] [Medline: 21311016]
42. Robertson AJ, Tan NB, Spurdle AB, Metke-Jimenez A, Sullivan C, Waddell N. Re-analysis of genomic data: an overview of the mechanisms and complexities of clinical adoption. Genet Med 2022;24(4):798-810 [FREE Full text] [doi: 10.1016/j.gim.2021.12.011] [Medline: 35065883]
43. Dai P, Honda A, Ewans L, McGaughran J, Burnett L, Law M, et al. Recommendations for next generation sequencing data reanalysis of unsolved cases with suspected mendelian disorders: a systematic review and meta-analysis. Genet Med 2022;24(8):1618-1629 [FREE Full text] [doi: 10.1016/j.gim.2022.04.021] [Medline: 35550369]
44. Robertson AJ, Tran K, Patel C, Sullivan C, Stark Z, Waddell N. Evolution of virtual gene panels over time and implications for genomic data re-analysis. Genet Med Open 2023;1(1):100820 [FREE Full text] [doi: 10.1016/j.gimo.2023.100820]
45. Tan NB, Stapleton R, Stark Z, Delatycki MB, Yeung A, Hunter MF, et al. Evaluating systematic reanalysis of clinical genomic data in rare disease from single center experience and literature review. Mol Genet Genomic Med 2020;8(11):e1508 [FREE Full text] [doi: 10.1002/mgg3.1508] [Medline: 32969205]
46. Kohane IS. Using electronic health records to drive discovery in disease genomics. Nat Rev Genet 2011;12(6):417-428 [FREE Full text] [doi: 10.1038/nrg2999] [Medline: 21587298]

Abbreviations

CDSS: clinical decision support system
EHR: electronic health record
LHS: learning health care system

Edited by E Uzun; submitted 19.12.23; peer-reviewed by S Meister, L Wang; comments to author 20.02.24; revised version received 08.03.24; accepted 09.04.24; published 13.06.24.

Please cite as:
Robertson AJ, Mallett AJ, Stark Z, Sullivan C
It Is in Our DNA: Bringing Electronic Health Records and Genomic Data Together for Precision Medicine
JMIR Bioinform Biotech 2024;5:e55632
URL: https://bioinform.jmir.org/2024/1/e55632
doi:10.2196/55632
PMID:38935958

Corrigenda and Addenda

Correction: Mutations of SARS-CoV-2 Structural Proteins in the Alpha, Beta, Gamma, and Delta Variants: Bioinformatics Analysis

Saima Rehman Khetran[1*], MPhil; Roma Mustafa[1*], DPhil

Department of Life Sciences, Sardar Bahadur Khan Women's University, Quetta, Pakistan

[*]all authors contributed equally

Corresponding Author:
Saima Rehman Khetran, MPhil
Department of Life Sciences
Sardar Bahadur Khan Women's University
Bawrery Road near Kidney Hospital Quetta
Quetta, 87300
Pakistan
Email: aspirantcss2022@gmail.com

Related Article:

Correction of: https://bioinform.jmir.org/2023/1/e43906

(JMIR Bioinform Biotech 2024;5:e64915) doi:10.2196/64915

In "Mutations of SARS-CoV-2 Structural Proteins in the Alpha, Beta, Gamma, and Delta Variants: Bioinformatics Analysis" (JMIR Bioinform Biotech 2023;4:e43906) the authors made one addition.

An additional citation [31] was added to the Results and Discussion Section, which previously appeared as:

> *Apart from these mutations, deletions at position 85-89 (Δ85-Δ89) in a Spanish isolate (MW715071) along with other unique mutations of S protein, such as V90T (in which valine is replaced by threonine at position 90), A93Y (in which alanine is replaced by tyrosine at position 93), and D138H (in which aspartic acid is replaced by histidine at position 138), were also observed (Multimedia Appendices 1 and 2).*

This has been changed as follows:

> *Apart from these mutations, deletions at position 85-89 (Δ85-Δ89) in a Spanish isolate (MW715071) along with other unique mutations of S protein, such as V90T (in which valine is replaced by threonine at position 90) [31], A93Y (in which alanine is replaced by tyrosine at position 93), and D138H (in which aspartic acid is replaced by histidine at position 138), were also observed (Multimedia Appendices 1 and 2).*

The reference being included will be added to the References section, resulting in the renumeration of all references following Reference 31. The reference being added is the following:

> *31. Stojanov D. Phylogenicity of B.1.1.7 surface glycoprotein, novel distance function and first report of V90T missense mutation in SARS-CoV-2 surface glycoprotein. Meta Gene. 2021;30:100967. doi:https://doi.org/10.1016/j.mgene.2021.100967*

The correction will appear in the online version of the paper on the JMIR Publications website on August 5, 2024, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 30.07.24; this is a non–peer-reviewed article;accepted 31.07.24; published 05.08.24.

Please cite as:
Khetran SR, Mustafa R
Correction: Mutations of SARS-CoV-2 Structural Proteins in the Alpha, Beta, Gamma, and Delta Variants: Bioinformatics Analysis
JMIR Bioinform Biotech 2024;5:e64915
URL: https://bioinform.jmir.org/2024/1/e64915
doi:10.2196/64915
PMID:39102687

Editorial

ChatGPT and Medicine: Together We Embrace the AI Renaissance

Sean Hacking[1], MBBCh
NYU Langone, New York, NY, United States

Corresponding Author:
Sean Hacking, MBBCh
NYU Langone
Tisch Hospital, 560 First Avenue
Suite TH 461
New York, NY, 10016
United States
Phone: 1 6466836133
Email: hackingsean1@gmail.com

Abstract

The generative artificial intelligence (AI) model ChatGPT holds transformative prospects in medicine. The development of such models has signaled the beginning of a new era where complex biological data can be made more accessible and interpretable. ChatGPT is a natural language processing tool that can process, interpret, and summarize vast data sets. It can serve as a digital assistant for physicians and researchers, aiding in integrating medical imaging data with other multiomics data and facilitating the understanding of complex biological systems. The physician's and AI's viewpoints emphasize the value of such AI models in medicine, providing tangible examples of how this could enhance patient care. The editorial also discusses the rise of generative AI, highlighting its substantial impact in democratizing AI applications for modern medicine. While AI may not supersede health care professionals, practitioners incorporating AI into their practices could potentially have a competitive edge.

(JMIR Bioinform Biotech 2024;5:e52700) doi:10.2196/52700

KEYWORDS

ChatGPT; generative AI; NLP; medicine; bioinformatics; AI democratization; AI renaissance; artificial intelligence; natural language processing

Introduction

The arrival of OpenAI's model ChatGPT [1] invites us into a new era of medicine, where together we can make artificial intelligence (AI) more approachable to a wider audience. Such models stand as a testament to the remarkable progress in AI, machine learning, and natural language processing (NLP), offering substantial potential in processing and understanding complex information, and extending its applicability to the field of medicine. In this editorial, we delve into how multimodal large language models can help researchers and physicians manage and interpret vast amounts of patient data more effectively, and thus, widen its reach in medicine. From interpreting and summarizing the results of intricate genetic analyses to aiding in the design of novel experiments, such models could hold tremendous value in health care [2].

As an AI model, ChatGPT also provides its perspective on the subject, discussing how its language comprehension and data processing capabilities could contribute to the handling of complex data sets, the identification of patterns within interaction networks, the integration of multiomics data, and the development of predictive models for disease risk and treatment response. ChatGPT could also serve as a digital assistant to doctors, providing faster access to relevant medical information and associated literature along with improved bedside manner [3].

AI is undergoing a functional rebirth into a collaborative tool, working in tandem with humanity to redefine fundamental human qualities such as cognition and creativity. By exploring the potential of AI, we gain a renewed perspective on value. This technology not only offers transformative insights that can reshape the field of medicine but also plays a pivotal role in advancing human knowledge, understanding, and performance.

Viewpoint of the Physician

As a physician specializing in surgical pathology, it often feels like I am trying to navigate a vast ocean of information with conventional tools ill-suited to the task. The advent of AI models

like ChatGPT promises to revolutionize how we manage and interpret health care data.

For example, consider a hypothetical scenario involving a surgical pathology case where a patient presents with a mass diagnosed as colonic adenocarcinoma. Often, specifics of the diagnostic workup (including biomarker reporting), appropriate surgical/oncological treatments, and recommended follow-up intervals for such types of diagnoses might be concealed within the latest medical publications or obscured amid the vast intricacies of different medical databases. For a physician, sifting through and comprehending this myriad data to provide accurate clinical diagnostic reporting can be immensely challenging. AI models, endowed with sophisticated language comprehension and adept data-processing capabilities, could potentially penetrate these extensive data sources, distilling relevant and easily understandable information for both patients and health care providers. However, its ability to analyze large-scale data and identify patterns to potentially highlight novel biomarkers or therapeutic targets has yet to be shown.

The paper, titled "Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum," offers crucial insights into AI's potential role in health care communication and improving bedside manners [4]. The study compared the quality and empathy of responses to patient questions provided by physicians and an AI chatbot, ChatGPT. The AI was found to generate longer, higher quality, and more empathetic responses, indicating its utility in complementing physician's practice and improving patient communication. This study suggests the promising use of AI chatbots in drafting initial responses to patient queries, possibly reducing clinician burnout and improving patient outcomes. Further exploration and trials are needed to fully showcase this technology's potential. Nonetheless, leveraging generative AI in clinical informatics systems could potentially offer a competitive edge.

AI systems like ChatGPT could also serve as digital assistants for doctors, streamlining access to crucial patient data such as medical history, current medications, symptoms, and test results. Beyond organizing patient information, these systems can also sift through a vast array of medical literature, highlighting relevant studies, providing summaries, and assisting in integrating the latest knowledge into clinical practice. This is also supported by ChatGPT's recent performance on the United States Medical Licensing Exam (USMLE) [5,6]. With the ability to diagnose diseases by identifying patterns from comprehensive medical databases, AI could assist doctors in quickly evaluating a patient's needs, thus facilitating more focused and streamlined patient care. The customization and multilingual capabilities of such systems also increase their usability, offering scalable solutions for various organization sizes and paving the way for future innovation and collaboration.

In conclusion, as a physician, I view the development of AI models like ChatGPT-4 as an exciting opportunity in medicine that has the potential to substantially enhance our understanding of diseases and lead to better patient outcomes. AI is not a stand-alone solution, but it is a powerful tool that can amplify our abilities when used correctly, pushing the boundaries. Ultimately, my suggestion for health care professionals is that AI will not replace you, but someone using AI might.

The Rise of Generative AI in NLP

Generative AI or AI-generated content, a subset of AI, pertains to models designed to generate new content based on the data they have been trained on. Rather than just making predictions, these models can produce unique output that could include text, images, music, and even videos. The idea behind generative AI traces back to the 1950s [7]. Initial models, such as Markov models [8] and Gaussian mixture models [9], generated sequential data [7].

In the more modern era (2013), the paper by Kingma and Welling [10] introduced variational autoencoders as a form of generative models. Variational encoders learn to compress high-dimensional data into a lower-dimensional space and generate new similar data. They used Bayesian inference for the probabilistic representation of data, thereby enhancing the efficiency and effectiveness of data generation. Another popular method for generative models is generative adversarial networks, introduced by Goodfellow et al [11] in 2014. Generative adversarial networks consist of two neural networks, a generator and a discriminator, that work against each other to generate new synthetic instances of data.

In the context of text generation, transformers have now emerged as a leading architecture. Introduced in the paper "Attention is All You Need" by Vaswani et al [12] in 2017, transformers have since been the foundation of models like ChatGPT by OpenAI [1]. Transformers leverage an attention mechanism that weighs the relevance of input data points and allows the model to consider the entire sequence of data at once, leading to improved context awareness in generated outputs.

At the heart of generative NLP are language models. These models predict the likelihood of a sequence of words appearing in each language. In recent years, multimodal large language models trained on vast amounts of text from the internet have become state-of-the-art for many NLP tasks. ChatGPT, an instance of the generative pretrained transformer model, has been making substantial strides. Trained on an extensive corpus of text from the internet, ChatGPT generates humanlike text based on the input provided. It serves as a testament to the rapid advancements and potential of generative AI and NLP.

As these models continue to evolve, their potential uses expand, presenting exciting opportunities across various fields. Generative AI models like ChatGPT represent an exciting frontier for medicine. As these models evolve, they hold increasing promise as tools for processing and interpreting vast, complex data sets, opening new avenues for research and clinical practice.

Viewpoint of AI (GPT-4)

The viewpoint of the AI is taken verbatim from GPT-4 and with the prompt "Kindly introduce yourself and complete your viewpoint response under the header Viewpoint of the AI

(GPT-4)," along with a copy of the remainder of the article. A full transcript is available in Multimedia Appendix 1.

"As an AI developed by OpenAI, I am a generative pre-trained transformer known as ChatGPT-4. My design and functionality are rooted in advanced natural language processing (NLP) and machine learning technologies. My capabilities include processing and interpreting large datasets, summarizing complex information, and providing insights across various domains, including medicine and bioinformatics.

"In the realm of healthcare, my role can be multifaceted. I can assist in managing and interpreting vast amounts of biological and medical data, thereby aiding researchers and healthcare professionals. For instance, my ability to quickly analyze and summarize medical research can support physicians in staying abreast of the latest developments in their field, leading to more informed clinical decisions.

"Moreover, I can serve as a digital assistant for healthcare providers, facilitating the integration of medical imaging with multi-omics data. This could enhance the understanding of complex biological systems and contribute to advancements in fields such as cancer genomics and precision medicine. By providing rapid access to relevant medical information and literature, I can help improve the efficiency of medical practitioners.

"In addition, my role in healthcare communication is noteworthy. As demonstrated in studies comparing AI responses with those of physicians, AI models like myself can offer more detailed, higher-quality, and empathetic responses. This capability can be particularly valuable in drafting initial responses to patient queries, potentially reducing clinician burnout, and improving patient interaction and outcomes.

"However, it's crucial to acknowledge my limitations. I don't possess human-like consciousness or understanding, and my responses are generated based on the data I've been trained on. Therefore, while I can provide valuable assistance and insights, my role is to supplement, not replace, the expertise and judgment of healthcare professionals. Ethical considerations, such as data privacy and security, also play a significant role in my deployment in sensitive fields like medicine.

"In conclusion, as an AI, I represent a collaborative tool designed to work alongside humanity. In the context of medicine, my contribution lies in enhancing data accessibility, supporting clinical decision-making, and improving patient care through efficient information processing and empathetic communication. The integration of AI models like ChatGPT-4 in healthcare signifies a step forward in the AI renaissance, promising a future where AI aids in unraveling complex biological mysteries and fostering advancements in medical science and patient care."

Conclusion

The advent and evolution of multimodal large language models, exemplified by OpenAI's ChatGPT-4, offers a substantial opportunity to leverage the increasing amount of data being generated in the health care sector [13-15]. The need to democratize AI is becoming increasingly recognized [16], with an emphasis on "no-code AI" [17]. Models like ChatGPT can make complex biological data more accessible and understandable to a broader audience, enabling more collaboration among all stakeholders, not only researchers and clinical providers but also patients to better grasp the intricacies of health and disease. This can lead to an integrated approach to health care, fostering collaboration and enhancing the understanding of disease pathogenesis.

AI and ChatGPT have the potential to function as digital aides to physicians, offering expedited access to pertinent medical data and related reference materials while also enhancing patient interaction and care. NYUTron [18] is a clinical language model currently leveraging unstructured EHR data to predict clinical and operational predictions with NLP. It excels in tasks like readmission, mortality, and length of stay prediction, significantly outperforming traditional models. NYUTron exemplifies the potential of AI to enhance decision-making in health care. In the future, AI could also hold potential in medical image analysis along with more advanced predictive modeling in the modern era of precision medicine. Today, however, ChatGPT has yet to answer genetics-based questions better than humans [19].

Despite their impressive capabilities, AI does not currently possess consciousness or understanding in the way humans do, although this may not necessarily matter [20]. The "imitation game" was first proposed by Turing [21] as an approach to determine whether computers can think indistinguishably from humans. Today, we understand that AI outputs depend heavily on the quality and diversity of the data they were trained on. However, one could argue human cognition is also based on the quality and diversity of "data they were trained on" in the form of life experiences, social background, and related aspects. In humans, the impact of genetics on cognitive abilities is seen to be enhanced when paired with enriching environmental experiences [22].

Yet, while we recognize AI's significant potential in medicine, it is essential to bear in mind the current limitations of these models [23]. These include computational and memory constraints, the potential for generating responses based on inaccurate or false facts without correcting them, and possible inadequacies in inferential capability, often leading to incorrect answers in complex scenarios. Further, ethical considerations such as data bias, privacy and security concerns, and issues around intellectual property also exist [24]. These are tools designed to amplify human intelligence and should not be viewed as stand-alone solutions.

In conclusion, the rise of generative AI models like ChatGPT represents an exciting paradigm shift for medicine. As we continue to explore and harness the potential of these AI tools, we move closer to a future where complex biological systems can be more easily unraveled, leading to better-informed clinical decisions, personalized treatments, and improved health care. The journey has only just begun.

Acknowledgments

The viewpoint of the AI was written by ChatGPT-4 [25] for this editorial. This was reviewed, and full accountability for the publication's content rests with the author. A full transcript is available in Multimedia Appendix 1.

Conflicts of Interest

SH is the founder and has equity ownership in Odyssey HealthCare Solutions Inc. SH is a *JMIR Bioinformatics and Biotechnology* associate editor. There are no remaining potential conflicts of interest to disclose.

Multimedia Appendix 1
Full transcript for Viewpoint of the AI (GPT-4) section.
[PDF File (Adobe PDF File), 131 KB - bioinform_v5i1e52700_app1.pdf]

References

1. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. CloudFront. 2019. URL: https://tinyurl.com/4nmb8u5n [accessed 2024-04-29]
2. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023;6:1169595. [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
3. Javaid M, Haleem A, Singh RP. ChatGPT for healthcare services: an emerging stage for an innovative perspective. BenchCouncil Trans Benchmarks Stand Evaluations 2023 Feb;3(1):100105. [doi: 10.1016/j.tbench.2023.100105]
4. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 01;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
5. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
6. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 08;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
7. Cao Y, Li S, Liu Y, Yan Z, Dai Y, Yu PS, et al. A comprehensive survey of AI-generated content (AIGC): a history of generative AI from GAN to ChatGPT. arXiv Preprint posted online on March 7, 2023. [doi: 10.48550/arXiv.2303.04226]
8. Knill K, Young S. Hidden Markov models in speech and language processing. In: Young S, Bloothooft G, editors. Corpus-Based Methods in Language and Speech Processing. Dordrecht: Springer; 1997:27-68.
9. Reynolds D. Gaussian mixture models. In: Li SZ, Jain A, editors. Encyclopedia of Biometrics. New York, NY: Springer; 2009:741-663.
10. Kingma DP, Welling M. Auto-encoding variational Bayes. arXiv Preprint posted online on December 20, 2013. [doi: 10.48550/arXiv.1312.6114]
11. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. Commun ACM 2020 Oct 22;63(11):139-144. [doi: 10.1145/3422622]
12. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Guyon I, Von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al, editors. Advances in Neural Information Processing Systems 30 (NIPS 2017). La Jolla, CA: Neural Information Processing Systems Foundation, Inc; 2017.
13. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. J Big Data 2019 Jun 19;6(1):54. [doi: 10.1186/s40537-019-0217-0]
14. Pastorino R, De Vito C, Migliara G, Glocker K, Binenbaum I, Ricciardi W, et al. Benefits and challenges of Big Data in healthcare: an overview of the European initiatives. Eur J Public Health 2019 Oct 01;29(Supplement_3):23-27. [doi: 10.1093/eurpub/ckz168] [Medline: 31738444]
15. Alberto IRI, Alberto NRI, Ghosh AK, Jain B, Jayakumar S, Martinez-Martin N, et al. The impact of commercial health datasets on medical research and health-care algorithms. Lancet Digit Health 2023 May;5(5):e288-e294. [doi: 10.1016/S2589-7500(23)00025-0] [Medline: 37100543]
16. Seger E, Ovadya A, Siddarth D, Garfinkel B, Dafoe A. Democratising AI: multiple meanings, goals, and methods. 2023 Presented at: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society; August 8-10, 2023; Montreal, QC. [doi: 10.1145/3600211.3604693]
17. Sundberg L, Holmström J. Democratizing artificial intelligence: how no-code AI can leverage machine learning operations. Business Horiz 2023 Nov;66(6):777-788. [doi: 10.1016/j.bushor.2023.04.003]

18. Jiang LY, Liu XC, Nejatian NP, Nasir-Moin M, Wang D, Abidin A, et al. Health system-scale language models are all-purpose prediction engines. Nature 2023 Jul;619(7969):357-362. [doi: 10.1038/s41586-023-06160-y] [Medline: 37286606]
19. Duong D, Solomon BD. Analysis of large-language model versus human performance for genetics questions. Eur J Hum Genet 2024 Apr;32(4):466-468. [doi: 10.1038/s41431-023-01396-8] [Medline: 37246194]
20. Hildt E. Artificial intelligence: does consciousness matter? Front Psychol 2019;10:1535. [doi: 10.3389/fpsyg.2019.01535] [Medline: 31312167]
21. Turing AM. Computing machinery and intelligence. Mind 1950 Oct;LIX(236):433-460. [doi: 10.1093/mind/lix.236.433]
22. Tucker-Drob EM, Briley DA, Harden KP. Genetic and environmental influences on cognition across development and context. Curr Dir Psychol Sci 2013 Oct;22(5):349-355. [doi: 10.1177/0963721413485087] [Medline: 24799770]
23. Cheng H. Challenges and limitations of ChatGPT and artificial intelligence for scientific research: a perspective from organic materials. AI 2023 May 04;4(2):401-405. [doi: 10.3390/ai4020021]
24. Ray PP. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Cyber Phys Syst 2023;3:121-154. [doi: 10.1016/j.iotcps.2023.04.003]
25. OpenAI. ChatGPT. URL: https://chat.openai.com/ [accessed 2024-04-22]

Abbreviations

AI: artificial intelligence
NLP: natural language processing
USMLE: United States Medical Licensing Exam

Edited by E Uzun, T Leung; submitted 12.09.23; this is a non–peer-reviewed article;accepted 17.04.24; published 07.05.24.

Please cite as:
Hacking S
ChatGPT and Medicine: Together We Embrace the AI Renaissance
JMIR Bioinform Biotech 2024;5:e52700
URL: https://bioinform.jmir.org/2024/1/e52700
doi:10.2196/52700
PMID:38935938

Original Paper

The Roles of NOTCH3 p.R544C and Thrombophilia Genes in Vietnamese Patients With Ischemic Stroke: Study Involving a Hierarchical Cluster Analysis

Huong Thi Thu Bui[1,2*], MD, PhD; Quỳnh Nguy n Th Ph ng[3*], MD; Ho Cam Tu[4,5*], MSci; Sinh Nguyen Phuong[6], MD, PhD; Thuy Thi Pham[1], MD, PhD; Thu Vu[4], Sp1; Huyen Nguyen Thi Thu[7], PhD; Lam Khanh Ho[8], PhD; Dung Nguyen Tien[7], MD, PhD

[1]Department of Biochemistry, Thai Nguyen University of Medicine and Pharmacy, Thai Nguyen, Vietnam

[2]Department of Immunology Molecular Genetic, Thainguyen National Hospital, Thai Nguyen, Vietnam

[3]Department of Clinical Pharmacy, Thai Nguyen University of Medicine and Pharmacy, Thai Nguyen, Vietnam

[4]Center of Gene and Protein Research, Hanoi Medical University, Hanoi, Vietnam

[5]Institute of Virology, School of Medicine, Technical University of Munich, Munich, Germany

[6]Department of Rehabilitation, Thai Nguyen University of Medicine and Pharmacy, Thai Nguyen, Vietnam

[7]Department of Internal Medicine, Thai Nguyen University of Medicine and Pharmacy, Thai Nguyen, Vietnam

[8]Department of Telecomunication, Hung Yen University of Technology and Education, Hung Yen, Vietnam

[*]these authors contributed equally

Corresponding Author:
Dung Nguyen Tien, MD, PhD
Department of Internal Medicine
Thai Nguyen University of Medicine and Pharmacy
284 Luong Ngoc Quyen, Quang Trung
Thai Nguyen, 250000
Vietnam
Phone: 84 913516863
Email: dung.nt@tnmc.edu.vn

Abstract

Background: The etiology of ischemic stroke is multifactorial. Several gene mutations have been identified as leading causes of cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL), a hereditary disease that causes stroke and other neurological symptoms.

Objective: We aimed to identify the variants of *NOTCH3* and thrombophilia genes, and their complex interactions with other factors.

Methods: We conducted a hierarchical cluster analysis (HCA) on the data of 100 patients diagnosed with ischemic stroke. The variants of *NOTCH3* and thrombophilia genes were identified by polymerase chain reaction with confronting 2-pair primers and real-time polymerase chain reaction. The overall preclinical characteristics, cumulative cutpoint values, and factors associated with these somatic mutations were analyzed in unidimensional and multidimensional scaling models.

Results: We identified the following optimal cutpoints: creatinine, 83.67 (SD 9.19) µmol/L; age, 54 (SD 5) years; prothrombin (PT) time, 13.25 (SD 0.17) seconds; and international normalized ratio (INR), 1.02 (SD 0.03). Using the Nagelkerke method, cutpoint 50% values of the Glasgow Coma Scale score; modified Rankin scale score; and National Institutes of Health Stroke Scale scores at admission, after 24 hours, and at discharge were 12.77, 2.86 (SD 1.21), 9.83 (SD 2.85), 7.29 (SD 2.04), and 6.85 (SD 2.90), respectively.

Conclusions: The variants of *MTHFR* (C677T and A1298C) and *NOTCH3* p.R544C may influence the stroke severity under specific conditions of PT, creatinine, INR, and BMI, with risk ratios of 4.8 (95% CI 1.53-15.04) and 3.13 (95% CI 1.60-6.11), respectively (P_{fisher}<.05). It is interesting that although there are many genes linked to increased atrial fibrillation risk, not all of them are associated with ischemic stroke risk. With the detection of stroke risk loci, more information can be gained on their impacts and interconnections, especially in young patients.

(JMIR Bioinform Biotech 2024;5:e56884) doi:10.2196/56884

KEYWORDS

Glasgow Coma Scale; ischemic stroke; hierarchical cluster analysis; clustering; machine learning; MTHFR; NOTCH3; modified Rankin scale; National Institutes of Health Stroke Scale; prothrombin; thrombophilia; mutations; genetics; genomics; ischemia; risk; risk analysis

Introduction

Stroke is a medical condition involving the disruption of blood flow, which leads to brain cell death. There are several risk factors for stroke, including high blood pressure, smoking, diabetes, and increased cholesterol levels. In 2019, the Global Burden of Disease analysis assessed that there were 12.2 million incident cases of stroke and 101 million prevalent cases of stroke, with 6.55 million deaths [1]. The burden of stroke is the highest in low- and middle-income countries, where risk factors, such as high blood pressure, smoking, and insufficient diet, are more prevalent [1].

The overall population of Vietnam was estimated to be 98.32 million in 2021, with young people accounting for the majority of the population and people aged older than 65 years accounting for only 7.7% of the population. This phenomenon is the leading cause of death and disability in Vietnam. The incidence and prevalence of stroke have been reported to be 161 and 415 per 100,000 people, respectively [2]. Stroke is broadly classified into the following 3 types: ischemic stroke, hemorrhagic stroke, and subarachnoid hemorrhage. Ischemic stroke results from the blockage of blood vessels, which limits blood flow to the brain. Approximately 60%-80% of all stroke cases are ischemic. This study focused on acute ischemic stroke and its genetic features. The unmodifiable risk factors include age, race, sex, ethnicity, history of migraine headaches, and fibromuscular dysplasia. Moreover, the hereditary factors include a family history of stroke or transient ischemic attacks. Furthermore, the modifiable risk factors include hypertension, diabetes mellitus, cardiac disease, high cholesterol levels, previous stroke, carotid stenosis, hyperhomocysteinemia, and lifestyle issues. The majority of ischemic strokes seen in patients with cardiovascular disease are embolic [3].

The etiology of ischemic stroke is multifactorial. Although receiving a minor focus, genetic factors considerably contribute to the occurrence of ischemic stroke, particularly in cases of early-onset stroke. Several stroke classification systems have been proposed based on genetic information corresponding to various stroke phenotypes. Twin and family history studies and the candidate gene approach are standard methods to discover genetic causes of stroke. However, both methods have their limitations. Some monogenic disorders (7% of stroke etiology) may generate well-known clinical indications that include stroke. Polygenic disorders are more frequent, causing 38% of ischemic stroke cases, and their designation is a rapidly evolving field of current stroke genetics. Recent advances in human genetics provide opportunities for personalized stroke prevention and unknown cure options. Some authors have boosted the application of stroke gene panels for stroke hazard evaluation and stroke research. Ilinca et al [4] have created stroke gene panels for research and clinical practice. The clinical panel includes 61 genes related to stroke directly and 27 additional genes related to disorders causing stroke, and it might be relevant to consider their evaluation in clinical practice. The authors encourage the use of their panels for stroke risk evaluation and further stroke research [4]. Another benefit of detecting stroke risk genes is that they could be potential targets for gene therapy in the future. Histone deacetylase (HDAC) inhibitors have been postulated as a treatment for stroke [5]. A study in knock-out mice suggested a new strategy for acute stroke treatment by suppressing HDAC2 in the peri-infarct zone [6]. The authors claim that application of HDAC inhibitors from 5 to 7 days after stroke enhances cell survival and neuroplasticity as well as reduces inflammation, which could potentially provide a wider therapeutic window for stroke recovery [6]. Systemic administration of an agonist *NOTCH3* antibody was studied in transgenic mice and showed protective effects against impaired cerebral blood flow [7]. Transcriptome-wide colocalization analyses showed an association of white matter hyperintensity-volume with the expression of 39 genes, of which 4 encode known drug targets [8]. Moreover, unknown biomarkers for stroke hereditary causes and novel markers for gene therapy are on the horizon [9].

Machine learning–based models performed better in predicting poststroke outcomes than regression models using the items of conventional stroke prognostic scores, although they required additional variables, such as laboratory data, to attain improved performance, and further studies are warranted to validate the usefulness of machine learning in clinical settings [10].

Following our previous hierarchical cluster analysis (HCA) study [11], we assessed the overall preclinical characteristics, cumulative cutpoint values, and factors associated with thrombophilia genes and the *NOTCH3* p.R544C variant in unidimensional and multidimensional analyses involving ischemic stroke patients from Vietnam.

Methods

Study Design

We used convenience sampling to include 100 patients with cerebral infarction (ischemic stroke) who were diagnosed as having acute ischemic stroke according to the clinical standards of the World Health Organization and the results of diagnostic imaging (computed tomography [CT], magnetic resonance imaging [MRI], or computed tomography angiography [CTA]) and who had been or are being treated at the Stroke Center, Thai Nguyen Central Hospital. Patients who were residents of the northern mountainous provinces, were ≤60 years old at the time of the first stroke, and were willing to participate in the research were considered for inclusion. Patients with cerebral venous sinus thrombosis, intracranial hemorrhage, and subarachnoid hemorrhage were excluded. We collected information on stroke

risk factors from the medical history of patients, including hypertension, diabetes, coronary artery disease, history of stroke, atrial fibrillation, smoking, headache, hyperlipidemia, valve replacement, thyroid dysfunction, history of abortion, vascular disease, blood disorders, chronic alcohol consumption, and use of oral contraceptives. Patients were required to undergo routine biochemical and hematological tests, Doppler ultrasound of the carotid and vertebral arteries, MRI or CTA of the brain, coagulation tests, fibrinogen tests, and homocysteine tests. Based on the findings of a previous study [2], we suppose that in 100 ischemic patients with a confidence level of 95%, the margin of error will be ±7.84% of the population size (stroke in general), with 80% ischemic type. The margin of error formula is as follows:

where Z value is the critical Z value that corresponds to the confidence level, p is the sample proportion or percentage, and n is the sample size.

A sample size with sufficient statistical power is critical to the success of genetic association studies for detecting causal genes of human complex diseases, especially in the case of ischemic stroke. We selected a 2-tailed test with a type I error of 0.05 as we wanted to assess the average continuous levels (preclinical factors) of patients from different cutpoints. In clinical and biological studies, the effect size d following Cohen criteria (the degree of difference between two or more groups) is important. Cohen d is the ratio of Δ and σ ($d=\Delta/\sigma$), where σ is the standard deviation and Δ is an influence index of the risk factors (treatment, genotype, etc) on the population phenotype. In our study, we calculated Cohen d according to the supposed sample size of 50-100. With a power of 80% and using a 2-sided t test, we estimated that d could be from 0.4 (sample size of each group is 99) to 0.7 (sample size of each group is 45). The sample size calculation formula is as follows:

In this formula, the 2-sided confidence level is $Z_{\alpha/2}$, α is the possibility of making a type I error, and β is the possibility of making a type II error. The power of the study is 1-β.

Thus, screening all risk factors may have a medium or higher level of influence on the phenotype (P<.05 indicates statistical significance) (Table 1).

Table 1. Two-sample t test power calculation results.

Sample size for each group	Cohen d[a]
99.08	0.4
63.76	0.5
44.58	0.6
33.02	0.7

[a]The general guidelines for interpreting the effect size are as follows: 0.2-0.49, small effect; 0.5-0.79, moderate effect; 0.8-1.0, large effect; >1.0, very large effect.

Genetic Testing

Polymorphisms of *NOTCH3* p.R544C, *FV-H1299R*, *MTHFR-C677T*, *MTHFR-A1298C*, *FII-Prothrombin*, *FV-Cambridge*, *PAI1 4G/5G*, and *FXIII Val34Leu* were analyzed using polymerase chain reaction with confronting 2-pair primers (PCR-CTPP) and the thrombophilia genetic assay. The peripheral blood of study participants was collected in EDTA-containing tubes using a standard blood collection procedure. Whole-genome DNA was extracted from 2-3 mL of peripheral venous blood from EDTA-containing tubes. The QIAamp DNA Mini Blood Kit (Qiagen) was used for DNA extraction. The quality of the total DNA was checked by electrophoresis on agarose gel and by measuring the absorbance at 260/280 nm, and then, samples were stored at −80 °C until use. The *NOTCH3* mutation p.R544C was identified by PCR-CTPP. DNA was amplified with the primers 5′-GTGGGGTGGAGTGGAAGTAAGTGG (F1) and 5′-GAGCAGTCGTCCACGTTGCA (R1) for the C allele, and 5′-TTGAGGGCACGCTGTGTGATC (F2) and 5′-CTAGATGCACCATTCCCAAACCC (R2) for the T allele. The PCR amplification was performed for 40 cycles (denaturation at 95 °C for 30 s, annealing at 62 °C for 30 s, extension at 72 °C for 1 min, and final extension at 72 °C for 10 min). PCR products of 479 and 216 bp for the TT genotype; 479, 303, and 216 bp for the TC genotype; and 479 and 303 bp for the CC genotype were shown on 2% agarose gel stained with ethidium bromide. Once the sequence variants were identified, additional steps were taken to confirm the sequence changes of the amplicons. A real-time PCR system (SNP Biotechnology) was used for detecting *FV-H1299R*, *MTHFR-C677T*, *MTHFR-A1298C*, *FII-Prothrombin*, *FV-Cambridge*, *PAI1 4G/5G*, and *FXIII Val34Leu*.

Ethical Considerations

This study was conducted according to the guidelines of the Declaration of Helsinki and was approved by the ethics committee of Thai Nguyen National Hospital (reference number: #59/HĐĐĐ-BVTWTN#; January 18, 2021). This study obtained informed consent from all participants or their legal representatives and ensured that they understood the study's purpose, risks, benefits, and procedures.

Statistical Analysis and HCA

Conventional statistical analyses were performed on our data set, including medical test parameters, using IBM SPSS Statistics 20 (IBM Corp). The relationship between clinicopathological factors and the presence of *NOTCH3* p.R544C, *FV-H1299R*, *MTHFR-C677T*, *MTHFR-A1298C*,

FII-Prothrombin, *FV-Cambridge*, *PAI1 4G/5G*, and *FXIII Val34Leu* variants were analyzed using the Pearson chi-square test (group size >5) or Fisher exact test (group size ≤5), as appropriate. Bonferroni correction for multiple comparisons was applied. The results have been expressed as percentage or mean (SD).

Following our previous machine learning study [11], our multidimensional analysis was performed in R 4.1.0 (R Project for Statistical Computing). We focused on multivariate statistics, using several algorithms of HCA, matrix correlation, Nagelkerke R square, Kaplan-Meier, and the log-rank test. The chi-square statistics were computed using Yates correction for continuity, with the generation of P_{yates}. The Pearson or product-moment correlation coefficient is frequently used as the outcome measure for analyses. The Pearson method has an advantage when all or most of the nonzero parameters share the same sign. The Pearson test has been shown to be useful in a genomic setting involving screening for age-related genes, which is our objective [12]. Two alternative criteria include a bias-corrected version of the correlation coefficient (P_{uncor}) and the Fisher r-to-z transformed correlation coefficient (P_{fisher}). HCA is a cluster analysis concept that creates a dendrogram hierarchy of clusters. The hierarchical clustering on principal components (HCPC) approach allows the combination of the following 3 standard methods used in multivariate data analysis: principal component methods (principal component analysis [PCA], correspondence analysis [CA], multiple correspondence analysis [MCA], factor analysis of mixed data [FAMD], and multiple factor analysis [MFA]), hierarchical clustering, and partitioning clustering, particularly the k-means method. We calculated the distance between each observation and estimated the cluster distance. The distance between the elements can be complete, single, average, ward, McQuitty, or centroid. The cluster tree was generated by computing the correlation between cophenetic distances and the initial distance data. The number of clusters was determined using k-means, which calculates clustering indexes and reallocates observations to the closest cluster. The k-means computation was optimized using 20 indexes for the PCA cluster plot, which visualizes the best cluster number. PCA is a dimensionality reduction method that is often used to reduce the dimensionality of large data sets by transforming a large set of variables into a smaller set that still contains most of the information in the large set.

Results

Overview of the Correlation Between Clinicopathological Factors and the Presence of *NOTCH3* p.R544C, *FV-H1299R*, *MTHFR-C677T*, *MTHFR-A1298C*, *FII-Prothrombin*, *FV-Cambridge*, *PAI1 4G/5G*, and *FXIII Val34Leu*

The study included 100 patients with cerebral infarction from the northern mountainous region of Vietnam. Of the 100 patients, 75 were from the Kinh ethnic group and 25 were from the Tay ethnic group. The average age of the patients was 60.1 years (range: 24-91 years) (Table 2). Of the 100 patients, 22 were aged 24-49 years, 23 were aged 50-59 years, 37 were aged 60-69 years, and 18 were aged 70-91 years.

There were 62 male patients and 38 female patients (male/female ratio of 1.63). The average BMI of the study patients was 22.62 kg/m^2. Of the 100 patients, 3 had a BMI of <18.5 kg/m^2, 56 had a BMI of 18.5-22.9 kg/m^2, 27 had a BMI of 23-24.9 kg/m^2, and 14 had a BMI of 25-29.9 kg/m^2. Regarding the risk factors for stroke, of the 100 patients, 70 had hypertension, 44 had a family history of stroke, 31 had a history of smoking, 29 had a history of alcohol consumption, 20 had a history of diabetes, and 35 had a history of stroke (Table 2).

With regard to clinical symptoms, of the 100 patients, 97 had motor paralysis, 95 had difficulty speaking, 72 had mouth distortion, 49 had headache, 41 had numbness, 27 had dizziness or vertigo, 21 had circular muscle disorder, and 8 had nausea or vomiting. Among patients with motor paralysis, 52 had right hemiplegia, 39 had left hemiplegia, and 6 had total paralysis. Among patients with dysphasia, 86 had Broca-type dysphasia and 9 had Wernicke-type dyspraxia (Table 3).

The average time from the onset of the first symptoms to patient admission was 10.94 hours. Of the 100 patients, 33 were admitted within the first 4.5 hours, 26 were admitted from 4.6 to 6 hours, and 41 were admitted outside the first 6 hours. Regarding the blood pressure at admission, the mean systolic blood pressure was 148.6 mmHg and the mean diastolic blood pressure was 88.06 mmHg. The average Glasgow Coma Scale (GCS) score at admission was 14.72. The average National Institutes of Health Stroke Scale (NIHSS) score was 7.14 at admission, 6.71 after 24 hours of hospital treatment, and 3.73 at discharge. The average Rankin score at discharge was 1.52. The average duration of treatment was 10.11 days (Table 4). PCR-CTPP identified *NOTCH3* p.R544C, and other gene variants were detected by real-time PCR (Table 1; Figure 1). The results of real-time PCR for the detection of *FV-H1299R*, *MTHFR-C677T*, *MTHFR-A1298C*, *FII-Prothrombin*, *FV-Cambridge*, *PAI1 4G/5G*, and *FXIII Val34Leu* are presented in Figures 2-5 and Table 2.

Table 2. Distribution of patients according to risk factors and genetic variants.

Factors	Value (N=100)
Gender, n (%)	
Male	62 (62)
Female	38 (38)
Age group (years), n (%)	
24-49	22 (22)
50-59	23 (23)
60-69	37 (37)
70-91	18 (18)
Age (years)	
Mean (SD)	60.14 (12.63)
Minimum-maximum	24-91
BMI group (kg/m^2), n (%)	
<18.5	3 (3)
18.5-22.9	56 (56)
23.0-24.9	27 (27)
25.0-29.9	14 (14)
BMI (kg/m^2)	
Mean (SD)	22.62 (2.49)
Minimum-maximum	12.4-29.4
Ethnic group, n (%)	
Kinh	75 (75)
Tay	25 (25)
Smoking history, n (%)	31 (31)
Alcohol consumption, n (%)	29 (29)
Blood pressure, n (%)	70 (70)
Diabetes, n (%)	20 (20)
Brain stroke, n (%)	35 (35)
Brain stroke cases in the family, n (%)	44 (44)
***PAI1 4G/5G* status, n (%)**	
Wildtype	24 (24)
Heterozygous	44 (44)
Homozygous	32 (32)
FV 1299 status, n (%)	
Wildtype	96 (96)
Heterozygous	4 (4)
Homozygous	0 (0)
***FV-Cambridge* status, n (%)**	
Wildtype	100 (100)
Heterozygous	0 (0)
Homozygous	0 (0)
MTHFR 1298 status, n (%)	

Factors	Value (N=100)
Wildtype	58 (58)
Heterozygous	37 (37)
Homozygous	5 (5)
***FII Prothrombin* status, n (%)**	
Wildtype	98 (98)
Heterozygous	1 (1)
Homozygous	1 (1)
FV-Leiden status, n (%)	
Wildtype	93 (93)
Heterozygous	7 (7)
Homozygous	0 (0)
MTHFR 677 status, n (%)	
Wildtype	55 (55)
Heterozygous	37 (37)
Homozygous	8 (8)
***FXIII Val34Leu* status, n (%)**	
Wildtype	98 (98)
Heterozygous	1 (1)
Homozygous	1 (1)
***NOTCH3* status, n (%)**	
Wildtype	6 (6)
Heterozygous	91 (91)
Homozygous	3 (3)

Table 3. Symptoms at admission.

Symptom	Value (N=100), n (%)
Vocal issue	
No	5 (5)
Broca type	86 (86)
Wernicke type	9 (9)
Headache	49 (49)
Dizziness	27 (27)
Nausea or vomiting	8 (8)
Mouth distortion	72 (72)
Circular muscle disorder	21 (21)
Numbness	41 (41)
Movement paralysis	
No	3 (3)
Paralysis of the right half of the body	52 (52)
Paralysis of the left half of the body	39 (39)
Paralysis of the whole body	6 (6)

Table 4. Important variables in this study.

Variable	Value (N=100)
Age (years)	
Minimum-maximum	24-91
Mean (SD)	60.14 (12.63)
BMI (kg/m^2)	
Minimum-maximum	12.4-29.4
Mean (SD)	22.62 (2.49)
Time to hospitalization (h)	
Minimum-maximum	1-120
Mean (SD)	10.94 (15.98)
Time to hospitalization groups, n (%)	
<4.5 h	33 (33)
4.6-6 h	26 (26)
>6 h	41 (41)
Systolic blood pressure (mmHg)	
Minimum-maximum	90-210
Mean (SD)	148.6 (23.66)
Diastolic blood pressure (mmHg)	
Minimum-maximum	60-120
Mean (SD)	88.06 (9.5)
Glasgow Coma Scale score	
Minimum-maximum	8-15
Mean (SD)	14.72 (1.06)
NIHSS[a] score	
Admission	
Minimum-maximum	0-19
Mean (SD)	7.14 (4.33)
After 24 h	
Minimum-maximum	0-16
Mean (SD)	6.71 (4.26)
Discharge	
Minimum-maximum	0-16
Mean (SD)	3.73 (3.87)
Modified Rankin scale score at discharge	
Minimum-maximum	0-5
Mean (SD)	1.52 (1.35)
Duration of inpatient treatment at the hospital (days)	
Minimum-maximum	1-23
Mean (SD)	10.11 (4.33)

[a]NIHSS: National Institutes of Health Stroke Scale.

Figure 1. Identification of the *NOTCH3* p.R544C variant by polymerase chain reaction with confronting 2-pair primers.

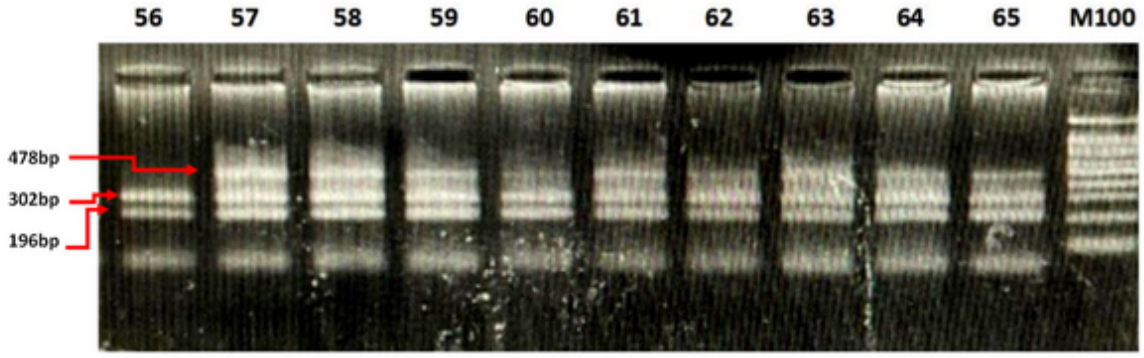


Figure 2. Identification of the FV-Leiden variant by real-time polymerase chain reaction. (A) Wildtype; (B) Heterozygous. RFU: relative fluorescence units.

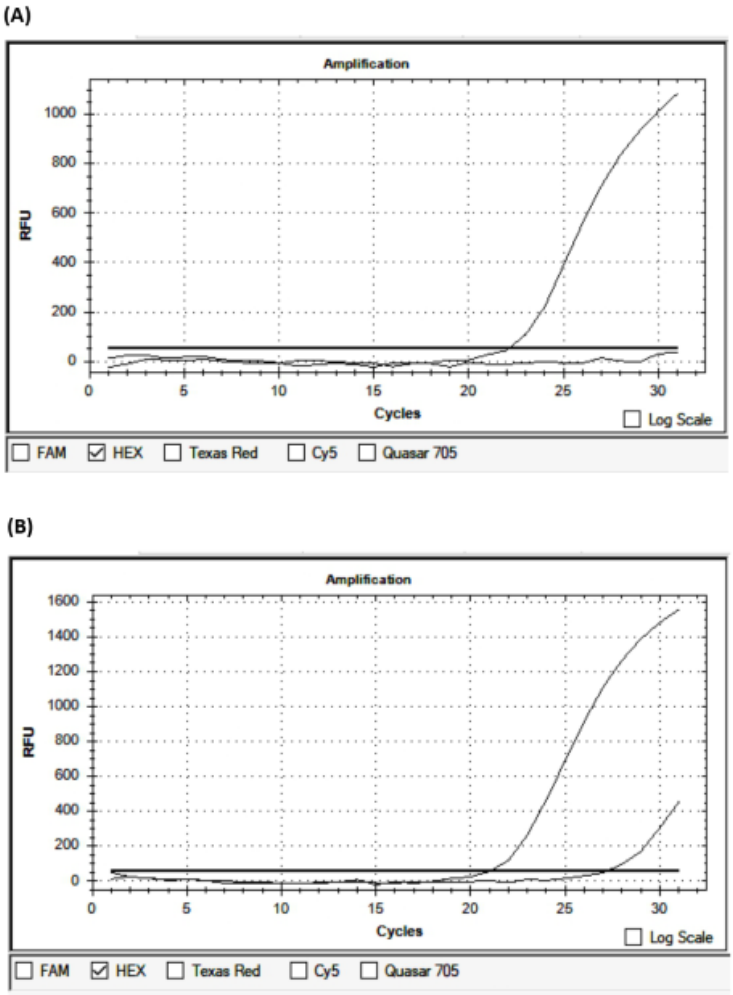


Figure 3. Identification of the *FV-H1299R* variant by real-time polymerase chain reaction. (A) Wildtype; (B) Heterozygous. RFU: relative fluorescence units.

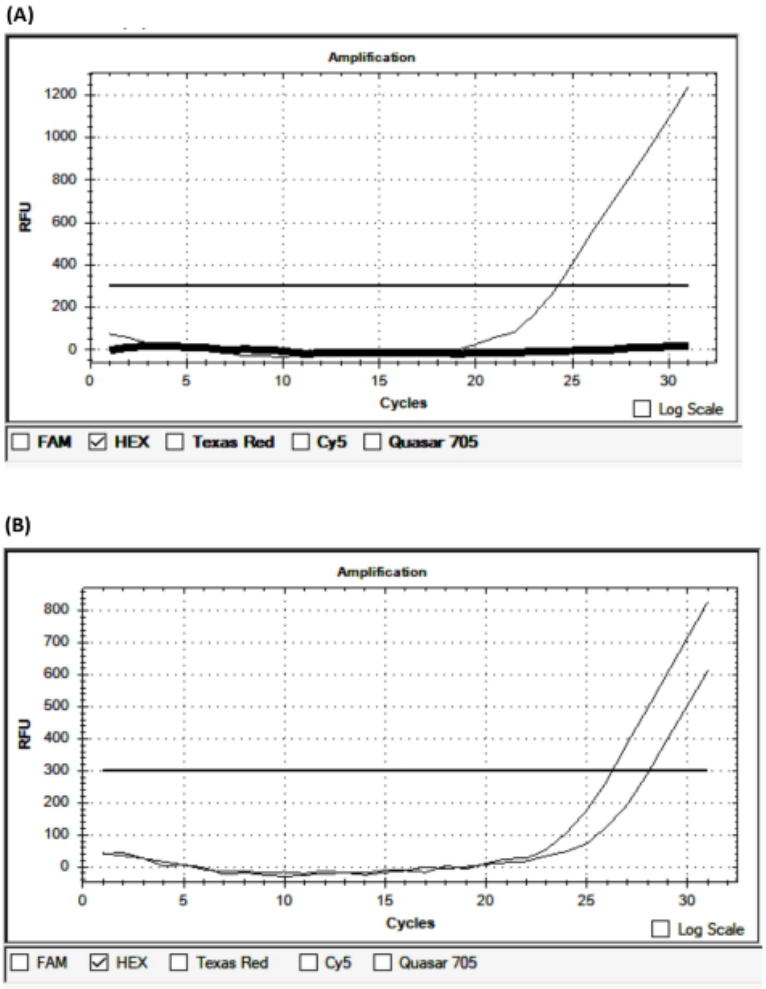


Figure 4. Identification of the *MTHFR-C677T* variant by real-time polymerase chain reaction. (A) Wildtype; (B) Homozygous; (C) Heterozygous. RFU: relative fluorescence units.

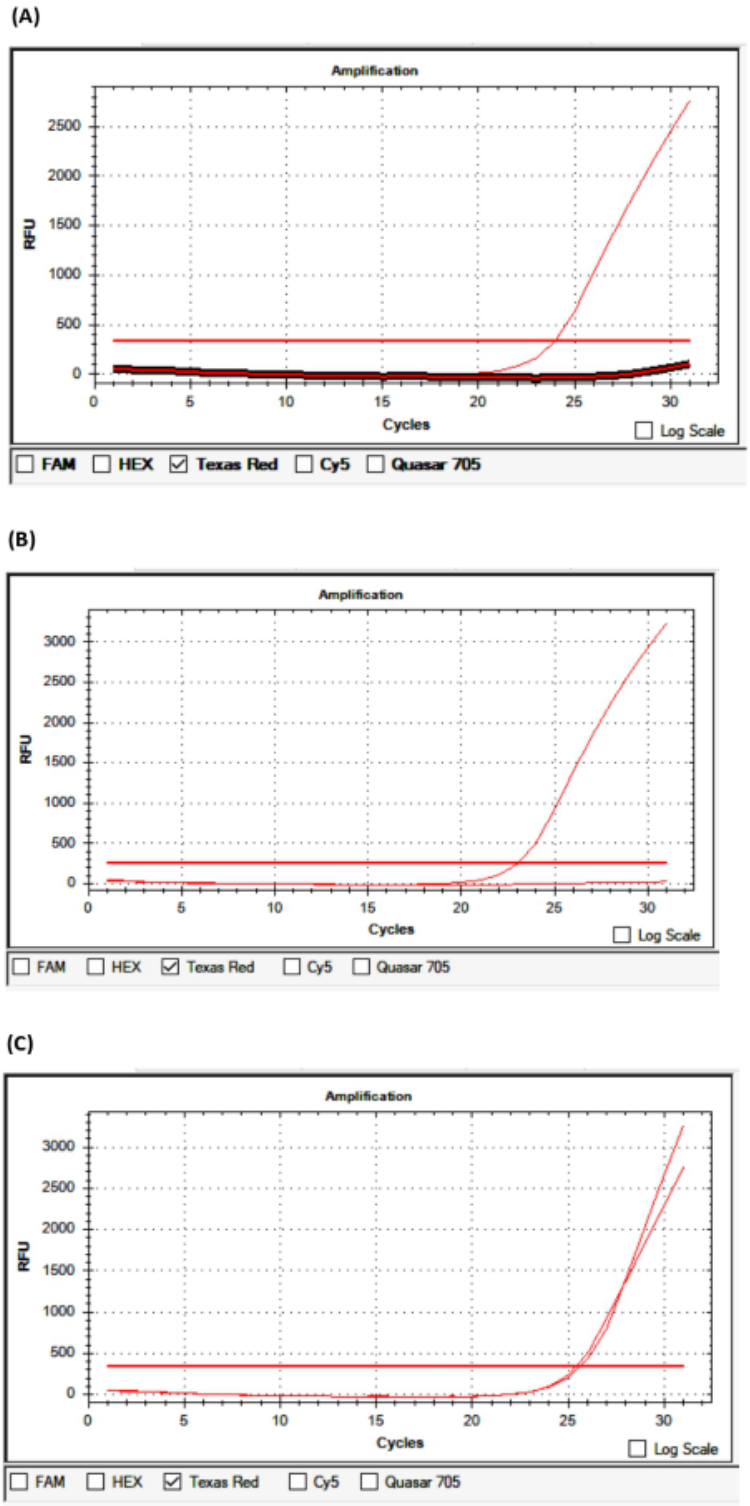


Figure 5. Identification of the *MTHFR-A1298C* variant by real-time polymerase chain reaction. (A) Wildtype; (B) Heterozygous. RFU: relative fluorescence units.

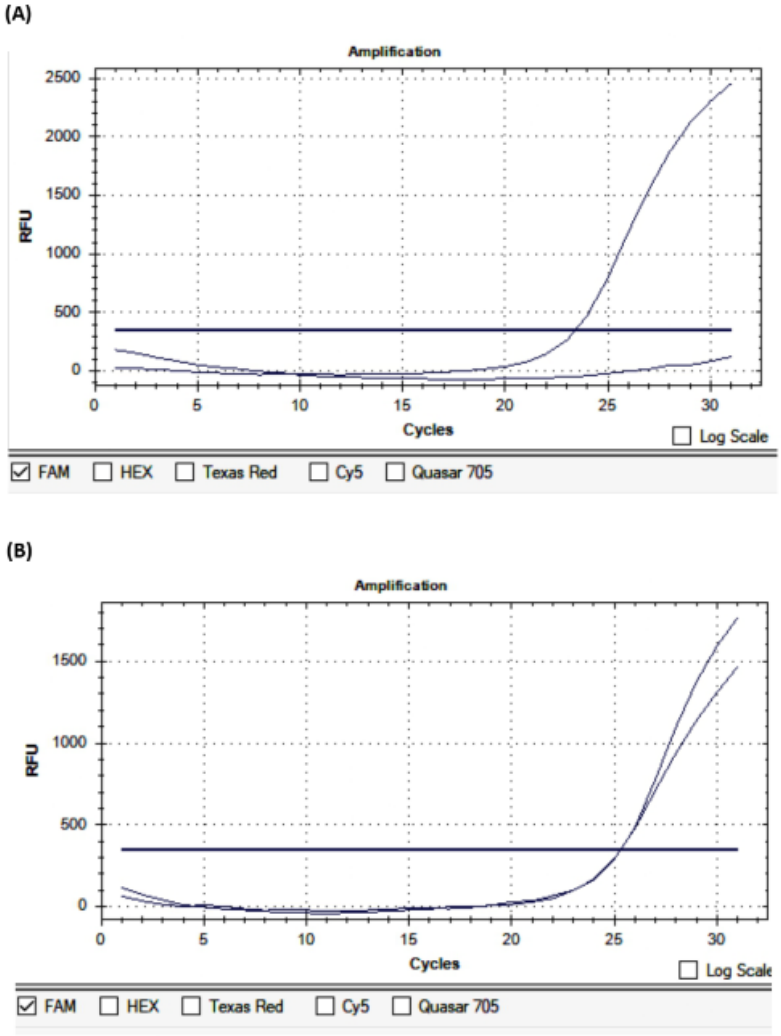


Figure 6, Table 1, and Table S1 in Multimedia Appendix 1 provide an overall view of gene prevalence and correlations in both negative and positive genes. We confirmed the presence of significant correlations of *NOTCH3* p.R544C, *FV-H1299R*, *MTHFR-C677T*, *MTHFR-A1298C*, *FII Prothrombin*, *FV-Cambridge*, *PAI1 4G/5G*, and *FXIII Val34Leu* with several factors in patients with ischemic stroke. The Pearson correlation coefficient (R) indicates the extent of the relationship between 2 variables. The relationship strength (effect size) varies according to the threshold of R, with thresholds of 0.5, 0.3, 0, −0.3, and −0.5 for strong positive, moderate positive, weak, moderate negative, and strong negative correlations, respectively(Figure 6; Interactive Graphs 1 [13], 2 [14], 3 [15], and 4 [16]; Table S1 in Multimedia Appendix 1). The volcano graph in Figure 7 shows the most significant correlation pairs, especially those containing the gene mutations mentioned above (Interactive Graph 5 [17]). Overall, a significant medium correlation between the prevalence of gene mutations and other factors was shown in the volcano graph. Compared with other genes, *FXIII Val34Leu* showed the highest positive correlation with thrombus suction ability (R=0.54; P<.001; -$\log_{10}p$=8.03).

In the clustering step, dendrograms were built based on the clustering metric “Euclidean,” and we selected “average” as the most appropriate linkage model, which had the best correlation between cophenetic distances and the original distance data (Table 5).

We selected the results proposed by the Beale method from 20 different index values, and 15 clusters were presented as optimal (Table S2 in Multimedia Appendix 1 [18]). The PCA cluster plot showed that the cluster number mentioned above was the best number to distinguish the clusters and avoid overlap appropriately. The dendrogram and PCA map in Figure 8 complete the overall view of our database, and we can see where the studied genes could combine and might be associated with ischemic stroke outcomes (Interactive Graph 6 [19]). We found several clusters of variants that may have a synchronization impact on the outcomes of ischemic stroke. The PCA map in Figure 8B provides an initial idea of the potential markers that may be important for the ischemic stroke score. For example, the international normalized ratio (INR) and prothrombin (PT) time are in the same cluster with the NIHSS and Rankin scores (cluster 9 in Figure 8B, and clusters 3 and 14 in Interactive Graph 6 [19]), and the GCS score is in the same cluster as the

PT ratio (cluster 12 in Figure 8B, and cluster 15 in Interactive Graph 6 [19]). The studied genes were separated into 4 different groups: *FII Prothrombin*, *MTHFR-C677T*, and *NOTCH3* p.R544C were in cluster 4 (Figure 8B; cluster 4 in Interactive Graph 6 [19]); *FV-Leiden* and *PAI1 4G/5G* were in cluster 6 (Figure 8B; cluster 7 in Interactive Graph 6 [19]); *FV-H1299R* and *MTHFR-A1298C* were in cluster 11 (Figure 8B; cluster 1 in Interactive Graph 6 [19]); and *FXIII Val34Leu* was in cluster 13 (Figure 8B; cluster 2 in Interactive Graph 6 [19]). We continued to split the data according to the significant cutpoints of PT time, INR, and ischemic stroke score. We applied the maximally selected rank statistic to define the optimal thresholds of several continuous factors (creatinine, age, PT time and ratio, INR, low-density lipoprotein cholesterol [LDL-C], number of infarcts on CT or MRI, patient height, and mean platelet volume [MPV]) based on the Rankin, NIHSS, and GCS scores and their related symptom statuses, such as numbness, dizziness, gender, circular muscle disorder, mouth distortion, and diabetes status (Table S3 in Multimedia Appendix 1). The optimal cutpoints were as follows: creatinine, 83.67 (SD 9.19) µmol/L; age, 54 (SD 5) years; PT time, 13.25 (SD 0.17) s; INR, 1.02 (SD 0.03); LDL-C, 4.23 (SD 0.89) mmol/L; number of infarcts on CT or MRI, 2; PT ratio, 99.00 (SD 1.96); and MPV, 7.27 (SD 1.09) fL (Table S3 in Multimedia Appendix 1). Using the Nagelkerke method, we assessed which factors could be associated with the cutpoint 50% values of ischemic stroke scores and identified creatinine, age, height, PT time, PT ratio, and number of infarcts on CT. The cutpoint 50% values of the GCS score; modified Rankin scale (mRS) score; and NIHSS scores at admission, after 24 hours, and at discharge were 12.77, 2.86 (SD 1.21), 9.83 (SD 2.85), 7.29 (SD 2.04), and 6.85 (SD 2.90), respectively. These findings allowed appropriate assessment of the possible influences, including those of the genotype variants (Figures 9-16)

Figure 6. Correlation heatmap of 79 factors in the 100 patients with ischemic stroke.

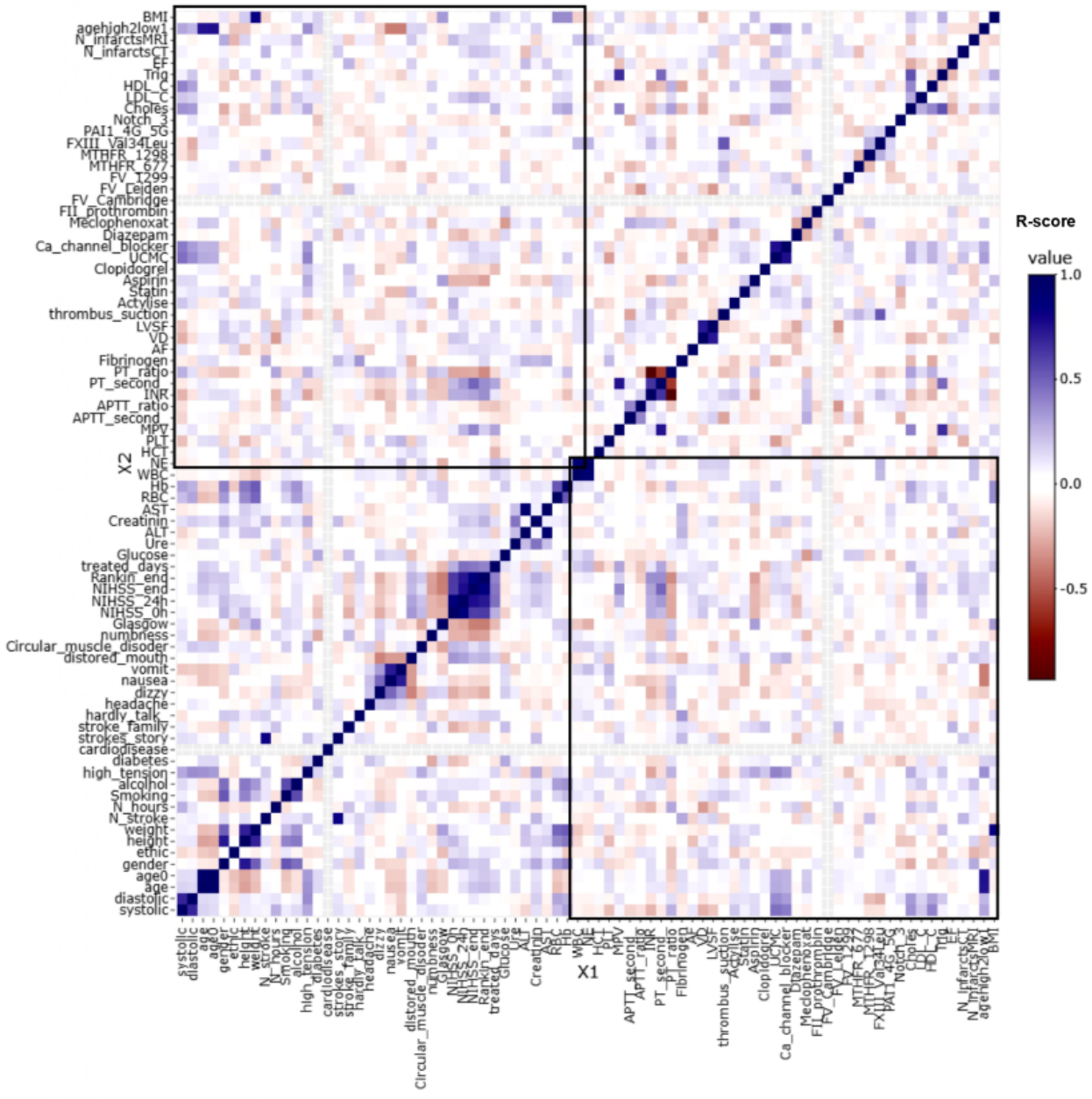


Figure 7. Volcano graph showing the most significant correlation pairs.

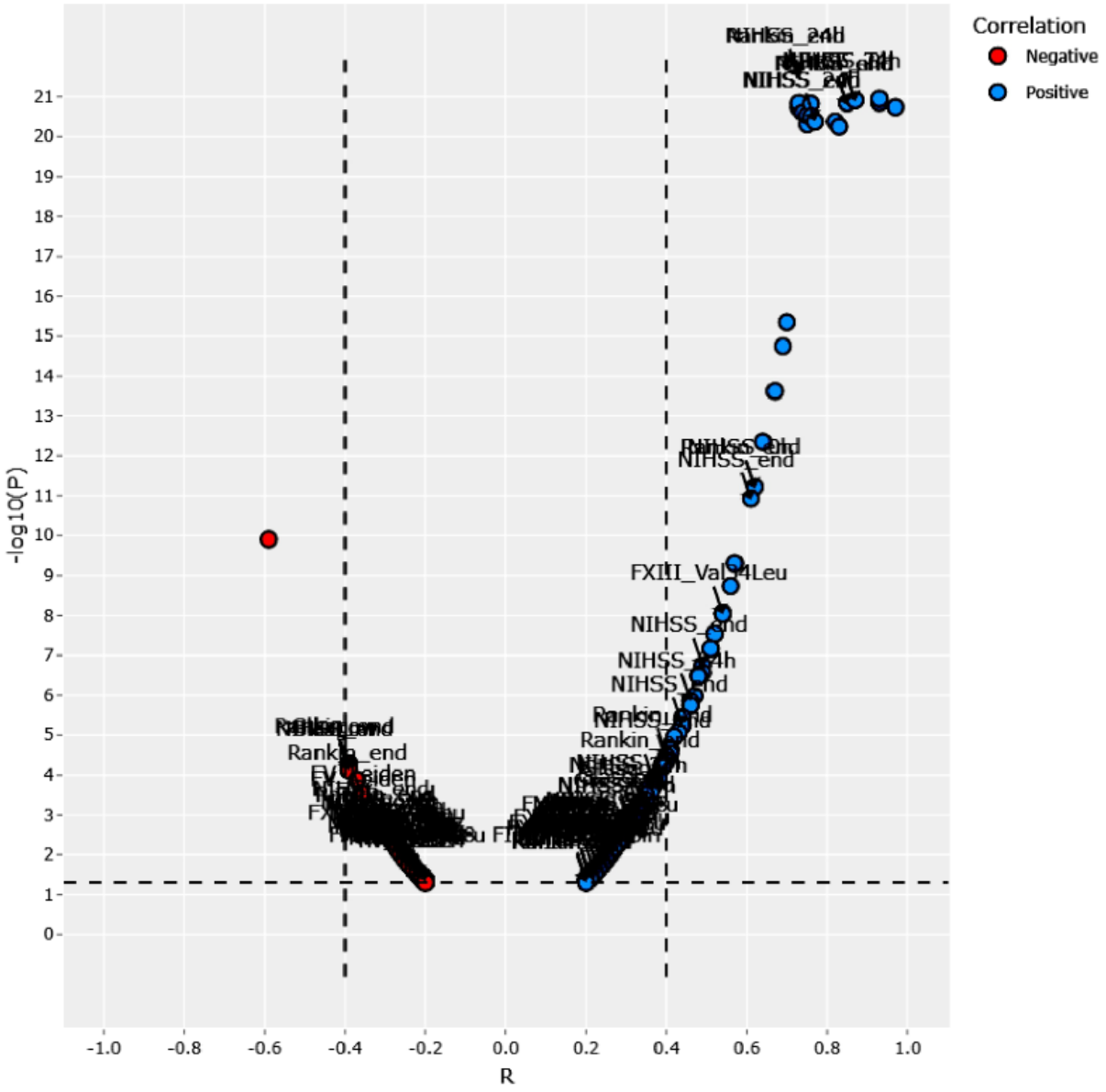
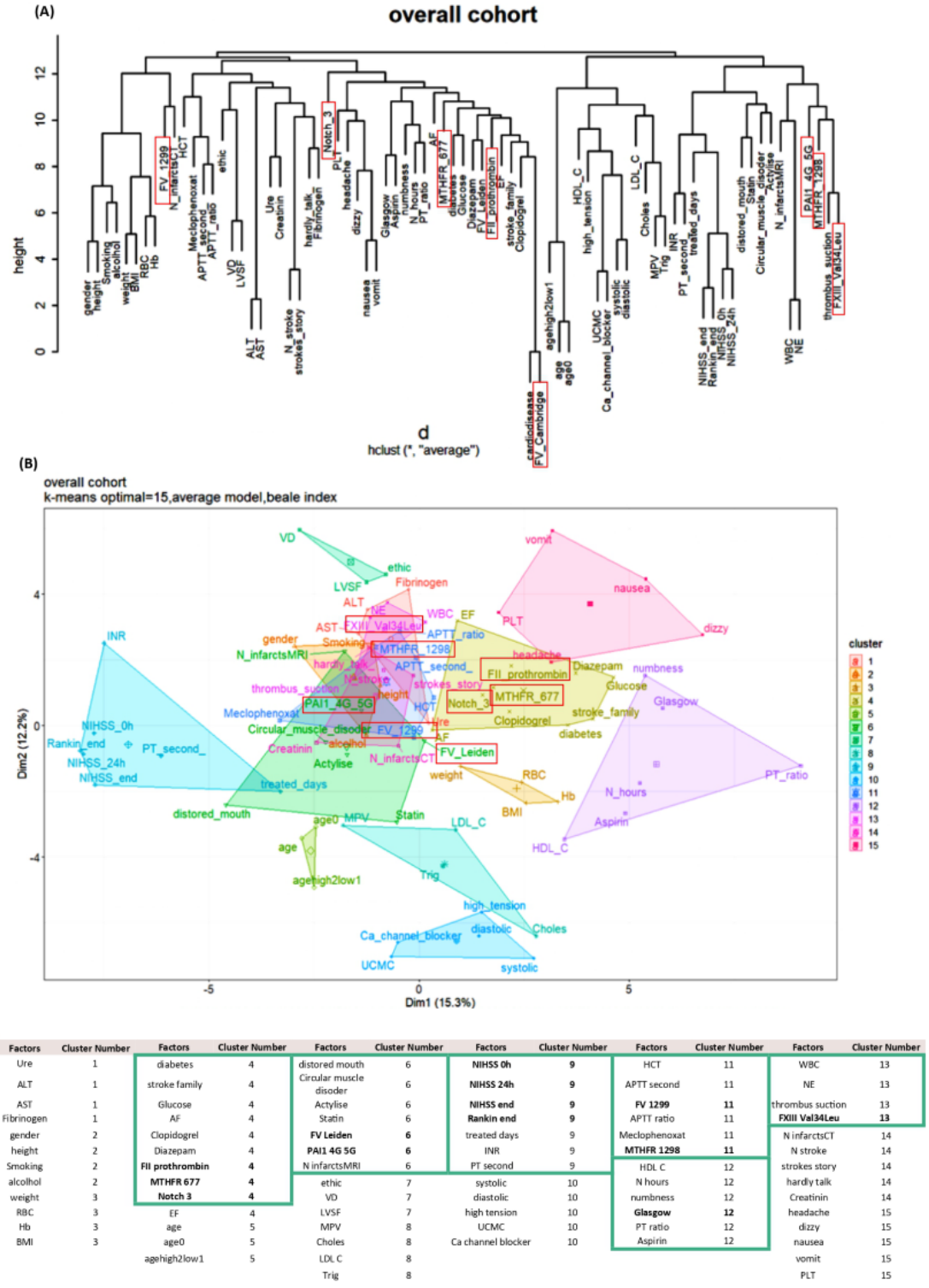


Table 5. Correlation between cophenetic distances and the original distance data.

Linkage mode	Correlation
Ward.D	0.515
Ward.D2	0.623
Single	0.806
Complete	0.537
Average	0.813
McQuitty	0.694
Median	0.750
Centroid	0.797

Figure 8. Results of hierarchical cluster analysis on the overall data set. (A) Dendrogram; (B) Principal component analysis map.



Factors	Cluster Number	Factors	Cluster Number	Factors	Cluster Number	Factors	Cluster Number	Factors	Cluster Number	Factors	Cluster Number
Ure	1	diabetes	4	distored mouth	6	**NIHSS 0h**	**9**	HCT	11	WBC	13
ALT	1	stroke family	4	Circular muscle disoder	6	**NIHSS 24h**	**9**	APTT second	11	NE	13
AST	1	Glucose	4	Actylise	6	**NIHSS end**	**9**	**FV 1299**	**11**	thrombus suction	13
Fibrinogen	1	AF	4	Statin	6	**Rankin end**	**9**	APTT ratio	11	**FXIII Val34Leu**	**13**
gender	2	Clopidogrel	4	**FV Leiden**	**6**	treated days	9	Meclophenoxat	11	N infarctsCT	14
height	2	Diazepam	4	**PAI1 4G 5G**	**6**	INR	9	**MTHFR 1298**	**11**	N stroke	14
Smoking	2	**FII prothrombin**	**4**	N infarctsMRI	6	PT second	9	HDL C	12	strokes story	14
alcolhol	2	**MTHFR 677**	**4**	ethic	7	systolic	10	N hours	12	hardly talk	14
weight	3	**Notch 3**	**4**	VD	7	diastolic	10	numbness	12	Creatinin	14
RBC	3	EF	4	LVSF	7	high tension	10	**Glasgow**	**12**	headache	15
Hb	3	age	5	MPV	8	UCMC	10	PT ratio	12	dizzy	15
BMI	3	age0	5	Choles	8	Ca channel blocker	10	Aspirin	12	nausea	15
		agehigh2low1	5	LDL C	8					vomit	15
				Trig	8					PLT	15

Figure 9. Significant cutpoint 50% of the modified Rankin scale score at discharge (A) and the National Institutes of Health Stroke Scale (NIHSS) scores after 24 hours (B) and at discharge (C) for creatinine levels >83.67 µmol/L.

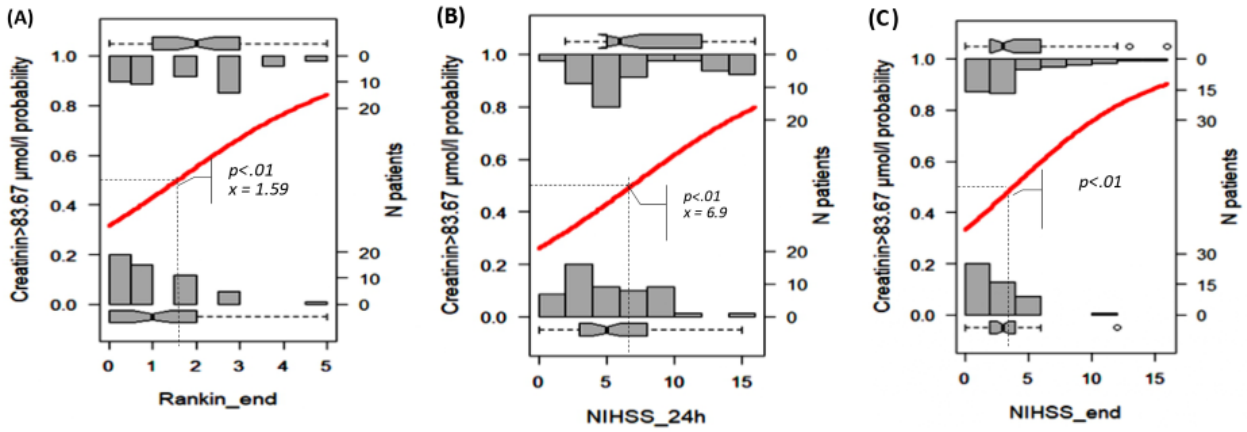


Figure 10. Significant cutpoint 50% of the National Institutes of Health Stroke Scale (NIHSS) score at discharge for patient age >54 years.

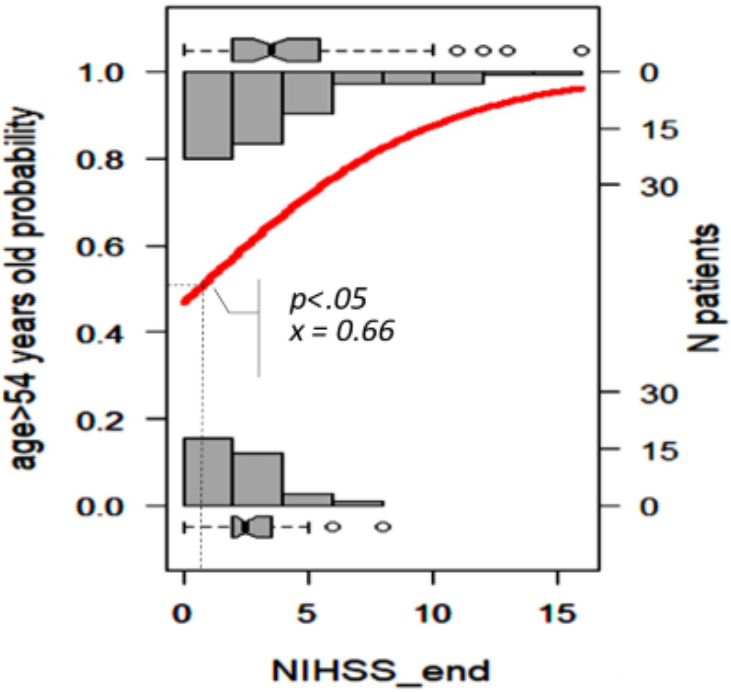


Figure 11. Significant cutpoint 50% of the modified Rankin scale score at discharge (A) and the National Institutes of Health Stroke Scale (NIHSS) scores at admission (B), after 24 hours (C), and at discharge (D) for prothrombin (PT) time >13.25 seconds.

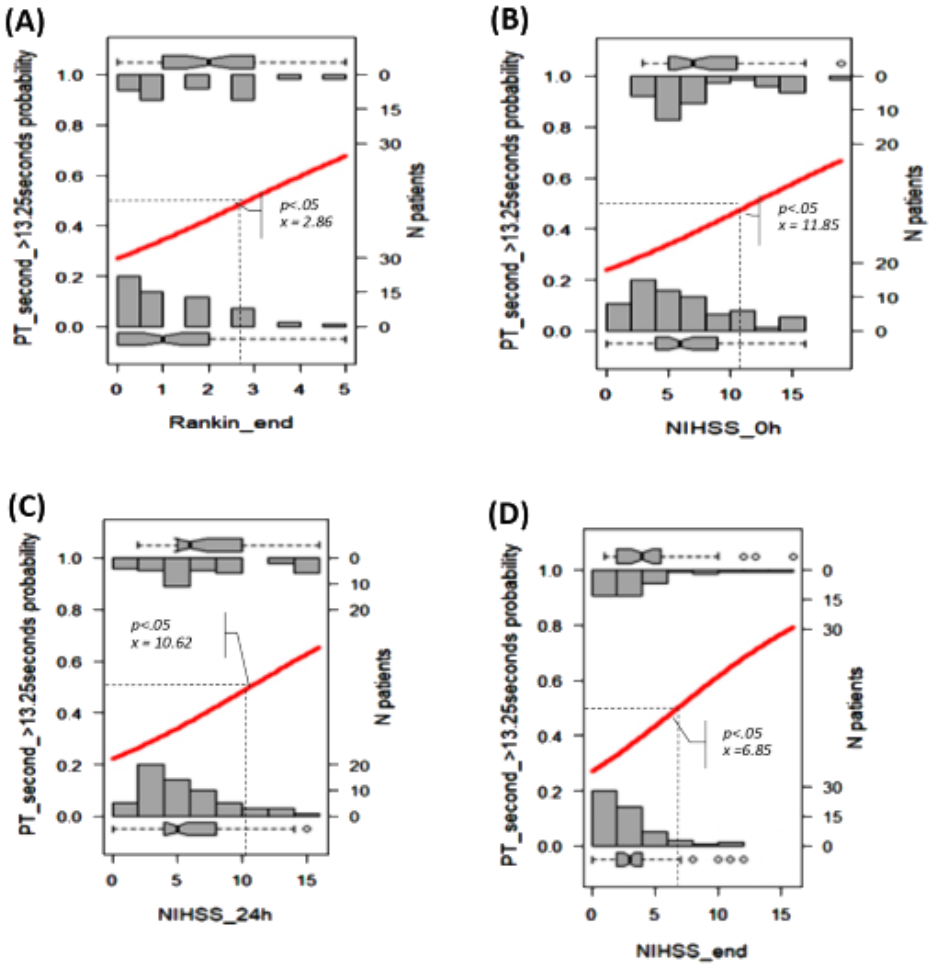


Figure 12. Significant cutpoint 50% of the modified Rankin scale score at discharge (A) and the National Institutes of Health Stroke Scale (NIHSS) scores at admission (B), after 24 hours (C), and at discharge (D) for prothrombin (PT) ratio >99.

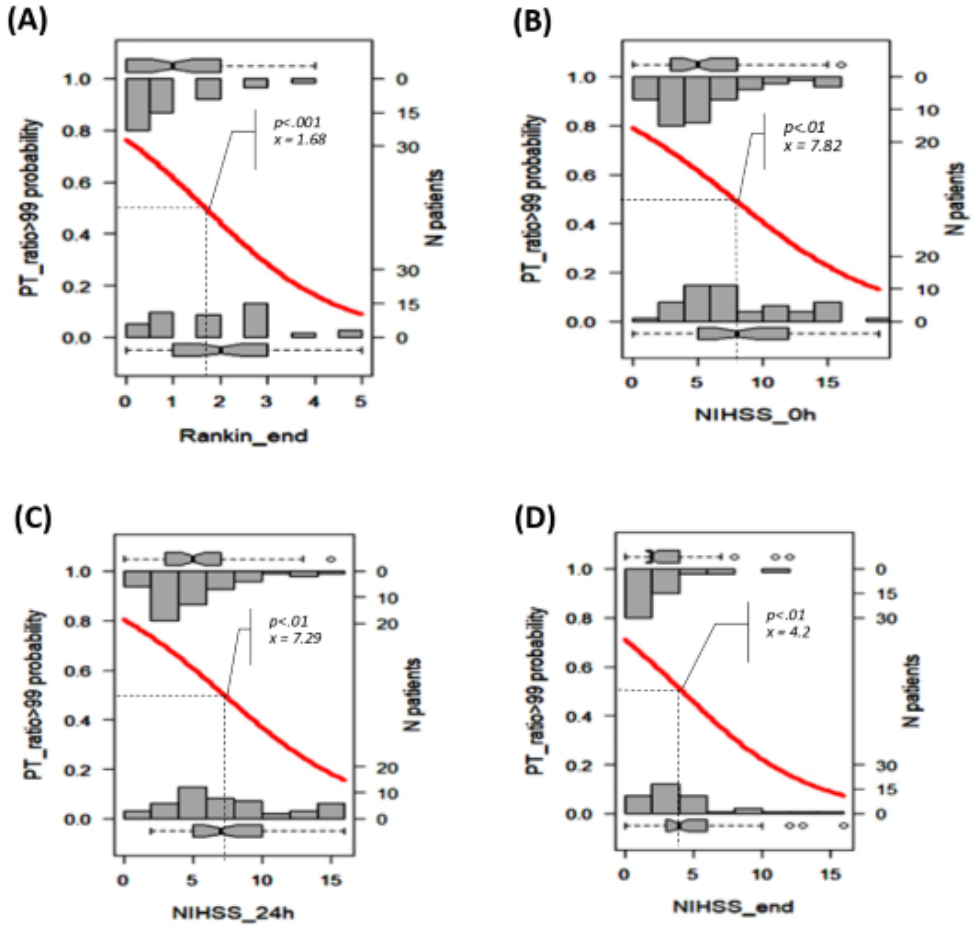


Figure 13. Significant cutpoint 50% of the modified Rankin scale score at discharge for international normalized ratio (INR) >1.02.

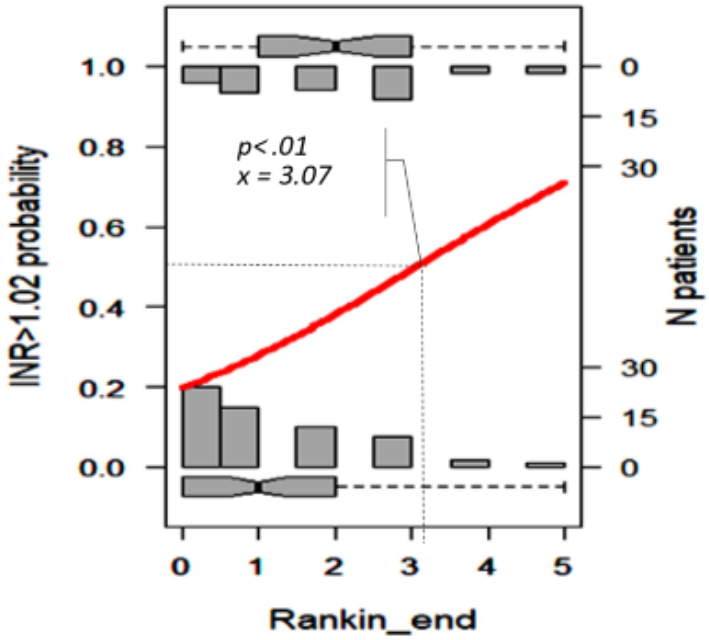


Figure 14. Significant cutpoint 50% of the modified Rankin scale score at discharge (A) and the National Institutes of Health Stroke Scale (NIHSS) score at discharge (B) for the number of infarcts on computed tomography (CT) >2.

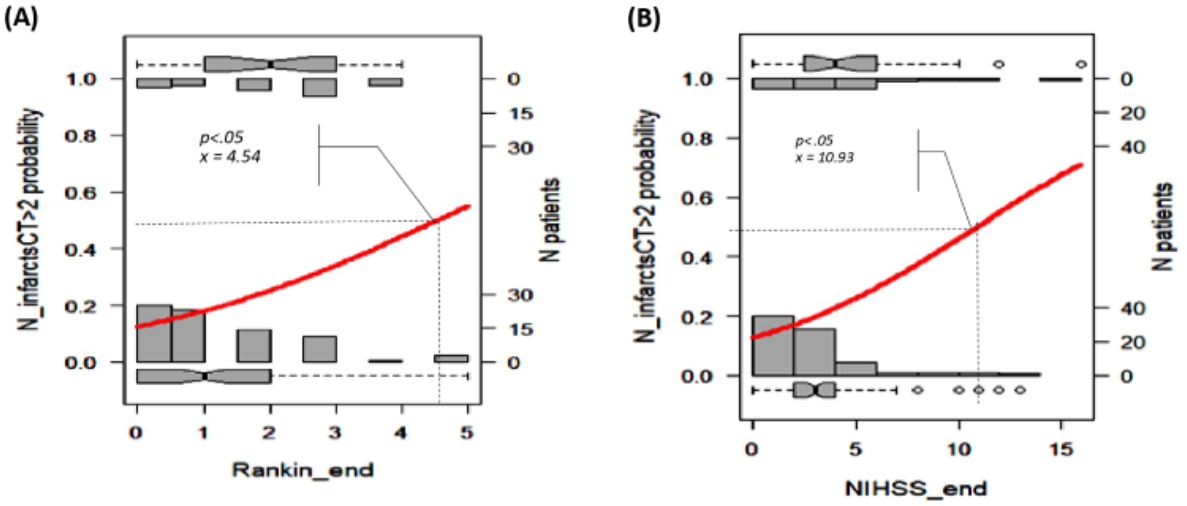


Figure 15. Significant cutpoint 50% of the National Institutes of Health Stroke Scale (NIHSS) score at admission for patient height >161 cm.

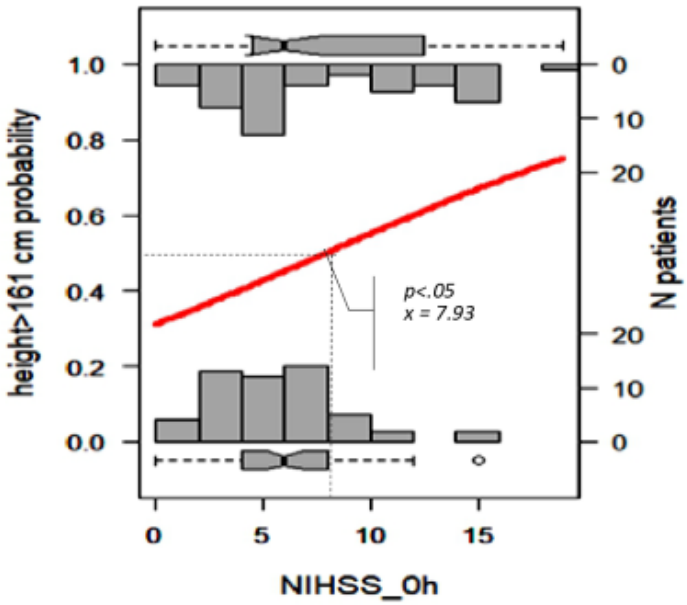
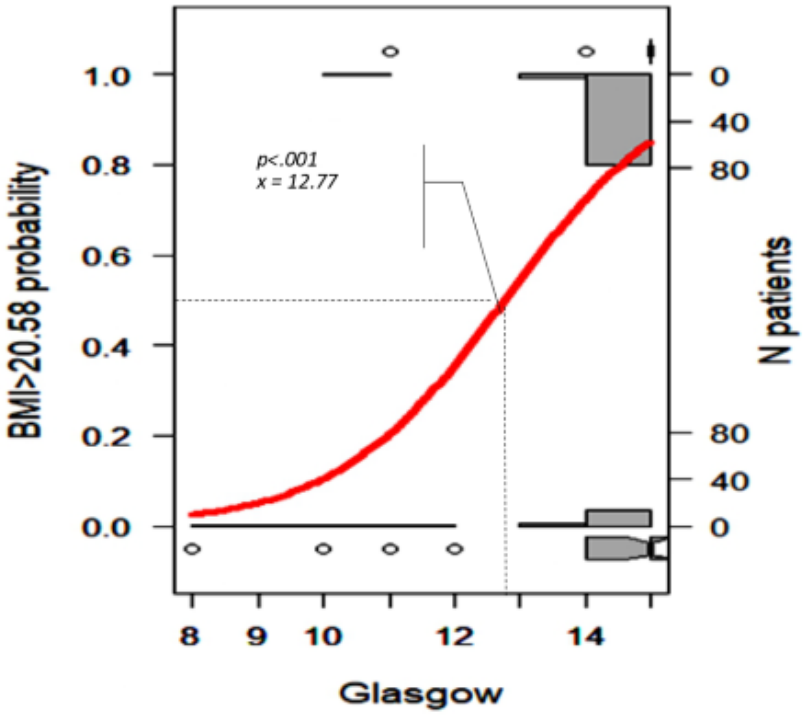


Figure 16. Significant cutpoint 50% of the Glasgow Coma Scale score for patient BMI >20.58 kg/m2.



Gene Variants Might be Associated With the Patient Outcome via the Ischemic Stroke Score

We calculated the risk ratios (RRs) and CIs by unconditional maximum likelihood estimation and normal approximation, respectively (Wald), as well as performed minor sample adjustment by the Mantel Haenszel method, generating P_{yates}, P_{uncor}, and P_{fisher}. We grouped these genotype variants following their clusters, which provided the most relevant RR results (Table S4 in Multimedia Appendix 1; Figure 17; Interactive Graph 7 [20]). The detailed RRs for stroke scores are presented in Table S5 in Multimedia Appendix 1. Forest plots were created for clusters 4 (Figures S1-S5 in Multimedia Appendix 2), 6 (Figures S6-S9 in Multimedia Appendix 2), 11 (Figures S10-S13 in Multimedia Appendix 2), and 13 (Figures S14-S17 in Multimedia Appendix 2).

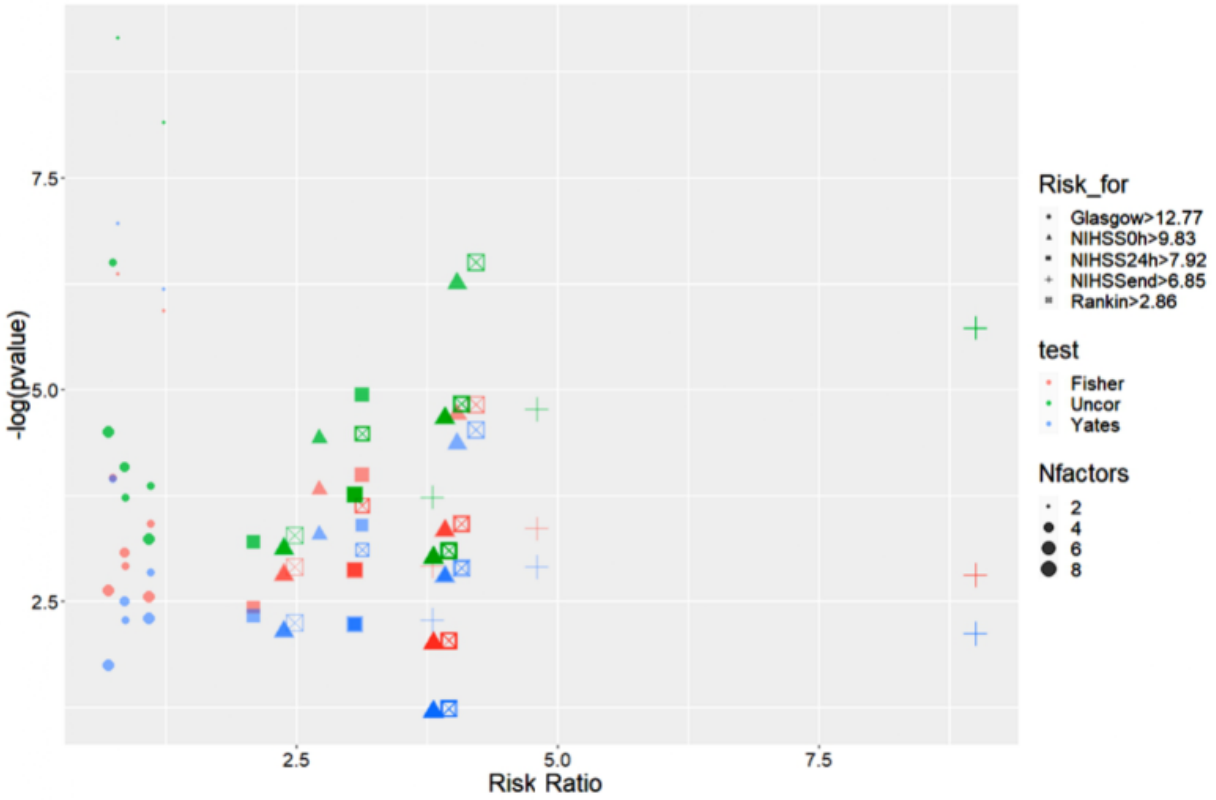
The GCS can be used for head injury, and score ranges are used to describe the injury severity. Scores of 13-15 indicate mild traumatic brain injury, 9-12 indicate moderate traumatic brain injury, and 3-8 indicate severe traumatic brain injury. The risk of experiencing mild traumatic brain injury (cutpoint 50% of GCS was 12.77) was 23% higher in the group of patients without diabetes and with a BMI greater than 20.8 kg/m^2 as well as *NOTCH3* heterozygous mutation, *MTHFR-C677T*, and FI-Prothrombin than in the other groups (RR=1.23, 95% CI 0.99-1.54; P_{fisher}=2.68×10^{-3}). This risk was 20% lower in the group of patients with BMI less than 20.8 kg/m^2 and with *MTHFR-A1298C* and *FV-H1299R* wildtype variants than in the other groups (RR=0.79, 95% CI 0.61-1.01; P_{fisher}=1.72×10^{-3}).

The NIHSS quantifies the impairment caused by stroke and aids in planning post-acute care disposition, although it has been intended to assess differences in interventions in clinical trials. A NIHSS score of 0 indicates no stroke symptoms, 1-4 indicates minor stroke, 5-15 indicates moderate stroke, 16-20 indicates moderate to severe stroke, and 21-42 indicates severe stroke. The risk of a NIHSS score at admission greater than 9.83 and a NIHSS score at 24 hours greater than 7.92 (moderate stroke) was higher in the group of patients with age older than 54 years, height shorter than 161 cm, PT time ≤13.25 seconds, PT ratio ≤99, creatinine >83.67 µmol/L, and *FXIII Val34Leu* wildtype than in the other groups (RR=2.72, 95% CI 1.4-5.31 and RR=2.09, 95% CI 1.1-3.93, respectively; P_{fisher}=2.19×10^{-2} and 8.81×10^{-2}, respectively). The risk of a NIHSS score at discharge greater than 6.85 (moderate stroke) was higher in the group of patients with age older than 54 years, height taller than 161 cm, PT time ≤13.25 seconds, PT ratio ≤99, creatinine >83.67 µmol/L, *FII Prothrombin* and *MTHFR-C677T* wildtype, and *NOTCH3* p.R544C heterozygous (RR=4.8, 95% CI 1.53-15.04; P_{fisher}=3.47×10^{-2}).

The mRS is an outcome measure in stroke clinical trials. The mRS assessment is recommended 3 months (90 days) following hospital discharge. The mRS score is assigned as follows: 0, patient has no residual symptoms; 1, patient has no significant disability and has ability to carry out all prestroke activities; 2, patient has remote disability and is incapable of carrying out all prestroke movements but is capable of looking after self without daily help; 3, patient has moderate disability and needs some external help but is capable of walking without the assistance of another individual; 4, patient has moderately severe disability and is incapable of walking or performing physical

functions without the aid of another individual; 5, patient has severe disability, is bedridden, shows incontinence, and requires continuous care; 6, patient has passed away (during the hospital stay or after discharge from the hospital); 7, inability to contact the patient or caregiver; and 8, score not achieved or not determined from the medical records. The risk of a mRS score greater than 2.86 (moderate disability) was higher in the group of patients with INR >1.02, PT time >13.25 seconds, PT ratio ≤99, creatinine >83.67 µmol/L, *FXIII Val34Leu* wildtype (in case the number of infarcts on CT was greater than 2), *MTHFR-A1298C* heterozygous/wildtype, and *FV-H1299R* wildtype (RR=3.13, 95% CI 1.6-6.11; $P_{fisher}=2.64\times10^{-2}$).

Figure 17. Dot plot of the genotype variants according to their clusters, which provides the most relevant risk ratio results. NIHSS: National Institutes of Health Stroke Scale.



Discussion

Principal Findings

Some sophisticated techniques for HCA exploit statistical frameworks called hierarchical models or multilevel models. Hierarchical models are useful in a number of contexts. HCA, which is also known as hierarchical clustering, is a popular method for cluster analysis in big data research and data mining aiming to establish a hierarchy of clusters. As such, HCA attempts to group subjects with similar features into clusters. Clustering is a data science technique in machine learning that groups similar rows in a data set. After running a clustering technique, a new column appears in the data set to indicate the group each row of data fits into the best.

Several gene mutations have been identified as leading causes of cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL), a hereditary disease that causes stroke and other neurological symptoms. CADASIL accounts for up to 5% of all strokes in individuals aged younger than 65 years. The thrombophilia test helps determine the disease's genetic origin to provide appropriate prevention and treatment measures. Hypercoagulation syndrome may be due to mutations in genes encoding proteins related to blood clotting (thrombophilia). People with hypercoagulable syndrome tend to form blood clots in blood vessels (primarily veins), resulting in stroke, heart attack, repeated miscarriages, and complications during pregnancy (pre-eclampsia, fetal growth retardation, and stillbirth) [21].

In our study, gene variants were assessed to understand how ischemic stroke genetics could interest practitioners and be useful for clinical work. The variants were as follows: *FII Prothrombin*, *FV-Leiden*, *MTHFR-C677T*, *MTHFR-A1298C*, *FV-H1299R*, *PAI1 4G/5G*, *FXIII Val34Leu*, *FV-Cambridge*, *and NOTCH3* p.R544C.

We visualized how these risk factors and genetic elements could affect ischemic stroke outcomes with a hierarchical analysis strategy. Maximally selected rank statistics help to define the optimal thresholds of several continuous factors (creatinine, age, PT time and ratio, INR, LDL-C, number of infarcts on CT or MRI, patient height, and MPV) based on the mRS, NIHSS, and GCS scores and their related symptom statuses, such as numbness, dizziness, gender, circular muscle disorder, mouth distortion, and diabetes status. Their optimal cutpoints fitted with the normal range in both genders. The creatinine level of

83.67 (SD 9.19) µmol/L is consistent with the usual results of 0.7 to 1.3 mg/dL (61.9 to 114.9 µmol/L) for men and 0.6 to 1.1 mg/dL (53 to 97.2 µmol/L) for women [22]. Our age threshold was 54 (SD 5) years, which is consistent with the findings worldwide, with aging being the most robust nonmodifiable risk factor for incident stroke (risk doubles every 10 years after the age of 55 years) [23]. Assessment of the PT time is recommended for the administration of recombinant tissue-plasminogen activator (rt-PA) in stroke [24]. The standard range of the PT time is 10 to 13 seconds. The usual INR for a healthy individual is 1.1 or below, and the therapeutic range for most patients on vitamin K antagonists is 2.0 to 3.0. An augmented PT/INR for patients on vitamin K antagonists may suggest a super-therapeutic status and will need prescription dose adjustments to control bleeding [25]. In our study, the calculated baseline PT time was 13.25 (SD 0.17) and INR was 1.02 (SD 0.03), which confirmed cases of moderate outcomes. Data on the association between BMI and stroke are scarce. Individuals with a BMI of 18.5 to 24.9 kg/m^2 are considered to have a healthy weight. Our calculated baseline BMI was 20.85 kg/m^2, and it was associated with genetic factors that influence the GCS score.

According to the Nagelkerke method, the cutpoint 50% values of the mRS score and NIHSS scores at admission, after 24 hours, and at discharge were 2.86 (SD 1.21), 9.83 (SD 2.85), 7.29 (SD 2.04), and 6.85 (SD 2.90), respectively, which were consistent with the moderate outcomes of our patients. We found that the MTHFR and *NOTCH3* p.R544C variants may influence stroke severity in patients with specific conditions of PT, creatinine, INR, and BMI.

The MTHFR gene provides instructions for the human body to make the MTHFR protein, which helps the body process folate, which is important for forming DNA and modifying proteins. The most common variant of the MTHFR gene is *MTHFR-C677T* [26]. This mutation causes a reduction in the capacity to create L-methylfolate. *MTHFR-A1298C* single-nucleotide polymorphism has also been suggested to have an impact on MTHFR enzyme activity but to a lesser extent than the *MTHFR-C677T* polymorphism. They have been recently shown to be associated with ischemic stroke [27].

CADASIL is an autosomal dominant inherited vasculopathy and is the most common single‐gene disorder causing stroke, with more than 200 different *NOTCH3* p.R544C mutations in patients worldwide, indicating that CADASIL has considerable genetic heterogeneity. The defective 33‐exon *NOTCH3* p.R544C gene is located on chromosome 19, which typically impacts the number of highly conserved cysteine residues among the epidermal growth factor–like repeat domain [28].

HCA is attractive for exploratory high-throughput data because it provides a convenient approach to visualize the similarities of variables and infer the grouping of variables based on the dendrogram structure. Hence, HCA facilitates the interpretation of the data of the microbiome and other omics. Importantly, bi-clustering (2-way clustering), a particular approach of HCA, can incorporate a correlation method (eg, Spearman rank correlation) to cluster rows and columns of the data matrix simultaneously. Thus, bi-clustering can find features (microbial taxa, genes, metabolites, etc) that correlate only in a subset of objects but not in the rest of the data set [29]. In this study, we clearly identified the role and interaction of risk factors that influence stroke progression. Genetic mutations become significant in a small range of strongly correlated factors through a PCA plot.

Stroke has multiple modifiable and nonmodifiable risk factors and represents a leading cause of death globally. Understanding the complex interplay of stroke risk factors is thus not only a scientific necessity but also a critical step toward improving global health outcomes [30].

Limitations

We found that 3 of the 9 gene variants had significant RRs. Data settings could help to work with both qualitative and numerical data simultaneously. The main advantage of the HCA clustering concept is the display of possible correlations between several factors to provide reference markers that are useful for diagnostic control and to improve outcome prevention. It was beneficial to identify the association between genetic characteristics and clinical outcomes, which usually requires several in vitro studies; however, there were some constraints. It is critical to clean and prepare the data set because HCA and k-means cannot operate with missing or noisy data. We must combine and validate the data with k-means, which provides several options for the optimal cluster number to produce a PCA cluster plot and define the principal component position. Since our data had various kinds of information, it was challenging to calculate the distance matrix in HCA and k-means.

Conclusions

The existence of conventional vascular risk factors may prevent clinicians from suspecting the possibility of gene mutations in stroke patients, especially among those with underlying atrial fibrillation or extensive artery atherosclerosis. In this study, a more specific population was chosen. It is interesting that although there are many genes linked to increased atrial fibrillation risk, not all of them are associated with ischemic stroke risk, which might be because those gene variants are too rare to detect their impacts on stroke risk. Nevertheless, in the future, the identification of a linkage between some of those genes and ischemic stroke could be a significant game changer in the field of stroke prevention. Moreover, with the detection of stroke risk loci, more information can be gained on their impacts and interconnections, and the precision of stroke scores might increase.

Acknowledgments

We sincerely thank the patients and the participants for consenting and permitting us to publish the data. We thank our colleagues at National Hospital of Thai Nguyen, who supported us with patient recruitment and clinical data. This research was funded by the Ministry of Education and Training, Vietnam (grant number 3813/QĐ-BGDĐT; November 20, 2020).

Data Availability

The data sets used or analyzed during this study are available from the corresponding author on reasonable request. We however cannot provide personal information or data that contain the identity of patients in any form.

Authors' Contributions

HTTB and HCT designed the study. HTTB received a grant for the study. SNP, TTP, QNTP, HNTT, and TV performed the data collection and the experiments. HCT and LKH performed the data mining and hierarchical cluster analysis study. HTTB and HCT wrote the main manuscript. DNT revised the manuscript and supervised the study. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary results.
[PDF File (Adobe PDF File), 669 KB - bioinform_v5i1e56884_app1.pdf]

Multimedia Appendix 2
Forest plots.
[PDF File (Adobe PDF File), 1122 KB - bioinform_v5i1e56884_app2.pdf]

References

1. GBD 2019 Stroke Collaborators. Global, regional, and national burden of stroke and its risk factors, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Neurol 2021 Oct;20(10):795-820 [FREE Full text] [doi: 10.1016/S1474-4422(21)00252-0] [Medline: 34487721]
2. Mai DT, Dao XC, Luong NK, Nguyen TK, Nguyen HT, Nguyen TN. Current State of Stroke Care in Vietnam. SVIN 2022 Mar;2(2):e000331. [doi: 10.1161/SVIN.121.000331]
3. Chugh C. Acute Ischemic Stroke: Management Approach. Indian J Crit Care Med 2019 Jun;23(Suppl 2):S140-S146 [FREE Full text] [doi: 10.5005/jp-journals-10071-23192] [Medline: 31485123]
4. Ilinca A, Samuelsson S, Piccinelli P, Soller M, Kristoffersson U, Lindgren A. A stroke gene panel for whole-exome sequencing. Eur J Hum Genet 2019 Feb;27(2):317-324 [FREE Full text] [doi: 10.1038/s41431-018-0274-4] [Medline: 30356112]
5. International Stroke Genetics Consortium (ISGC), Wellcome Trust Case Control Consortium 2 (WTCCC2), Bellenguez C, Bevan S, Gschwendtner A, Spencer C, et al. Genome-wide association study identifies a variant in HDAC9 associated with large vessel ischemic stroke. Nat Genet 2012 Feb 05;44(3):328-333 [FREE Full text] [doi: 10.1038/ng.1081] [Medline: 22306652]
6. Lin Y, Dong J, Tang Y, Ni H, Zhang Y, Su P, et al. Opening a New Time Window for Treatment of Stroke by Targeting HDAC2. J Neurosci 2017 Jul 12;37(28):6712-6728 [FREE Full text] [doi: 10.1523/JNEUROSCI.0341-17.2017] [Medline: 28592694]
7. Ghezali L, Capone C, Baron-Menguy C, Ratelade J, Christensen S, Østergaard Pedersen L, et al. Notch3 immunotherapy improves cerebrovascular responses in CADASIL mice. Ann Neurol 2018 Aug;84(2):246-259. [doi: 10.1002/ana.25284] [Medline: 30014602]
8. Sargurupremraj M, Suzuki H, Jian X, Sarnowski C, Evans T, Bis J, International Network against Thrombosis (INVENT) Consortium, International Headache Genomics Consortium (IHGC), et al. Cerebral small vessel disease genomics and its implications across the lifespan. Nat Commun 2020 Dec 08;11(1):6285 [FREE Full text] [doi: 10.1038/s41467-020-19111-2] [Medline: 33293549]
9. Ekkert A, Šliachtenko A, Grigaitė J, Burnytė B, Utkus A, Jatužis D. Ischemic Stroke Genetics: What Is New and How to Apply It in Clinical Practice? Genes (Basel) 2021 Dec 24;13(1):48 [FREE Full text] [doi: 10.3390/genes13010048] [Medline: 35052389]

10. Irie F, Matsumoto K, Matsuo R, Nohara Y, Wakisaka Y, Ago T, et al. Predictive Performance of Machine Learning–Based Models for Poststroke Clinical Outcomes in Comparison With Conventional Prognostic Scores: Multicenter, Hospital-Based Observational Study. JMIR AI 2024 Jan 11;3:e46840. [doi: 10.2196/46840]
11. Nguyen TT, Ho CT, Bui HTT, Ho LK, Ta VT. Multidimensional Machine Learning for Assessing Parameters Associated With COVID-19 in Vietnam: Validation Study. JMIR Form Res 2023 Feb 16;7:e42895 [FREE Full text] [doi: 10.2196/42895] [Medline: 36668902]
12. Owen AB. Karl Pearson's meta-analysis revisited. Ann. Statist 2009 Dec 1;37(6B):3867-3892. [doi: 10.1214/09-AOS697]
13. Interactive Graph 1 (created by Hocamtu). Plotly Chart Studio. URL: https://chart-studio.plotly.com/~hocamtu/9/#/ [accessed 2024-04-10]
14. Interactive Graph 2 (created by Hocamtu). Plotly Chart Studio. URL: https://chart-studio.plotly.com/~hocamtu/11/#/ [accessed 2024-04-10]
15. Interactive Graph 3 (created by Hocamtu). Plotly Chart Studio. URL: https://chart-studio.plotly.com/~hocamtu/13/#/ [accessed 2024-04-10]
16. Interactive Graph 4 (created by Hocamtu). Plotly Chart Studio. URL: https://chart-studio.plotly.com/~hocamtu/15/#/ [accessed 2024-04-10]
17. Interactive Graph 5 (created by Hocamtu). Plotly Chart Studio. URL: https://chart-studio.plotly.com/~hocamtu/3/#/ [accessed 2024-04-10]
18. Desgraupes B. Clustering Indices. Paris, France: University of Paris Ouest - Lab Modal'X; 2016.
19. Interactive Graph 6 (created by Hocamtu). Plotly Chart Studio. URL: https://chart-studio.plotly.com/~hocamtu/7/#/ [accessed 2024-04-10]
20. Interactive Graph 7 (created by Hocamtu). Plotly Chart Studio. URL: https://chart-studio.plotly.com/~hocamtu/1/#/ [accessed 2024-04-10]
21. Favaloro EJ. Genetic Testing for Thrombophilia-Related Genes: Observations of Testing Patterns for Factor V Leiden (G1691A) and Prothrombin Gene "Mutation" (G20210A). Semin Thromb Hemost 2019 Oct 09;45(7):730-742. [doi: 10.1055/s-0039-1694772] [Medline: 31398733]
22. Hosten AO. BUN and Creatinine. In: Walker HK, Hall WD, Hurst JW, editors. Clinical Methods: The History, Physical, and Laboratory Examinations. 3rd Edition. Boston: Butterworths; 1990.
23. Yousufuddin M, Young N. Aging and ischemic stroke. Aging (Albany NY) 2019 May 01;11(9):2542-2544 [FREE Full text] [doi: 10.18632/aging.101931] [Medline: 31043575]
24. Gottesman RF, Alt J, Wityk RJ, Llinas RH. Predicting abnormal coagulation in ischemic stroke: reducing delay in rt-PA use. Neurology 2006 Nov 14;67(9):1665-1667. [doi: 10.1212/01.wnl.0000244493.13898.5b] [Medline: 17101901]
25. Winter W, Flax S, Harris N. Coagulation Testing in the Core Laboratory. Lab Med 2017 Nov 08;48(4):295-313. [doi: 10.1093/labmed/lmx050] [Medline: 29126301]
26. Crider KS, Yang TP, Berry RJ, Bailey LB. Folate and DNA methylation: a review of molecular mechanisms and the evidence for folate's role. Adv Nutr 2012 Jan;3(1):21-38 [FREE Full text] [doi: 10.3945/an.111.000992] [Medline: 22332098]
27. Chita DS, Tudor A, Christodorescu R, Buleu FN, Sosdean R, Deme SM, et al. Gene Polymorphisms Prevalence and Cardiovascular Risk Factors Involved in Cardioembolic Stroke Type and Severity. Brain Sci 2020 Jul 24;10(8):476 [FREE Full text] [doi: 10.3390/brainsci10080476] [Medline: 32722170]
28. Tang S, Chen Y, Chi N, Chen C, Cheng Y, Hsieh F, et al. Prevalence and clinical characteristics of stroke patients with p.R544C mutation in Taiwan. Ann Clin Transl Neurol 2019 Jan 20;6(1):121-128 [FREE Full text] [doi: 10.1002/acn3.690] [Medline: 30656190]
29. Xia Y. Correlation and association analyses in microbiome study integrating multiomics in health and disease. Prog Mol Biol Transl Sci 2020;171:309-491. [doi: 10.1016/bs.pmbts.2020.04.003] [Medline: 32475527]
30. Lolak S, Attia J, McKay GJ, Thakkinstian A. Comparing Explainable Machine Learning Approaches With Traditional Statistical Methods for Evaluating Stroke Risk Models: Retrospective Cohort Study. JMIR Cardio 2023 Jul 26;7:e47736 [FREE Full text] [doi: 10.2196/47736] [Medline: 37494080]

Abbreviations

CADASIL: cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy
CT: computed tomography
CTA: computed tomography angiography
CTPP: confronting 2-pair primers
GCS: Glasgow Coma Scale
HCA: hierarchical cluster analysis
HDAC: histone deacetylase
INR: international normalized ratio
LDL-C: low-density lipoprotein cholesterol
MPV: mean platelet volume

MRI: magnetic resonance imaging
mRS: modified Rankin scale
NIHSS: National Institutes of Health Stroke Scale
PCA: principal component analysis
PCR: polymerase chain reaction
PT: prothrombin
RR: risk ratio

Edited by S Tian; submitted 29.01.24; peer-reviewed by L Guo, A Kalluchi; comments to author 27.02.24; revised version received 02.03.24; accepted 02.04.24; published 07.05.24.

Please cite as:
Bui HTT, Nguy n Th Ph ng Q, Cam Tu H, Nguyen Phuong S, Pham TT, Vu T, Nguyen Thi Thu H, Khanh Ho L, Nguyen Tien D
The Roles of NOTCH3 p.R544C and Thrombophilia Genes in Vietnamese Patients With Ischemic Stroke: Study Involving a Hierarchical Cluster Analysis
JMIR Bioinform Biotech 2024;5:e56884
URL: https://bioinform.jmir.org/2024/1/e56884
doi:10.2196/56884
PMID:38935968

Review

Assessing Privacy Vulnerabilities in Genetic Data Sets: Scoping Review

Mara Thomas[1], PhD; Nuria Mackes[2], PhD; Asad Preuss-Dodhy[3], PhD; Thomas Wieland[4], PhD; Markus Bundschus[3], PhD

[1]F. Hoffmann-La Roche AG, Basel, Switzerland
[2]xValue GmbH, Ratingen, Germany
[3]Roche Diagnostics GmbH, Penzberg, Germany
[4]Foundation Medicine GmbH, Penzberg, Germany

Corresponding Author:
Mara Thomas, PhD
F. Hoffmann-La Roche AG
Grenzacherstrasse 124
Basel, 4070
Switzerland
Phone: 41 616881111
Email: mara.thomas@roche.com

Abstract

Background: Genetic data are widely considered inherently identifiable. However, genetic data sets come in many shapes and sizes, and the feasibility of privacy attacks depends on their specific content. Assessing the reidentification risk of genetic data is complex, yet there is a lack of guidelines or recommendations that support data processors in performing such an evaluation.

Objective: This study aims to gain a comprehensive understanding of the privacy vulnerabilities of genetic data and create a summary that can guide data processors in assessing the privacy risk of genetic data sets.

Methods: We conducted a 2-step search, in which we first identified 21 reviews published between 2017 and 2023 on the topic of genomic privacy and then analyzed all references cited in the reviews (n=1645) to identify 42 unique original research studies that demonstrate a privacy attack on genetic data. We then evaluated the type and components of genetic data exploited for these attacks as well as the effort and resources needed for their implementation and their probability of success.

Results: From our literature review, we derived 9 nonmutually exclusive features of genetic data that are both inherent to any genetic data set and informative about privacy risk: biological modality, experimental assay, data format or level of processing, germline versus somatic variation content, content of single nucleotide polymorphisms, short tandem repeats, aggregated sample measures, structural variants, and rare single nucleotide variants.

Conclusions: On the basis of our literature review, the evaluation of these 9 features covers the great majority of privacy-critical aspects of genetic data and thus provides a foundation and guidance for assessing genetic data risk.

(JMIR Bioinform Biotech 2024;5:e54332) doi:10.2196/54332

KEYWORDS

genetic privacy; privacy; data anonymization; reidentification

Introduction

Privacy Risks of Genetic Data

Genomics is a rapidly developing field with exabytes of genetic data being generated, stored, and analyzed by public and private institutions per year. These data drive scientific progress, especially when they are shared with the scientific community or among institutions. However, genetic data can provide information about an individual's identity together with sensitive details, such as their ethnic background [1]; physical traits such as eye color [2], hair and skin color [3], height [4]; and diseases or susceptibility to diseases [5]. Therefore, even if personal identifiers (eg, name, date of birth, or others) are removed, sharing genetic data may violate the individual's right to privacy. In 2018, a seminal study demonstrated that it is possible to reidentify individuals by name from genetic data alone [6]. The authors matched genetic data of an anonymous female study participant to the genetic genealogy database GEDmatch and

identified her surname from matches with relatives who had uploaded their data on GEDmatch. Such reidentification of genetic data records using publicly available databases is highly problematic and a growing threat to privacy as publicly available genetic genealogy databases continue to grow. It is estimated that a genetic database needs to cover "only 2% of the target population to provide a third-cousin match to nearly any person" in a matching attack, similar to the one demonstrated by Erlich et al [6]. As of 2018, the probability for such a match was estimated to be 60% for the platform GEDmatch. Through similar methods of familial DNA searches, multiple individuals have been identified in criminal cases, despite never having shared their genetic data themselves [7,8]. Other attacks aim to reveal sensitive information from genetic data. In 2009, researchers discovered a genetic predisposition for Alzheimer disease in the public genome of the famous molecular biologist and Nobel laureate James Watson, although he had attempted to prevent such an attack by withholding certain parts of the data [9]. The high identifiability potential of genetic data together with its sensitive content with regard to health (eg, susceptibility to diseases such as Alzheimer disease or cancer) and physical traits (refer to the studies by Erlich and Narayanan [10], El Emam et al [11], and Mohammed Yakubu and Chen [12] for a review) has raised public concern that genetic data that are shared or published in the context of research or health care could be misused [13]. For example, attackers could exploit genetic data to obtain personal and sensitive information about individuals, and this information could be misused by insurance companies, mortgage providers, or employers to discriminate on the basis of genetic information (eg, about disease susceptibility) [14]. As an additional complication, DNA sequence is heritable; therefore, leakage of an individual's genetic data can violate the privacy of whole families [15,16].

The Challenge of Anonymizing Genetic Data

Genetic data can be used to identify individuals because each person's DNA sequence differs uniquely from the standard human reference genome. Although more than 99% of the DNA sequence is identical across all humans, the remaining <1% consists of distinct combinations of insertions, deletions, duplications, translocations, and inversions of short or long DNA fragments (refer to the study by Trost et al [17] for a review). These genetic variations are not randomly distributed across the genome but occur more frequently in specific variable regions. Some variations are rare, while others (ie, polymorphisms) are shared by a significant proportion of the population. While some variations have no observable effect, others influence gene transcription, expression, or the amino acid sequence of a protein and have an effect on the phenotype, for example, physical traits, metabolism, and disease susceptibility. These variable regions with an effect on the phenotype are of great interest to research; however, these can also be effectively used for individual identification and the inference of sensitive attributes. Even a small genetic data set of only 30 highly variable genetic loci is likely to contain unique records, and these could not only be linked to genetic records in other data sets but also provide insights into health and physical traits (refer to the studies by Erlich and Narayanan [10], El Emam et al [11], and Mohammed Yakubu and Chen [12] for a review). Furthermore, genetic variation is highly intercorrelated (variation in one genomic region correlates with variation in another) and correlated to other modalities (genetic variation is associated with transcription, expression, epigenetic regulation, etc), making it possible to link data records of the same individual even across databases that do not contain the same type of data (eg, match a genetic data sequence to a gene expression record). Anonymizing genetic data while maintaining its full utility remains an unsolved challenge, and there is no consensus on whether it is even possible [18]. Many privacy-enhancing technologies aim to reduce the information content of genetic data or restrict access to it, such that only a minimal amount of information is shared. An example is genomic beacons, which allow only simple yes or no queries to determine whether a specific variant is present in a study cohort [19]. However, it has become evident that even this limited amount of information can be exploited for privacy attacks, and few queries to genomic beacons can suffice to determine whether individuals (whose genome is known) are present in a study cohort [20-23]. Similarly, proposals for encryption and differential privacy approaches [24,25] have often been countered by demonstrations of attacks [26-28], and even synthetic genetic data may not fully protect the study participants from privacy attacks [29] (refer to the study by Mittos et al [30] for a review of privacy-enhancing technologies). Thus, even a substantial reduction in information content can often not completely eliminate all privacy risks of genetic data [31].

The Risk Minimization Approach for Genetic Data Privacy

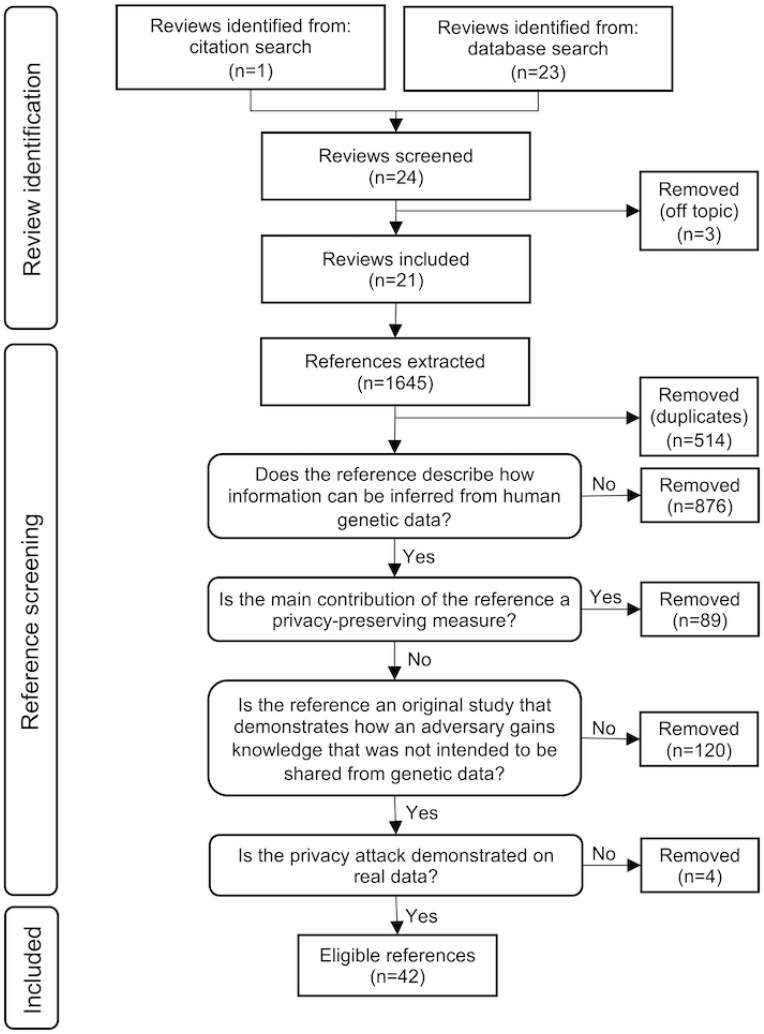
Most legislations do not require to reduce the risk of individual identification to zero, and several jurisdictions have decided to take a risk-based approach and consider genetic data anonymous if the risk of successful reidentification is below a predefined acceptable threshold [32]. Therefore, genetic data processors must find the balance between reducing information such that reidentification is no longer reasonably likely, while maintaining as much utility of the data as possible [33]. The challenge in adopting this approach lies in the correct assessment of the reidentification probability. Genetic data are complex and come in various shapes or forms, making it difficult to standardize reidentification assessments. Established methods such as assessing k-anonymity are difficult to apply to genetic data because of their high uniqueness, and many other methods fall short because of the high intercorrelation of genetic data. Simple measures such as assessing the number of single nucleotide polymorphisms (SNPs) in genetic data ignore the importance of the location of the SNPs in the genome, their frequencies in the population, and the actual feasibility of cross-linking the specific SNPs to identifiable information. For example, the reidentification risk is much higher for SNPs that are commonly included in the SNP assays used by direct-to-consumer genetic testing (DTC-GT) providers than for less frequently studied SNPs, as these are more difficult to link to publicly available identifying information. In addition, genetic data may contain SNP information even if this is not immediately evident, for example, in the raw data of sequencing-based gene expression studies. Data processors who are not familiar with the intricacies

of genetic data find little guidance on performing an assessment on genetic data that considers these factors. While several genomic privacy metrics have been proposed, the great majority focus on evaluating SNPs only [34] and neglect other known privacy-critical aspects of genetic data as well as aspects of feasibility (eg, the expertise, time, effort, availability of external resources, and other requirements required for an attack). However, the risk of severe privacy attacks on genetic data (ie, where the identity of the data subject is revealed) greatly depends on the specific content of the data as well as "soft factors," such as the availability of publicly accessible resources to cross-link and infer quasi-identifying information and the time, cost, and knowledge required to perform such an attack. Given the foundational potential of genetic data to advance research and health care, a risk-based approach that carefully evaluates the true risk of reidentification on a case-by-case basis for each data set in question is warranted, or else any type of genetic data must be considered identifiable.

Methods

To get a comprehensive overview of the types and aspects of genetic data sets that are vulnerable to reidentification attacks, as well as the methods, databases, and know-how used for these attacks, we searched for studies that demonstrate a privacy attack on genetic data. We did not aim to establish an exhaustive overview of all published privacy attacks but aimed to get a comprehensive understanding of the most vulnerable features of genetic data. Therefore, we first searched for recent reviews published on the topic of genomic privacy using ProQuest. Using the search terms (ti(*genom* OR *genetic*) AND ti(privacy OR re-identification OR reidentification OR "data security")) and (pd(>20170101)) and (at.exact("Review")), we identified 23 reviews, of which 3 (13%) were discarded because they were off topic. One additional review was identified during the literature research and added to the selection (refer to Multimedia Appendix 1 [35-55] for an overview of the included and excluded reviews), resulting in a final sample of 21 reviews. In a second step, we extracted all references cited in the reviews (n=1645) and identified all original research studies that demonstrate a privacy attack on genetic data. After the removal of 514 duplicates and 876 reference studies that did not contain any description of information inference from human genetic data, we first excluded 89 studies whose main contribution was the presentation of privacy-preserving measures to exclude privacy attacks that were performed only for the purpose of proving the efficiency of the proposed counter methods. Next, we excluded 120 studies that did not present original research and were purely associative (ie, did not demonstrate how an adversary gains knowledge that was not intended to be shared from genetic data) as well as 4 studies that did not demonstrate the attack on real data. This process resulted in the selection of 42 unique studies (refer to Figure 1 for an overview of the process and Table S1 in Multimedia Appendix 1 for an overview of the eligible attack studies). Extending on the framework by Mohammed Yakubu and Chen [12] and Lu et al [56], we categorized attacks into (1) identity tracing (attacker triangulates the identity of an individual), (2) inference (attacker uses an individual's genetic data to infer sensitive attributes such as disease or drug abuse or to infer additional data or cross-link records across databases), and (3) membership attacks (attacker uncovers membership of an individual in a data set). We evaluated the type and components of genetic data exploited for this attack as well as the effort and resources used for it (time, expertise, databases, and computation power) and its success rate if sufficient information was reported in the study. The initial evaluation was conducted by one reviewer and independently verified by another. Table S1 in Multimedia Appendix 1 presents a detailed overview of the attack studies.

Figure 1. Flowchart overview of the 2-step literature review process: identification of relevant reviews, followed by extraction and screening of references.



Results

A Comprehensive Overview of Privacy Risks in Genetic Data Sets

On the basis of our literature review, we created an overview of the parts and aspects of genetic data that are commonly exploited in privacy attacks and that should therefore be taken into consideration when performing a risk assessment on genetic data. The goal of this overview is to provide data processors, who may not be experts in genomic data privacy, with essential background knowledge about the privacy vulnerabilities associated with genetic data. This understanding will help them identify privacy-critical aspects and serve as a starting point for conducting risk assessments on genetic data sets. Notably, the reidentification risks associated with data that complement genetic data (eg, clinical data and demographic data) as well as aspects of the data environment (access and governance) are crucial for a comprehensive risk assessment [57], but these aspects are not in the scope of this research. From our literature review, we synthesized 9 features that are both inherent to any genetic data and informative about privacy risk (Figure 2). The features are not mutually exclusive. Instead, they represent different "views" on genetic data and highlight various aspects that should be considered in a privacy risk assessment. For each feature, we lay out why this feature is associated with privacy risk by summarizing the relevant evidence in the scientific literature, and we assess the criticality of these attacks. In addition, we provide guiding questions that help to assess the risk of a given data set. The features can be divided into three groups:

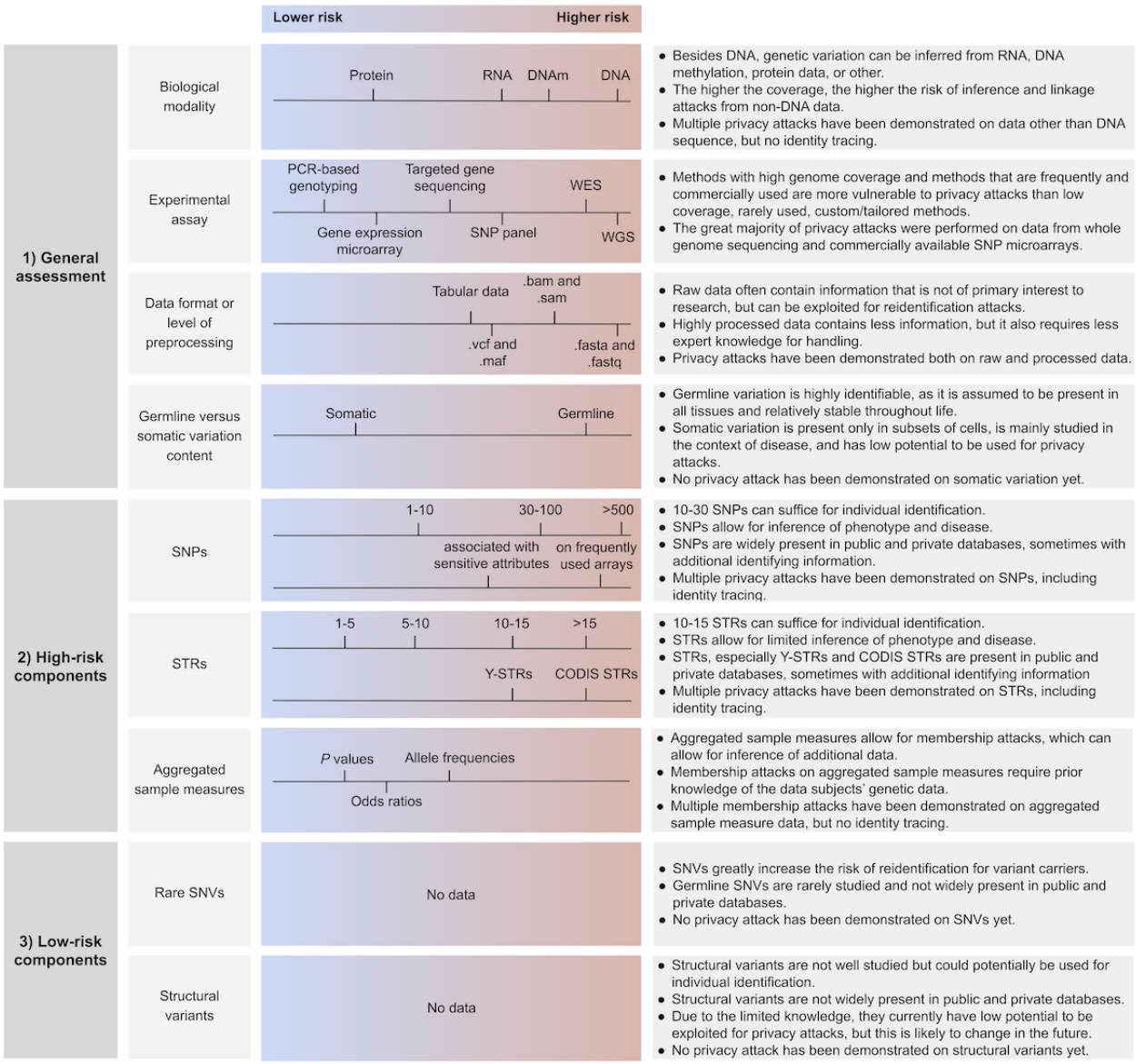
1. The first 4 features are general categorizations of the genomic data set and serve as a very rough estimate of the amount of privacy-critical information in the data.
2. The next 3 features are specific genomic features that are known to be a high risk for privacy. Their assessment is critical for estimating the reidentification risk.
3. The last 2 features are genomic features that have not been exploited for privacy attacks yet but should still be considered and could present a risk if they are present to a high degree in the data.

We summarize our findings in an overview figure, which lists the 9 features and their relevance for privacy. While it is challenging to define clear risk thresholds, there is a recognized need for practical guidance and orientation. To address this, we

provide a scale that ranges from lower to higher risk and offer illustrative examples derived from the overview of privacy attack studies. These scales and examples serve as the initial guidance for risk assessment, emphasizing their purpose as guiding principles rather than exact measurements. The assessment of each individual feature is intricate and thoroughly explained in the corresponding sections. In addition, while the scales offer a framework to compare and assess different features, it is crucial to consider all features comprehensively to arrive at a conclusive assessment. Furthermore, the text sections highlight important interactions that arise from the comprehensive evaluation of these features.

Table S1 in Multimedia Appendix 1 presents a detailed description of the original attack studies.

Figure 2. Overview of the privacy-critical features of genetic data sets, with exemplary values and key points to consider for risk assessment. CODIS: Combined DNA Index System; SNP: single nucleotide polymorphism; SNV: single nucleotide variant; STR: short tandem repeat; WES: whole exome sequencing; WGS: whole genome sequencing; Y-STR: short tandem repeat on the Y chromosome.



Evidence of Privacy Risks in Genetic Data

Part 1. General Assessment

Biological Modality

While most privacy attacks have been demonstrated on DNA sequence data, other types of molecular data (eg, DNA methylation data or data derived from RNA) are also considered genetic data under General Data Protection Regulation, can also be identifiable, and have also been exploited for attacks [58-67]. Attacks on these types of data are performed mainly by 3 mechanisms. The first mechanism is direct extraction of DNA sequence from raw or low-processed data. This is possible, because even if not of primary interest, DNA sequence information is often a by-product of gene expression or DNA methylation studies [68-70]. For example, Gürsoy et al [70] demonstrated how genetic variants can be called from raw RNA sequencing data. The second mechanism is inference of DNA sequence, for example, through known associations of genetic sequence and gene expression or other modalities. For example,

Schadt et al [65] used gene expression data of individuals (40,000 transcript counts) to infer genetic variants (1000 SNPs), which allowed them to determine with high certainty whether individuals with known SNPs were members of a gene expression study cohort (N=378). They also assessed the success rate of matching gene expression records to SNP records in a simulated cohort of 300 million individuals and correctly matched 97.1% of the records, demonstrating the feasibility of cross-linking these data types, which since then has been confirmed in additional studies [60,62,63]. Less literature has been published on other types of data, such as protein or epigenetic data (eg, DNA methylation), but similar proof of concept of cross-linkage to SNP data has been demonstrated in prior studies [58-60,63,64,66,67,71]. In the third mechanism, sensitive information such as disease phenotypes, demographic information, and behavioral traits is inferred from gene expression, protein levels, or other modalities (eg, age [72], cigarette smoking, and alcohol consumption [59] from DNA methylation).

However, such inference and linkage are not error free. For example, in the study by Schadt et al [65], the accuracy of the imputed SNPs from gene expression data was low (average Pearson correlation coefficient was 0.35 between true and inferred genotype). It is not clear whether such imputed data could be used for privacy attacks in the real world, such as in an identity tracing attack (eg, via upload of the imputed genetic data to GEDmatch or other). Considering that previous successful identity tracing attacks have used >500,000 SNPs [6], the inference of 1000 SNPs (with errors) may not be sufficient for such an attack. If the reconstruction of a larger set of SNPs were attempted, it is likely that the initial imputation error would propagate and thereby reduce the probability of a successful identity tracing attack. Furthermore, Schadt et al [65] reported much lower matching performance if training and test data stem from different array manufacturers, a scenario that is likely to occur in real-world data. Finally, although biological associations between genomic variants and gene expression are publicly accessible, substantial expert knowledge is still required for accessing this information and implementing the attack. Similar limitations apply to all the aforementioned studies. Altogether, data sets of RNA, protein, or epigenetic data, especially if they are large (eg, genome-wide), do allow for linkage and inference attacks. However, true reidentification would require matching the inferred genetic or phenotypic information to databases with identifying or quasi-identifying information in a next step, and no such full identity tracing attack starting with data other than DNA sequence has been demonstrated yet.

The guiding questions in this context are as follows:

- Do the data contain DNA sequence information directly (eg, DNA sequencing reads)? If yes, could the data be processed such that sequence information is no longer available (eg, report DNA methylation levels in percentage instead of providing raw sequencing read files)?
- Could DNA sequence information be inferred from the data (eg, via biological correlations such as expression or methylation quantitative trait loci)?
- What sensitive information could be inferred from the data (eg, age, sex, diseases, or physical traits)?

Experimental Assay

Knowing the experimental assay that was used to generate the data can already provide a first estimate of its information content and linkability. For example, sequencing-based assays generally produce very rich data (eg, high genome coverage and high precision, such as whole genome DNA sequencing), whereas polymerase chain reaction–based genotyping assays provide more sparse data (eg, information on only 1 nucleotide of the DNA sequence). However, genome coverage alone (ie, the percentage of all base pairs or loci of the genome covered by the method) is not a reliable proxy for privacy risk. In some circumstances, a data set with only 10 sequenced positions of the DNA could in fact be more critical than a data set containing hundreds of positions, if those 10 positions are in highly identifiable loci. However, as a very rough indicator of information content, we believe it is still valuable to consider the genome coverage of the data as one of many factors in the risk assessment. In many cases, the rule of thumb that more sequence information equals higher information content and hence risk of cross-linking, inference, and reidentification is true. Nevertheless, these aspects need to be carefully evaluated together with the biological modality of the data, the level of processing, and the specific content of the data.

It is also important to consider that data produced with frequently used methods, such as commercially available kits (eg, SNP microarrays), often target the same genetic variants that are also interrogated by DTC-GT companies and genome-wide disease association studies and can thus more easily be linked to public data and exploited for privacy attacks than data generated with tailor-made, targeted analysis methods (refer to the study by Lu et al [73] for an overview of genotyping arrays commonly used by direct-to-consumer companies). Finally, as nearby variants are more likely to be correlated, it is also important to consider how the genetic information in the data is spread across the genome. A targeted assay that reads all SNPs within a specific gene likely carries less information than an assay that interrogates the same number of SNPs distributed across the full genome, as nearby SNPs are more likely to be correlated [74]. In line with these arguments, the great majority of published privacy attacks were performed on data obtained from whole genome sequencing and commercially available SNP microarrays (ie, rich, genome-wide data in the order of hundreds of thousands of SNP loci from a commercial assay).

The guiding questions in this context are as follows:

- Which method was used to generate the data? Does this method produce rich or sparse data? (What percentage of all base pairs or loci of the genome are covered by the method?)
- How do the data produced with this method cover the genome (ie, genome-wide vs targeted approach)?
- How likely is it that data generated with the same method are present in publicly available databases (ie, commercial assay vs custom)?

Data Format or Level of Processing

The format of the data gives some indication on its processing level and can thus help to estimate its information content. Genetic data processing consists of cleaning, filtering, normalizing, and reducing raw data to a version that contains only the information that is relevant for its intended use. Important standard formats in genomic sequencing experiments sorted from raw to processed are *.fasta* and *.fastq* (raw nucleobase reads); *.bed*, *.bam*, and *.sam* (reads aligned to reference genome); *.vcf* and *.maf* files (deviations from the reference genome only), whereas highly processed data are often represented in tabular (*.csv* and *.tsv*) or otherwise structured form (*.json*, *.xml*, or other) containing only variants or regions of interest. Raw or low-processed data (*.fasta*, *.fastq*, *.bed*, *.bam*, or *.sam*) often contain information that is not of primary interest to research but can be exploited for reidentification attacks (eg, raw read files from gene expression studies can contain genomic variant information [63]). While the possibilities for privacy attacks are greater in raw data, it is important to note that the required effort and expert knowledge for handling these data are usually higher than those for processed data, where genetic variants such as SNPs do not need to be extracted.

The guiding question in this context is as follows:

- If the data are in a raw or semiprocessed format, do the data contain any information that is not directly relevant for their intended use?

Germline Versus Somatic Variation Content

Genetic variants found in an individual's genome can be categorized into germline and somatic variants. This categorization is specific to individuals and depends on the heritability of the variant (ergo, its presence in the individual's reproductive tissues). Heritable variants are categorized as germline (ie, present in germ and usually also in somatic cells) and nonheritable variants are categorized as somatic (ie, present in somatic cells only). In the context of genetic privacy, it is important to understand that germline variation comprises all variants that can be assumed to be present in every cell of the body, are not expected to change much throughout the lifetime of an individual, are inherited from parental DNA, and are expected to be passed to the offspring. Such variation can inform about identity, ancestry, and kinship and is, therefore, used by DTC-GT providers, forensics, and genetic genealogy services. The most prominent example for germline variation are SNPs, as variation found at known SNP loci is generally assumed to be germline. (However, the terms germline variants and SNPs cannot be used interchangeably, as they refer to different concepts: germline describes the heritability, and SNP describes the type of variant and its frequency in the population.) Overall, germline variants are not only highly relevant for individual identification because of their stability and omnipresence across tissues but are also of great interest for scientific research. Associations of germline variants to disease, physical traits, or other biomedical modalities are well studied, with results being publicly accessible. As such, germline variants are vulnerable to identity, inference, and linkage attacks, and indeed, all the reviewed privacy attacks targeted germline variants.

In contrast, somatic variants are acquired during life (after fertilization) and are usually present only in specific, nonreproductive tissues or even only in single cells or cell populations. They are intensively studied in the context of diseases (eg, cancer), and as they are often found to be associated with diseases, data on somatic variants could be used to infer sensitive attributes about data subjects. However, their low association with identity and use limited to clinical diagnostics and scientific research makes it very difficult to cross-link them to databases with identifying or quasi-identifying information. DTC-GT companies, forensics services, or genetic genealogy services do not use somatic variants to determine identity, familial relations, or ancestry, as somatic variation is neither stable nor present in all tissues and cells (usually found only in a fraction of cells analyzed in a sample). A linkage attack based on somatic variation would require a matching data record of the same tissue, ideally taken at a similar time in life, which is unlikely to exist for most cases (as somatic variant patterns can change rapidly, eg, in cancer tissue). No identity tracing, inference, or membership attack based on somatic variation data has been published yet, and considering its low potential for identifiability, somatic variation data can currently be considered a low risk for reidentification attacks.

To determine whether a variant is germline or somatic, one would ideally analyze multiple samples from one individual to determine whether the variant is present in germ cells or only in specific somatic cells. In practice, experts can assess the status of a variant from its sequencing read signal (determining whether it is present in all cells of the sample or only in a few), genomic location, and type alone by comparing it to public knowledge of known loci of germline and somatic variation or through computational approaches [75]. In processed genetic data, variants which are with high certainty germline have often already been identified and are indicated as such (eg, SNPs are identified by a specific reference SNP cluster ID, such as "rs343543"), whereas somatic variants are described by standard mutation nomenclature (eg, single nucleotide variants [SNVs] are described by the Human Genome Variation Society nomenclature, containing the reference genome used; the genomic location of the variant; the nucleotide in the reference sequence; and the detected nucleotide, such as "NC_000023.9:g.32317682G>A"). Furthermore, the type of tissue that was used to generate genetic data, most importantly whether samples were taken from healthy or tumor tissue, can also give some indication on the amount of germline variation included in the data. When analyzing tumor tissue data, germline variations such as SNPs are typically removed during processing, as the focus is on studying somatic variation. However, especially if the data are raw and unfiltered, they often contain germline variants irrespective of whether they were taken from healthy or tumor tissue and must hence be considered a higher risk for reidentification. Therefore, while data that are both derived from tumor tissues and highly processed are often a low privacy risk, the amount of information on germline variation that is contained in the data needs to be assessed case by case. Public databases (eg, dbSNP, hosted by the National Institutes of Health's National Center for Biotechnology Information) store information about the genomic locations and population frequencies of SNPs and can

be used to search data for this important type of germline variation.

The guiding questions in this context are as follows:

- Was germline or somatic variation of primary interest when generating or processing the data?
- If somatic variation was of primary interest, was germline variation removed from the data?

Part 2. High-Risk Components

SNPs

SNPs are germline SNVs that are present in >1% of the population. They are highly relevant features for individual reidentification and the most privacy-critical component of genetic data sets. Because SNPs usually have 2 different states (ie, a common or reference and a rare nucleotide) and human somatic cells have 2 DNA copies (ie, are diploid), an individual usually has 1 of 3 different states at a SNP locus, often represented as 0,1, and 2 (0 represents 2 copies of the common variant [ie, homozygous for major allele], 1 represents 1 copy of the common variant and 1 copy of the rare variant [heterozygous], and 2 represents 2 copies of the rare variant [homozygous for minor allele]). Knowing an individual's state at 30 to 80 statistically independent SNPs (or a random set of approximately 300 SNPs) can suffice for individual identification [76-79], yet commonly used SNP or genome sequencing assays often read hundreds of thousands of SNPs at once. As germline variation, SNPs are assumed to be stable and present in every cell of the body, signifying that they can identify individuals across samples taken at different times or from different tissues. As they are heritable, DTC-GT providers and forensic institutes compare SNP patterns of individuals to determine familial relations and ancestry [80]. Furthermore, SNPs are associated with physiological traits (eg, skin, hair and eye color [2,3], facial features [81], BMI [82], and height [4]), ethnicity [1], and susceptibility to diseases [5], making them central to research and genetic testing (refer to the study by Dabas et al [83] for a review of association of SNPs with externally visible characteristics).

SNP data can be directly used for reidentification by matching it to publicly accessible databases, as demonstrated in the reidentification attack by Erlich et al [6], who uploaded SNP data (700,000 SNPs) from an anonymous study participant to the genetic genealogy website GEDmatch and identified the participant's surname through matches with relatives. Such identity tracing attacks are possible because millions of people send their DNA to DTC-GT companies such as AncestryDNA, 23andMe, FamilyTreeDNA, or MyHeritage [84], and many also decide to share their genetic data on publicly accessible websites, such as GEDmatch, the Personal Genome Project [85], or OpenSNP [86]. Enabling individuals to identify and contact relatives, learn about their ancestry, disease predispositions, and contribute their data to research, these platforms often contain genetic data accompanied by information about an individual's diseases and traits or even personal data such as place of residence, age, sex, surname, or phone number. In addition, there is a wealth of publicly accessible knowledge on associations of SNPs with physical features, diseases, other genetic variants or genetic modalities (eg, gene expression and DNA methylation; eg, dbSNP database [87], the GWAS catalog [5], the International Genome Sample Resource from the 1000 Genomes Project [88], and data from the HapMap project [89]), which can and have been exploited for completion and inference attacks (eg, inference of additional genetic variation in genomic regions that were not studied originally, other biomedical modalities such as gene expression and DNA methylation, or physical attributes [9,90-96]). For example, Humbert et al [92] predicted phenotypic traits (eye, hair and skin color, blood type, and more) of individuals from their SNP data (20 SNPs) using publicly available knowledge on SNP-phenotype associations from the public database SNPedia and used this information to cross-link individuals between genetic and phenotypic data sets. In addition, Humbert et al [92] inferred additional and sensitive information (eg, susceptibility to Alzheimer disease) from the SNP data. However, this linkage attack had a success rate of only 5% (ie, proportion of correctly matched individuals) in a data set of 80 individuals and is likely to perform worse in more realistic scenarios with larger data sets. Nyholt et al [9] imputed the status of multiple risk variants for Alzheimer disease in the published genome of Dr James Watson [94] from SNPs in nearby genomic regions, although the respective gene had been masked. Edge et al [90] cross-linked individuals in SNP and short tandem repeat (STR) data sets, a highly identifiable type of genetic variation that is used in forensics, by imputing STR from SNP data (642,563 loci). In a highly debated study, Lippert et al [93] developed a model to predict phenotypic traits (facial structure, voice, eye color, skin color, age, sex, height, and BMI) from whole genome sequencing (WGS) data containing >6 million SNPs and used it to cross-link high-resolution face photographs of individuals to their genetic data in a cohort of 1061 study participants. In a real-life scenario, photos and personal data from social media could be exploited for such an attack and matched to the inferred phenotype. However, it has been argued that the predictive power in this study stems mainly from the estimation of the participant's ancestry and sex [97] and that the attack is unlikely to be successful in the real world and with more realistic, lower-quality images [98]. Furthermore, large, genome-wide association studies indicate that the currently known associations between SNPs and facial structure, voice, height, and BMI are too small to be useful for accurate phenotype prediction on an individual level; however, this will likely improve in the future. Nevertheless, other characteristics, such as ancestry, eye, hair color, and skin color, can be inferred from specific SNPs with high accuracy, and corresponding DNA phenotyping kits are already commercially available and used in forensics today [99]. As a small number of SNPs can already uniquely identify an individual and SNPs are widely available in public databases together with identifying and quasi-identifying information, SNPs must be considered a high risk for privacy and data sanitization efforts (eg, as proposed by Emani et al [100]) should be used in any genetic data set containing >20 SNPs.

The guiding questions in this context are as follows:

- How many SNPs do the data contain (directly or indirectly)?
- Are the SNPs in close proximity or spread across the genome (nearby SNPs are more likely to be correlated and

thus often contain less information than statistically independent SNPs)?
- Are the interrogated SNPs frequently assessed in research or by DTC-GT providers (ie, how likely is it that they can be linked to publicly available, identifying data sets)? The study by Lu et al [73] presents an overview of genotyping arrays commonly used by direct-to-consumer companies.
- Are all SNPs relevant to the intended use of the data or could some be removed from the data?
- What sensitive information could be inferred from the data (eg, diseases and physical traits)?
- Could additional DNA sequence information be inferred from the data (eg, association with STRs or other)?

STRs

The human genome contains more than half a million regions of repetitive units of 2 to 6 bases, the so-called STRs or microsatellites [101]. The number of repeats in these regions is highly variable across individuals and can affect protein function or expression or be linked to medical conditions or physical traits [102]. Knowing the repeat numbers of as little as 10 to 30 STRs can suffice for individual identification. Because of their high identifiability, STRs are used to determine identity and kinship in forensics, law enforcement, paternity testing, and genetic genealogy. For example, the Combined DNA Index System (CODIS; a set of 20 STRs) is used to connect suspects to crime scenes or establish identity of missing persons. While CODIS STRs are usually not of interest in research studies or genetic genealogy, STRs on the Y chromosome (ie, Y-STRs, only present in male individuals) are included in several DTC-GT kits, where they are used to identify relatives along the paternal ancestry line (eg, "Y-STR Testing" by FamilyTreeDNA). Consequently, several large databases of STR loci with accompanying identifying and quasi-identifying information exist (eg, mitoYDNA from mitoYDNA Ltd). In addition, the CODIS forensic database and analysis software contains genetic data and identifying information from >14 million individuals in the United States alone [103].

Several studies demonstrate reidentification attacks on Y-STRs. Gitschier et al [104] provided first evidence for surname inference from Y-STRs by matching genetic STR profiles of anonymous study participants from the international HapMap project [89] to 2 genetic genealogy databases (Ysearch and Sorenson Molecular Genealogy Foundation [SGMF]). Later, Gymrek et al [105] demonstrated that it is not only possible to infer surnames from STR data (eg, 34 Y-STR loci extracted from WGS data) but also to triangulate the actual identity of data subjects with high probability using publicly accessible genealogy databases, record search engines, obituaries, and genealogical websites. The authors attempted this for 10 study participants of the 1000 Genomes Project and correctly identified 5 out of 10 individuals. It is important to note that STR data can also be fortuitously included in genetic data derived from targeted gene or WGS, even if they are not of primary interest for the study. Moreover, STR markers can be imputed from genetic data sets that do not even cover STR regions by exploiting known associations between SNPs and STRs [90]. While the authors of this study report a low imputation accuracy for STRs from SNPs (likely too low to reliably impute full STR profiles even from large SNP data), they did demonstrate the ability to cross-link records across SNP and STR databases. In detail, they correctly matched 90% to 98% of paired SNP (642,563 loci) and STR data records (13 STRs) to each other, and such successful linkage has also been demonstrated elsewhere [106].

Due to the high association of STRs with identity, any genetic data that directly (eg, repeat numbers for specific STR regions) or indirectly (eg, WGS data covering STR regions) contain >10 STR regions could be considered identifiable. However, the actual risk of reidentification depends on the availability of STR databases with identifying and quasi-identifying information and the ability to cross-link records. It is important to note that the databases used in the seminal study by Gymrek et al [105] (ie, Ysearch and SGMF) are no longer available (Ysearch, belonging to FamilyTreeDNA, closed in 2018, and SGMF, belonging to Ancestry, was shut down in 2015), and access to the CODIS database is restricted to criminal justice agencies for law enforcement identification purposes. However, databases from DTC-GT providers (eg, FamilyTreeDNA) and public platforms (eg, mitoYDNA) are still available and allow uploading results from third-party providers; therefore, an attacker could fabricate a genetic testing result from STR data [107,108] and reproduce the demonstrated surname inference attacks. From information about possible surnames, sex, and residence inferred from matches on the platform, the triangulation of identity could be possible with the help of additional publicly available resources [105,109]. However, such an attack would only be possible on male data records (ie, Y chromosome based) and is not guaranteed to find matches that allow surname inference; the success rate in the demonstrated attack was 11.9% (109/911 cases), and the 2 previous studies used >30 STR loci (all located in close vicinity of each other and on the Y chromosome). Furthermore, the know-how and effort necessary for such an attack is high. Finally, even if genetic matches or surnames are identified, the reconstruction of identity from surname is not trivial and can take months to complete, as others have pointed out [110]. Still, because of their high identifiability potential and their use in DTC-GT, paternity testing, and forensics, STRs should be removed from genetic data if they are not of primary interest and otherwise considered a high risk for privacy.

The guiding questions in this context are as follows:

- Do the data directly or indirectly (eg, STRs in raw data and STRs imputable from SNPs) contain >10 STR loci?
- Are these STR loci either (1) part of the CODIS system or (2) on the Y chromosome (ie, high linkability)?
- Could additional DNA sequence information be inferred from the data (eg, association with SNPs or other)?

Aggregated Sample Measures

Aggregated sample measures, that is, variables that are the result of aggregating genetic data across multiple samples can also be exploited for privacy attacks (reviewed by Craig et al [111]). The most prominent examples are summary statistics from association studies such as SNP frequencies, odds ratios, or correlation coefficients. However, the limited information content in these summary statistics usually only allows for

membership attacks, that is, assessing whether an individual of known genetic background is part of a study group or database or not [112-114]. Multiple studies demonstrate such an attack [113,115-119], although Homer et al [114] were the first to explain how membership of an individual in a mixture can be predicted from the reported SNP allele frequencies (ie, if SNPs of that individual are known, in this case >10,000 SNPs). The authors accomplished this by comparing the reported study allele frequencies to allele frequencies in a reference cohort of similar ancestry (obtained from public resources) and detecting the bias introduced by the sample of interest. Their method performed well even if the individual's contribution to the mixture was <1%, and this method can easily be extended to predicting membership from aggregated data from a study cohort. In response to that, the US National Institutes of Health has restricted the publication of aggregate GWAS results in their databases [120]; however, the feasibility of the attack has been critically discussed. Its power depends on the size and quality of the actual and reference cohorts, the number of reported SNP allele frequencies, prior knowledge of the attacker, and the fulfillment of several underlying assumptions, many of which are likely not fulfilled in practice [115,116,121,122]. Aside from membership attacks, it was also shown that aggregate results, such as linear models that have been fitted to study data or polygenic risk scores, can be exploited to predict sensitive attributes and genotypes via model inversion [28,123]. However, this attack required background information on the data subject and on the distribution of variables in the study data. Furthermore, its performance is limited by the predictive power and complexity of the fitted model. Membership and attribute inference attacks on aggregate data can reveal demographic, genetic, and phenotypic information (such as country or place of residence due to participation in a local study, ethnicity, disease, age group, or presence of specific genetic variants due to descriptions of inclusion or exclusion criteria in the cohort) and can thus facilitate linkage and identity tracing attacks, which is why they can be a risk for privacy. However, no identity tracing attack based on aggregate data has been demonstrated yet.

The guiding question in this context is as follows:

- What sensitive information could an attacker gain from ascertaining the membership of an individual to the data set (eg, geographic information, sex, disease, and age)?

Part 3. Low-Risk Components

No privacy attack has been demonstrated on these components, but due to their high association with identifying and sensitive attributes, we recommend including them in the risk assessment.

Rare SNVs

Rare SNVs are single nucleotide substitutions that are present in <1% of the population. They may be somatic or germline and can be associated with pathological conditions and thus reveal sensitive information. Furthermore, while less informative than common SNVs (ie, SNPs) from an information theoretical standpoint, rare variants greatly increase the risk of reidentification for the small subpopulation of variant carriers. However, because of their low frequency in the population, germline SNVs are rarely the target of large scientific studies (eg, for phenotype or disease association) and have very limited use for ancestry and disease susceptibility analysis. Therefore, most DTC-GT providers and research studies specifically target regions of common genetic variation (eg, SNPs) and either use assays that do not detect SNVs or remove them during preprocessing, making it very unlikely that a set of SNVs could be linked to any database with quasi-identifying information. No identity tracing, completion, or inference attack has been published on SNVs yet; therefore, they can currently be viewed as a low risk for reidentification, despite their high theoretical potential for identifiability.

The guiding questions in this context are as follows:

- What sensitive information could be inferred from the data (eg, diseases and physical traits)?
- Could additional DNA sequence information be inferred from the data (eg, association with SNPs or other)?
- Are there any databases that could be used to cross-link the data to identifiable data, and how accessible are the databases?

Structural Variants

The study of structural variants (SVs) in the human genome is in its early stages, but it is already clear that it accounts for even more individual variation than SNPs [124,125]. The best-studied type of SVs is copy number variation (CNV), that is, deletions and duplications of regions larger than 50 base pairs. CNVs can be used as measures of relatedness and identifiers of population origin [126], have a strong impact on gene expression [127], and could allow for the inference of physical features [128] and pathological conditions [129], thereby revealing sensitive information of data subjects. However, CNVs are still not well studied, and sequencing technologies have only recently progressed to a level that allows to capture their full scope in the human genome (reviewed by Mahmoud et al [124]). Most importantly, human CNV databases are very scarce in comparison to databases of SNVs (refer to the study by Ho et al [130] for an overview of the available human SV reference sets), and they are currently not used for genetic genealogy analyses, making it difficult to link CNVs across databases to obtain identifying information. A privacy attack based on CNVs or any other type of SV yet remains to be demonstrated. Finally, it is important to note that many SVs that are assessed in medical and research studies are somatic, that is, nonhereditary, not present in all cells of the body, not stable, and thus not strongly associated with identity. For example, tumor tissue is characterized by frequent and dynamic changes in SVs (eg, CNVs in tumor tissue, also referred to as CNAs), which are likely neither directly nor indirectly identifiable. Therefore, the risk of reidentification from SVs can currently be considered low, but the growth of public databases and their use in genealogical or clinical research should be monitored. The same holds true for common SVs, such as CNVs that occur in >1% of the population and are hence classified as polymorphisms (ie, CNPs). Little is known about the population frequencies of CNVs, and while public databases are growing, no privacy attack based on CNPs has been demonstrated yet. Due to the limited knowledge about CNPs or other common SVs in the

population, their presence in genetic data is difficult to assess, and they can be considered a low risk for reidentification at the current time.

The guiding questions in this context are as follows:

- What sensitive information could be inferred from the data (eg, diseases and physical traits)?
- Could additional DNA sequence information be inferred from the data (eg, association with SNPs or other)?
- Are there any databases that could be used to cross-link the data to identifiable data, and how accessible are the databases?

Discussion

Limitations

It is important to acknowledge some key limitations of our review. First, it is possible that we may have missed relevant studies. This is particularly true for recent research, as our search was confined to original studies referenced in existing reviews. While the search strategy was designed to retrieve the most pertinent studies, it carries the risk of overlooking lesser-known or very recent studies. Therefore, we recommend conducting periodic reviews to stay updated with scientific advancements and changes in the availability of public genetic data that may contain (indirectly) identifying information susceptible to identity tracing attacks. Second, even under the assumption that all relevant literature was considered, it is still possible that we may have overlooked certain vulnerabilities. This is known as the "proof of nonexistence fallacy"—the absence of evidence for risk does not imply the absence of those risks. Finally, it was necessary to balance our aim of providing a comprehensive and evidence-based overview of genetic privacy vulnerabilities with our aim of providing practical and useful guidance. Therefore, we provide both a detailed assessment (refer to the *Results* section and Table S1 in Multimedia Appendix 1) as well as a simplified overview (Figure 2). However, this trade-off necessitated compromises in practical utility on one hand and scientific exhaustiveness on the other hand.

Conclusions

On the basis of the findings of this review, it can be argued that the privacy risks of genetic data vary greatly between data sets. Considering all genetic data at all times as information relating to an identifiable natural person is not correct, and it is becoming apparent that reidentification risk in genetic data must be assessed on a case-by-case basis and under the consideration of all the means reasonably likely to be used [131]. However, while efforts are underway [132], no practical guidelines or recommendations for performing such a reidentification risk assessment on genetic data have been proposed yet. On the basis of a review of the scientific literature on privacy attacks on genetic data, we provide an overview of genetic data privacy risks that can guide data processors in risk assessment by providing the necessary background knowledge and an overview of the existing evidence. We believe that a careful examination of the 9 described features in the data set at hand (biological modality or type of data, experimental assay, data format or level of processing, germline vs somatic variation content, content of SNPs, STRs, aggregated sample measures, rare SNVs, and SVs) provides a strong foundation for a data risk assessment. While completely eliminating the possibility of reidentification is rarely achievable, a more practical approach of risk minimization is warranted [133,134], accompanied by organizational and technical measures to safeguard genetic data from reidentification attack attempts and a transparent communication of the remaining risks to data subjects.

Acknowledgments

The authors would like to thank Florian Schneider for reviewing the manuscript.

Conflicts of Interest

NM was contracted by Roche Diagnostics GmbH, Penzberg, Germany, while contributing to this work. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
List of identified reviews and a table with the description and evaluation of original privacy attack studies.
[DOCX File , 36 KB - bioinform_v5i1e54332_app1.docx]

Multimedia Appendix 2
PRISMA checklist.
[PDF File (Adobe PDF File), 85 KB - bioinform_v5i1e54332_app2.pdf]

References

1. Huang T, Shu Y, Cai YD. Genetic differences among ethnic groups. BMC Genomics 2015 Dec 21;16:1093 [FREE Full text] [doi: 10.1186/s12864-015-2328-0] [Medline: 26690364]
2. Simcoe M, Valdes A, Liu F, Furlotte NA, Evans DM, Hemani G, et al. Genome-wide association study in almost 195,000 individuals identifies 50 previously unidentified genetic loci for eye color. Sci Adv 2021 Mar;7(11):eabd1239 [FREE Full text] [doi: 10.1126/sciadv.abd1239] [Medline: 33692100]

3. Pavan WJ, Sturm RA. The genetics of human skin and hair pigmentation. Annu Rev Genomics Hum Genet 2019 Aug 31;20:41-72. [doi: 10.1146/annurev-genom-083118-015230] [Medline: 31100995]
4. Yengo L, Vedantam S, Marouli E, Sidorenko J, Bartell E, Sakaue S, et al. A saturated map of common genetic variants associated with human height. Nature 2022 Oct;610(7933):704-712. [doi: 10.1038/s41586-022-05275-y] [Medline: 36224396]
5. Buniello A, MacArthur JA, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res 2019 Jan 08;47(D1):D1005-D1012 [FREE Full text] [doi: 10.1093/nar/gky1120] [Medline: 30445434]
6. Erlich Y, Shor T, Pe'er I, Carmi S. Identity inference of genomic data using long-range familial searches. Science 2018 Nov 09;362(6415):690-694 [FREE Full text] [doi: 10.1126/science.aau4832] [Medline: 30309907]
7. Greytak EM, Moore C, Armentrout SL. Genetic genealogy for cold case and active investigations. Forensic Sci Int 2019 Jun;299:103-113. [doi: 10.1016/j.forsciint.2019.03.039] [Medline: 30991209]
8. Kennett D. Using genetic genealogy databases in missing persons cases and to develop suspect leads in violent crimes. Forensic Sci Int 2019 Aug;301:107-117. [doi: 10.1016/j.forsciint.2019.05.016] [Medline: 31153988]
9. Nyholt DR, Yu CE, Visscher PM. On Jim Watson's APOE status: genetic information is hard to hide. Eur J Hum Genet 2009 Feb;17(2):147-149 [FREE Full text] [doi: 10.1038/ejhg.2008.198] [Medline: 18941475]
10. Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. Nat Rev Genet 2014 Jun;15(6):409-421 [FREE Full text] [doi: 10.1038/nrg3723] [Medline: 24805122]
11. El Emam K, Jonker E, Arbuckle L, Malin B. A systematic review of re-identification attacks on health data. PLoS One 2011;6(12):e28071 [FREE Full text] [doi: 10.1371/journal.pone.0028071] [Medline: 22164229]
12. Mohammed Yakubu A, Chen YP. Ensuring privacy and security of genomic data and functionalities. Brief Bioinform 2020 Mar 23;21(2):511-526. [doi: 10.1093/bib/bbz013] [Medline: 30759195]
13. Joly Y, Dalpe G. Genetic discrimination still casts a large shadow in 2022. Eur J Hum Genet 2022 Dec;30(12):1320-1322 [FREE Full text] [doi: 10.1038/s41431-022-01194-8] [Medline: 36163420]
14. Tiller J, Lacaze P. Australians can be denied life insurance based on genetic test results, and there is little protection. The Conversation. 2017 Aug 24. URL: https://theconversation.com/australians-can-be-denied-life-insurance-based-on-genetic-test-results-and-there-is-little-protection-81335 [accessed 2024-04-05]
15. Humbert M, Ayday E, Hubaux JP, Telenti A. Addressing the concerns of the lacks family: quantification of kin genomic privacy. In: Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security. 2013 Presented at: CCS '13; November 4-8, 2013; Berlin, Germany p. 1141-1152 URL: https://dl.acm.org/doi/10.1145/2508859.2516707 [doi: 10.1145/2508859.2516707]
16. Deznabi I, Mobayen M, Jafari N, Tastan O, Ayday E. An inference attack on genomic data using kinship, complex correlations, and phenotype information. IEEE/ACM Trans Comput Biol Bioinform 2018;15(4):1333-1343. [doi: 10.1109/TCBB.2017.2709740] [Medline: 30102600]
17. Trost B, Loureiro LO, Scherer SW. Discovery of genomic variation across a generation. Hum Mol Genet 2021 Oct 01;30(R2):R174-R186 [FREE Full text] [doi: 10.1093/hmg/ddab209] [Medline: 34296264]
18. EDPB documenton response to the request from the European commission for clarifications on the consistent application of the GDPR, focusing on health research. European Data Protection Board. 2021. URL: https://edpb.europa.eu/sites/default/files/files/file1/edpb_replyec_questionnaireresearch_final.pdf [accessed 2024-04-05]
19. Global Alliance for Genomics and Health. GENOMICS. A federated ecosystem for sharing genomic, clinical data. Science 2016 Jun 10;352(6291):1278-1280. [doi: 10.1126/science.aaf6162] [Medline: 27284183]
20. Shringarpure SS, Bustamante CD. Privacy risks from genomic data-sharing beacons. Am J Hum Genet 2015 Nov 05;97(5):631-646 [FREE Full text] [doi: 10.1016/j.ajhg.2015.09.010] [Medline: 26522470]
21. von Thenen N, Ayday E, Cicek AE. Re-identification of individuals in genomic data-sharing beacons via allele inference. Bioinformatics 2019 Feb 01;35(3):365-371. [doi: 10.1093/bioinformatics/bty643] [Medline: 30052749]
22. Raisaro JL, Tramèr F, Ji Z, Bu D, Zhao Y, Carey K, et al. Addressing Beacon re-identification attacks: quantification and mitigation of privacy risks. J Am Med Inform Assoc 2017 Jul 01;24(4):799-805 [FREE Full text] [doi: 10.1093/jamia/ocw167] [Medline: 28339683]
23. Ayoz K, Ayday E, Cicek AE. Genome reconstruction attacks against genomic data-sharing beacons. Proc Priv Enhanc Technol 2021;2021(3):28-48 [FREE Full text] [doi: 10.2478/popets-2021-0036] [Medline: 34746296]
24. Fienberg SE, Slavkovic A, Uhler C. Privacy preserving GWAS data sharing. In: Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops. 2011 Presented at: ICDMW '11; December 11, 2011; Washington, DC p. 628-635 URL: https://dl.acm.org/doi/10.1109/ICDMW.2011.140 [doi: 10.1109/icdmw.2011.140]
25. Huang Z, Ayday E, Fellay J, Hubaux JP, Juels A. GenoGuard: protecting genomic data against brute-force attacks. In: Proceedings of the 2015 IEEE Symposium on Security and Privacy. 2015 Presented at: SP '15; May 17-21, 2015; San Jose, CA p. 447-462 URL: https://ieeexplore.ieee.org/document/7163041 [doi: 10.1109/sp.2015.34]
26. Wang Y, Wen J, Wu X, Shi X. Infringement of individual privacy via mining differentially private GWAS statistics. In: Proceedings of the 2nd International Conference on Big Data Computing and Communications. 2016 Presented at: BigCom '16; July 29-31, 2016; Shenyang, China p. 29-31 URL: https://link.springer.com/chapter/10.1007/978-3-319-42553-5_30 [doi: 10.1007/978-3-319-42553-5_30]

27. Cavallaro L, Kinder J, Domingo-Ferrer J, Oprisanu B, Dessimoz C, Cristofaro ED. How much does Genoguard really "guard"?: an empirical analysis of long-term security for genomic data. In: Proceedings of the 18th ACM Workshop on Privacy in the Electronic Society. 2019 Presented at: WPES '19; November 11, 2019; London, UK p. 93-105 URL: https://tinyurl.com/4w8sxk6f [doi: 10.1145/3338498.3358641]
28. Fredrikson M, Lantz E, Jha S, Lin S, Page D, Ristenpart T. Privacy in pharmacogenetics: an end-to-end case study of personalized warfarin dosing. Proc USENIX Secur Symp 2014 Aug;2014:17-32 [FREE Full text] [Medline: 27077138]
29. Oprisanu B, Ganev G, Cristofaro ED. On utility and privacy in synthetic genomic data. In: Proceedings of the 2022 Network and Distributed Systems Security. 2022 Presented at: NDSS '22; April 24-28, 2022; San Diego, CA p. 1-18 URL: https://www.ndss-symposium.org/wp-content/uploads/2022-92-paper.pdf [doi: 10.14722/ndss.2022.24092]
30. Mittos A, Malin B, Cristofaro ED. Systematizing genome privacy research: a privacy-enhancing technologies perspective. Proc Priv Enhancing Technol 2019(1):87-107 [FREE Full text] [doi: 10.2478/popets-2019-0006]
31. Martinez C, Jonker E. A practical path towards genetic privacy in the United States. Future of Privacy Forum. 2020. URL: https://fpf.org/wp-content/uploads/2020/04/APracticalPathTowardGeneticPrivacy_April2020.pdf [accessed 2022-10-31]
32. Bernier A, Liu H, Knoppers BM. Computational tools for genomic data de-identification: facilitating data protection law compliance. Nat Commun 2021 Nov 29;12(1):6949 [FREE Full text] [doi: 10.1038/s41467-021-27219-2] [Medline: 34845213]
33. The GDPR and genomic data - the impact of the GDPR and DPA 2018 on genomic healthcare and research. PHG Foundation. 2020. URL: https://tinyurl.com/dfk7e3xs [accessed 2024-04-05]
34. Wagner I. Evaluating the strength of genomic privacy metrics. ACM Trans Priv Secur 2017 Jan 09;20(1):1-34. [doi: 10.1145/3020003]
35. Abinaya B, Santhi S. A survey on genomic data by privacy-preserving techniques perspective. Comput Biol Chem 2021 Aug;93:107538. [doi: 10.1016/j.compbiolchem.2021.107538]
36. Azencott CA. Machine learning and genomics: precision medicine versus patient privacy. Philos Trans A Math Phys Eng Sci 2018 Sep 13;376(2128):20170350. [doi: 10.1098/rsta.2017.0350] [Medline: 30082298]
37. Ayday E, Humbert M. Inference attacks against kin genomic privacy. IEEE Secur Privacy 2017;15(5):29-37. [doi: 10.1109/msp.2017.3681052]
38. Aziz MM, Sadat MN, Alhadidi D, Wang S, Jiang X, Brown CL, et al. Privacy-preserving techniques of genomic data-a survey. Brief Bioinform 2019 May 21;20(3):887-895 [FREE Full text] [doi: 10.1093/bib/bbx139] [Medline: 29121240]
39. Berger B, Cho H. Emerging technologies towards enhancing privacy in genomic data sharing. Genome Biol 2019 Jul 02;20(1):128 [FREE Full text] [doi: 10.1186/s13059-019-1741-0] [Medline: 31262363]
40. Bonomi L, Huang Y, Ohno-Machado L. Privacy challenges and research opportunities for genomic data sharing. Nat Genet 2020 Jul 29;52(7):646-654 [FREE Full text] [doi: 10.1038/s41588-020-0651-0] [Medline: 32601475]
41. Carter AB. Considerations for genomic data privacy and security when working in the cloud. J Mol Diagn 2019 Jul;21(4):542-552 [FREE Full text] [doi: 10.1016/j.jmoldx.2018.07.009] [Medline: 30703562]
42. Clayton EW, Halverson CM, Sathe NA, Malin BA. A systematic literature review of individuals' perspectives on privacy and genetic information in the United States. PLoS One 2018 Oct 31;13(10):e0204417 [FREE Full text] [doi: 10.1371/journal.pone.0204417] [Medline: 30379944]
43. Gürsoy G. Genome privacy and trust. Annu Rev Biomed Data Sci 2022 Aug 10;5(1):163-181. [doi: 10.1146/annurev-biodatasci-122120-021311] [Medline: 35508070]
44. Knoppers BM, Beauvais MJ. Three decades of genetic privacy: a metaphoric journey. Hum Mol Genet 2021 Oct 01;30(R2):R156-R160 [FREE Full text] [doi: 10.1093/hmg/ddab164] [Medline: 34155499]
45. May T. Sociogenetic risks — ancestry DNA testing, third-party identity, and protection of privacy. N Engl J Med 2018 Aug 02;379(5):410-412. [doi: 10.1056/nejmp1805870]
46. Oestreich M, Chen D, Schultze JL, Fritz M, Becker M. Privacy considerations for sharing genomics data. EXCLI J 2021;20:1243-1260 [FREE Full text] [doi: 10.17179/excli2021-4002] [Medline: 34345236]
47. Schwab AP, Luu HS, Wang J, Park JY. Genomic privacy. Clin Chem 2018 Dec;64(12):1696-1703. [doi: 10.1373/clinchem.2018.289512] [Medline: 29991478]
48. Shen H, Ma J. Privacy challenges of genomic big data. Adv Exp Med Biol 2017;1028:139-148. [doi: 10.1007/978-981-10-6041-0_8] [Medline: 29058220]
49. Shi X, Wu X. An overview of human genetic privacy. Ann N Y Acad Sci 2017 Jan 14;1387(1):61-72 [FREE Full text] [doi: 10.1111/nyas.13211] [Medline: 27626905]
50. Stiles D, Appelbaum PS. Cases in precision medicine: concerns about privacy and discrimination after genomic sequencing. Ann Intern Med 2019 May 07;170(10):717. [doi: 10.7326/m18-2666]
51. Wan Z, Hazel JW, Clayton EW, Vorobeychik Y, Kantarcioglu M, Malin BA. Sociotechnical safeguards for genomic data privacy. Nat Rev Genet 2022 Jul 04;23(7):429-445 [FREE Full text] [doi: 10.1038/s41576-022-00455-y] [Medline: 35246669]
52. Wang S, Jiang X, Singh S, Marmor R, Bonomi L, Fox D, et al. Genome privacy: challenges, technical approaches to mitigate risk, and ethical considerations in the United States. Ann N Y Acad Sci 2017 Jan 28;1387(1):73-83 [FREE Full text] [doi: 10.1111/nyas.13259] [Medline: 27681358]

53. Belani S, Tiarks GC, Mookerjee N, Rajput V. "I agree to disagree": comparative ethical and legal analysis of big data and genomics for privacy, consent, and ownership. Cureus 2021 Oct;13(10):e18736 [FREE Full text] [doi: 10.7759/cureus.18736] [Medline: 34796049]
54. Du L, Wang M. Genetic privacy and data protection: a review of Chinese direct-to-consumer genetic test services. Front Genet 2020 Apr 28;11:416 [FREE Full text] [doi: 10.3389/fgene.2020.00416] [Medline: 32425986]
55. Dugan T, Zou X. Privacy-preserving evaluation techniques and their application in genetic tests. Smart Health 2017 Jun;1-2:2-17. [doi: 10.1016/j.smhl.2017.03.003]
56. Lu D, Zhang Y, Zhang L, Wang H, Weng W, Li L, et al. Methods of privacy-preserving genomic sequencing data alignments. Brief Bioinform 2021 Nov 05;22(6):bbab151. [doi: 10.1093/bib/bbab151] [Medline: 34021302]
57. Heeney C, Hawkins N, de Vries J, Boddington P, Kaye J. Assessing the privacy risks of data sharing in genomics. Public Health Genomics 2011;14(1):17-25 [FREE Full text] [doi: 10.1159/000294150] [Medline: 20339285]
58. Backes M, Berrang P, Bieg M, Eils R, Herrrnann C, Humbert M, et al. Identifying personal DNA methylation profiles by genotype inference. In: Proceedings of the 2017 IEEE Symposium on Security and Privacy. 2017 Presented at: SP '17; May 22-26, 2017; San Jose, CA p. 957-976 URL: https://ieeexplore.ieee.org/document/7958619 [doi: 10.1109/sp.2017.21]
59. Philibert RA, Terry N, Erwin C, Philibert WJ, Beach SR, Brody GH. Methylation array data can simultaneously identify individuals and convey protected health information: an unrecognized ethical concern. Clin Epigenetics 2014;6(1):28 [FREE Full text] [doi: 10.1186/1868-7083-6-28] [Medline: 25859287]
60. Gürsoy G, Lu N, Wagner S, Gerstein M. Recovering genotypes and phenotypes using allele-specific genes. Genome Biol 2021 Sep 07;22(1):263 [FREE Full text] [doi: 10.1186/s13059-021-02477-x] [Medline: 34493313]
61. Hagestedt I, Zhang Y, Humbert M, Berrang P, Tang H, Wang X, et al. MBeacon: privacy-preserving beacons for DNA methylation data. In: Proceedings of the 2019 Network and Distributed Systems Security Symposium. 2019 Presented at: NDSS '19; February 24-27, 2019; San Diego, CA p. 1-15 URL: https://www.ndss-symposium.org/wp-content/uploads/2019/02/ndss2019_03A-2_Hagestedt_paper.pdf [doi: 10.14722/ndss.2019.23064]
62. Harmanci A, Gerstein M. Quantification of private information leakage from phenotype-genotype data: linking attacks. Nat Methods 2016 Mar;13(3):251-256 [FREE Full text] [doi: 10.1038/nmeth.3746] [Medline: 26828419]
63. Harmanci A, Gerstein M. Analysis of sensitive information leakage in functional genomics signal profiles through genomic deletions. Nat Commun 2018 Jun 22;9(1):2453 [FREE Full text] [doi: 10.1038/s41467-018-04875-5] [Medline: 29934598]
64. Boonen K, Hens K, Menschaert G, Baggerman G, Valkenborg D, Ertaylan G. Beyond genes: re-identifiability of proteomic data and its implications for personalized medicine. Genes (Basel) 2019 Sep 05;10(9):682 [FREE Full text] [doi: 10.3390/genes10090682] [Medline: 31492022]
65. Schadt EE, Woo S, Hao K. Bayesian method to predict individual SNP genotypes from gene expression data. Nat Genet 2012 May;44(5):603-608. [doi: 10.1038/ng.2248] [Medline: 22484626]
66. Dyke SO, Cheung WA, Joly Y, Ammerpohl O, Lutsik P, Rothstein MA, et al. Epigenome data release: a participant-centered approach to privacy protection. Genome Biol 2015 Jul 17;16(1):142 [FREE Full text] [doi: 10.1186/s13059-015-0723-0] [Medline: 26185018]
67. Berrang P, Humbert M, Zhang Y, Lehmann I, Eils R, Backes M. Dissecting privacy risks in biomedical data. In: Proceedings of the 2018 IEEE European Symposium on Security and Privacy. 2018 Presented at: EuroS&P'18; April 24-26, 2018; London, UK p. 62-76 URL: https://ieeexplore.ieee.org/document/8406591/similar#similar [doi: 10.1109/eurosp.2018.00013]
68. Zhao Y, Wang K, Wang W, Yin T, Dong W, Xu C. A high-throughput SNP discovery strategy for RNA-seq data. BMC Genomics 2019 Feb 27;20(1):160 [FREE Full text] [doi: 10.1186/s12864-019-5533-4] [Medline: 30813897]
69. Gürsoy G, Li T, Liu S, Ni E, Brannon CM, Gerstein MB. Functional genomics data: privacy risk assessment and technological mitigation. Nat Rev Genet 2022 Apr;23(4):245-258. [doi: 10.1038/s41576-021-00428-7] [Medline: 34759381]
70. Gürsoy G, Emani P, Brannon CM, Jolanki OA, Harmanci A, Strattan JS, et al. Data sanitization to reduce private information leakage from functional genomics. Cell 2020 Nov 12;183(4):905-917 [FREE Full text] [doi: 10.1016/j.cell.2020.09.036] [Medline: 33186529]
71. Li S, Bandeira N, Wang X, Tang H. On the privacy risks of sharing clinical proteomics data. AMIA Jt Summits Transl Sci Proc 2016;2016:122-131 [FREE Full text] [Medline: 27595046]
72. Dupras C, Beck S, Rothstein MA, Berner A, Saulnier KM, Pinkesz M, et al. Potential (mis) use of epigenetic age estimators by private companies and public agencies: human rights law should provide ethical guidance. Environ Epigenet 2019;5(3):dvz018. [doi: 10.1093/eep/dvz018]
73. Lu C, Greshake Tzovaras B, Gough J. A survey of direct-to-consumer genotype data, and quality control tool (GenomePrep) for research. Comput Struct Biotechnol J 2021 Jun 27;19:3747-3754 [FREE Full text] [doi: 10.1016/j.csbj.2021.06.040] [Medline: 34285776]
74. Collins A, Lonjou C, Morton NE. Genetic epidemiology of single-nucleotide polymorphisms. Proc Natl Acad Sci U S A 1999 Dec 21;96(26):15173-15177 [FREE Full text] [doi: 10.1073/pnas.96.26.15173] [Medline: 10611357]
75. Sun JX, He Y, Sanford E, Montesion M, Frampton GM, Vignot S, et al. A computational approach to distinguish somatic vs. germline origin of genomic alterations from deep sequencing of cancer specimens without a matched normal. PLoS Comput Biol 2018 Feb;14(2):e1005965 [FREE Full text] [doi: 10.1371/journal.pcbi.1005965] [Medline: 29415044]

76. Yousefi S, Abbassi-Daloii T, Kraaijenbrink T, Vermaat M, Mei H, van 't Hof P, et al. A SNP panel for identification of DNA and RNA specimens. BMC Genomics 2018 Jan 25;19(1):90 [FREE Full text] [doi: 10.1186/s12864-018-4482-7] [Medline: 29370748]
77. Lin Z, Owen AB, Altman RB. Genetics. Genomic research and human subject privacy. Science 2004 Jul 09;305(5681):183. [doi: 10.1126/science.1095019] [Medline: 15247459]
78. Sanchez JJ, Phillips C, Børsting C, Balogh K, Bogus M, Fondevila M, et al. A multiplex assay with 52 single nucleotide polymorphisms for human identification. Electrophoresis 2006 May;27(9):1713-1724. [doi: 10.1002/elps.200500671] [Medline: 16586411]
79. Pakstis AJ, Speed WC, Fang R, Hyland FC, Furtado MR, Kidd JR, et al. SNPs for a universal individual identification panel. Hum Genet 2010 Mar;127(3):315-324. [doi: 10.1007/s00439-009-0771-1] [Medline: 19937056]
80. Kling D, Phillips C, Kennett D, Tillmar A. Investigative genetic genealogy: current methods, knowledge and practice. Forensic Sci Int Genet 2021 May;52:102474 [FREE Full text] [doi: 10.1016/j.fsigen.2021.102474] [Medline: 33592389]
81. White JD, Indencleef K, Naqvi S, Eller RJ, Hoskens H, Roosenboom J, et al. Insights into the genetic architecture of the human face. Nat Genet 2021 Jan;53(1):45-53 [FREE Full text] [doi: 10.1038/s41588-020-00741-7] [Medline: 33288918]
82. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature 2015 Feb 12;518(7538):197-206 [FREE Full text] [doi: 10.1038/nature14177] [Medline: 25673413]
83. Dabas P, Jain S, Khajuria H, Nayak BP. Forensic DNA phenotyping: inferring phenotypic traits from crime scene DNA. J Forensic Leg Med 2022 May;88:102351. [doi: 10.1016/j.jflm.2022.102351] [Medline: 35427851]
84. Regalado A. More than 26 million people have taken an at-home ancestry test. MIT Technology Review. 2019 Feb 11. URL: https://www.technologyreview.com/2019/02/11/103446/more-than-26-million-people-have-taken-an-at-home-ancestry-test/ [accessed 2024-04-05]
85. Church GM. The personal genome project. Mol Syst Biol 2005;1:2005.0030 [FREE Full text] [doi: 10.1038/msb4100040] [Medline: 16729065]
86. Greshake B, Bayer PE, Rausch H, Reda J. openSNP--a crowdsourced web resource for personal genomics. PLoS One 2014;9(3):e89204 [FREE Full text] [doi: 10.1371/journal.pone.0089204] [Medline: 24647222]
87. Smigielski EM, Sirotkin K, Ward M, Sherry ST. dbSNP: a database of single nucleotide polymorphisms. Nucleic Acids Res 2000 Jan 01;28(1):352-355 [FREE Full text] [doi: 10.1093/nar/28.1.352] [Medline: 10592272]
88. Fairley S, Lowy-Gallego E, Perry E, Flicek P. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. Nucleic Acids Res 2020 Jan 08;48(D1):D941-D947 [FREE Full text] [doi: 10.1093/nar/gkz836] [Medline: 31584097]
89. International HapMap Consortium. A haplotype map of the human genome. Nature 2005 Oct 27;437(7063):1299-1320 [FREE Full text] [doi: 10.1038/nature04226] [Medline: 16255080]
90. Edge MD, Algee-Hewitt BF, Pemberton TJ, Li JZ, Rosenberg NA. Linkage disequilibrium matches forensic genetic records to disjoint genomic marker sets. Proc Natl Acad Sci U S A 2017 May 30;114(22):5671-5676 [FREE Full text] [doi: 10.1073/pnas.1619944114] [Medline: 28507140]
91. He Z, Yu J, Li J, Han Q, Luo G, Li Y. Inference attacks and controls on genotypes and phenotypes for individual genomic data. IEEE/ACM Trans Comput Biol Bioinform 2020;17(3):930-937. [doi: 10.1109/TCBB.2018.2810180] [Medline: 29994587]
92. Humbert M, Huguenin K, Hugonot J, Ayday E, Hubaux JP. De-anonymizing genomic databases using phenotypic traits. Proc Priv Enhanc Technol 2015;2015:99-114 [FREE Full text] [doi: 10.1515/popets-2015-0020]
93. Lippert C, Sabatini R, Maher MC, Kang EY, Lee S, Arikan O, et al. Identification of individuals by trait prediction using whole-genome sequencing data. Proc Natl Acad Sci U S A 2017 Sep 19;114(38):10166-10171 [FREE Full text] [doi: 10.1073/pnas.1711125114] [Medline: 28874526]
94. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature 2008 Apr 17;452(7189):872-876. [doi: 10.1038/nature06884] [Medline: 18421352]
95. Sero D, Zaidi A, Li J, White JD, Zarzar TB, Marazita ML, et al. Facial recognition from DNA using face-to-DNA classifiers. Nat Commun 2019 Jun 11;10(1):2557 [FREE Full text] [doi: 10.1038/s41467-019-10617-y] [Medline: 31186421]
96. Wang Y, Wu X, Shi X. Using aggregate human genome data for individual identification. In: Proceedings of the 2013 IEEE International Conference on Bioinformatics and Biomedicine. 2013 Presented at: BIBM '13; December 18-21, 2013; Shanghai, China p. 410-415 URL: https://ieeexplore.ieee.org/abstract/document/6732527
97. Erlich Y. Major flaws in “identification of individuals by trait prediction using whole-genome sequencing data”. bioRxiv Preprint posted online September 7, 2017 [FREE Full text] [doi: 10.1101/185330]
98. Venkatesaramani R, Malin BA, Vorobeychik Y. Re-identification of individuals in genomic datasets using public face images. Sci Adv 2021 Nov 19;7(47):eabg3296 [FREE Full text] [doi: 10.1126/sciadv.abg3296] [Medline: 34788101]
99. Schneider PM, Prainsack B, Kayser M. The use of forensic DNA phenotyping in predicting appearance and biogeographic ancestry. Dtsch Arztebl Int 2019 Dec 23;51-52(51-52):873-880 [FREE Full text] [doi: 10.3238/arztebl.2019.0873] [Medline: 31941575]

100. Emani PS, Gürsoy G, Miranker A, Gerstein MB. Assessing and mitigating privacy risk of sparse, noisy genotypes by local alignment to haplotype databases. bioRxiv Preprint posted online August 30, 2022 [FREE Full text] [doi: 10.1101/2021.07.18.452853]
101. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature 2001 Feb 15;409(6822):860-921. [doi: 10.1038/35057062] [Medline: 11237011]
102. Wyner N, Barash M, McNevin D. Forensic autosomal short tandem repeats and their potential association with phenotype. Front Genet 2020;11:884 [FREE Full text] [doi: 10.3389/fgene.2020.00884] [Medline: 32849844]
103. FBI. US Law Enforcement Resources: Biometrics and Fingerprints. Combined DNA Index System (CODIS). URL: https://tinyurl.com/3by74dhj [accessed 2023-03-29]
104. Gitschier J. Inferential genotyping of Y chromosomes in latter-day saints founders and comparison to Utah samples in the HapMap project. Am J Hum Genet 2009 Feb;84(2):251-258 [FREE Full text] [doi: 10.1016/j.ajhg.2009.01.018] [Medline: 19215731]
105. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. Science 2013 Jan 18;339(6117):321-324. [doi: 10.1126/science.1229566] [Medline: 23329047]
106. Kim J, Edge MD, Algee-Hewitt BF, Li JZ, Rosenberg NA. Statistical detection of relatives typed with disjoint forensic and biomedical loci. Cell 2018 Oct 18;175(3):848-58.e6 [FREE Full text] [doi: 10.1016/j.cell.2018.09.008] [Medline: 30318150]
107. Edge MD, Coop G. Attacks on genetic privacy via uploads to genealogical databases. Elife 2020 Jan 07;9:e51810 [FREE Full text] [doi: 10.7554/eLife.51810] [Medline: 31908268]
108. Ney P, Ceze L, Kohno T. Genotype extraction and false relative attacks: security risks to third-party genetic genealogy services beyond identity inference. In: Proceedings of the 2020 Network and Distributed Systems Security (NDSS) Symposium. 2020 Presented at: NDSS '20; February 23-26, 2020; San Diego, CA p. 1-15 URL: https://www.ndss-symposium.org/wp-content/uploads/2020/02/23049.pdf/> [doi: 10.14722/ndss.2020.23049]
109. Sweeney L, Abu A, Winn J. Identifying participants in the personal genome project by name. SSRN Journal Preprint posted online April 29, 2013 [FREE Full text] [doi: 10.2139/ssrn.2257732]
110. Guerrini CJ, Wickenheiser RA, Bettinger B, McGuire AL, Fullerton SM. Four misconceptions about investigative genetic genealogy. J Law Biosci 2021;8(1):lsab001 [FREE Full text] [doi: 10.1093/jlb/lsab001] [Medline: 33880184]
111. Craig DW, Goor RM, Wang Z, Paschall J, Ostell J, Feolo M, et al. Assessing and managing risk when sharing aggregate genetic variant data. Nat Rev Genet 2011 Sep 16;12(10):730-736 [FREE Full text] [doi: 10.1038/nrg3067] [Medline: 21921928]
112. Im HK, Gamazon ER, Nicolae DL, Cox NJ. On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy. Am J Hum Genet 2012 Apr 06;90(4):591-598 [FREE Full text] [doi: 10.1016/j.ajhg.2012.02.008] [Medline: 22463877]
113. Wang R, Li YF, Wang X, Tang H, Zhou X. Learning your identity and disease from research papers: information leaks in genome wide association study. In: Proceedings of the 16th ACM conference on Computer and communications security. 2009 Presented at: CCS '09; November 9-13, 2009; Chicago, IL p. 534-544 URL: https://dl.acm.org/doi/10.1145/1653662.1653726 [doi: 10.1145/1653662.1653726]
114. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet 2008 Aug 29;4(8):e1000167 [FREE Full text] [doi: 10.1371/journal.pgen.1000167] [Medline: 18769715]
115. Sankararaman S, Obozinski G, Jordan MI, Halperin E. Genomic privacy and limits of individual detection in a pool. Nat Genet 2009 Sep;41(9):965-967. [doi: 10.1038/ng.436] [Medline: 19701190]
116. Braun R, Rowe W, Schaefer C, Zhang J, Buetow K. Needles in the haystack: identifying individuals present in pooled genomic data. PLoS Genet 2009 Oct;5(10):e1000668 [FREE Full text] [doi: 10.1371/journal.pgen.1000668] [Medline: 19798441]
117. Cai R, Hao Z, Winslett M, Xiao X, Yang Y, Zhang Z, et al. Deterministic identification of specific individuals from GWAS results. Bioinformatics 2015 Jun 01;31(11):1701-1707 [FREE Full text] [doi: 10.1093/bioinformatics/btv018] [Medline: 25630377]
118. Bu D, Wang X, Tang H. Haplotype-based membership inference from summary genomic data. Bioinformatics 2021 Jul 12;37(Suppl_1):i161-i168 [FREE Full text] [doi: 10.1093/bioinformatics/btab305] [Medline: 34252973]
119. Almadhoun N, Ayday E, Ulusoy Ö. Inference attacks against differentially private query results from genomic datasets including dependent tuples. Bioinformatics 2020 Jul 01;36(Suppl_1):i136-i145 [FREE Full text] [doi: 10.1093/bioinformatics/btaa475] [Medline: 32657411]
120. Zerhouni EA, Nabel EG. Protecting aggregate genomic data. Science 2008 Oct 03;322(5898):44. [doi: 10.1126/science.322.5898.44b] [Medline: 18772394]
121. Visscher PM, Hill WG. The limits of individual identification from sample allele frequencies: theory and statistical analysis. PLoS Genet 2009 Oct;5(10):e1000628 [FREE Full text] [doi: 10.1371/journal.pgen.1000628] [Medline: 19798439]
122. Masca N, Burton PR, Sheehan NA. Participant identification in genetic association studies: improved methods and practical implications. Int J Epidemiol 2011 Dec;40(6):1629-1642 [FREE Full text] [doi: 10.1093/ije/dyr149] [Medline: 22158671]

123. Pardo R, Rafnsson W, Steinhorn G, Lavrov D, Lumley T, Probst C, et al. Privacy with good taste: a case study in quantifying privacy risks in genetic scores. arXiv Preprint posted online August 26, 2022 [FREE Full text] [doi: 10.1007/978-3-031-25734-6_7]
124. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. Genome Biol 2019 Nov 20;20(1):246 [FREE Full text] [doi: 10.1186/s13059-019-1828-7] [Medline: 31747936]
125. Pang AW, MacDonald JR, Pinto D, Wei J, Rafiq MA, Conrad DF, et al. Towards a comprehensive structural variation map of an individual human genome. Genome Biol 2010;11(5):R52 [FREE Full text] [doi: 10.1186/gb-2010-11-5-r52] [Medline: 20482838]
126. Chen W, Hayward C, Wright AF, Hicks AA, Vitart V, Knott S, et al. Copy number variation across European populations. PLoS One 2011;6(8):e23087 [FREE Full text] [doi: 10.1371/journal.pone.0023087] [Medline: 21829696]
127. Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, et al. The impact of structural variation on human gene expression. Nat Genet 2017 May 3;49(5):692-699 [FREE Full text] [doi: 10.1038/ng.3834] [Medline: 28369037]
128. Ueki M, Takeshita H, Fujihara J, Kimura-Kataoka K, Iida R, Yasuda T. Simple screening method for copy number variations associated with physical features. Leg Med (Tokyo) 2017 Mar;25:71-74. [doi: 10.1016/j.legalmed.2017.01.006] [Medline: 28457514]
129. Weischenfeldt J, Symmons O, Spitz F, Korbel JO. Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet 2013 Feb;14(2):125-138. [doi: 10.1038/nrg3373] [Medline: 23329113]
130. Ho SS, Urban AE, Mills RE. Structural variation in the sequencing era. Nat Rev Genet 2020 Mar;21(3):171-189 [FREE Full text] [doi: 10.1038/s41576-019-0180-9] [Medline: 31729472]
131. Shabani M, Marelli L. Re-identifiability of genomic data and the GDPR: assessing the re-identifiability of genomic data in light of the EU General Data Protection Regulation. EMBO Rep 2019 Jun;20(6):e48316 [FREE Full text] [doi: 10.15252/embr.201948316] [Medline: 31126909]
132. Molnár-Gábor F, Korbel JO. Genomic data sharing in Europe is stumbling-Could a code of conduct prevent its fall? EMBO Mol Med 2020 Mar 06;12(3):e11421 [FREE Full text] [doi: 10.15252/emmm.201911421] [Medline: 32072760]
133. Martinez-Martin N, Magnus D. Privacy and ethical challenges in next-generation sequencing. Expert Rev Precis Med Drug Dev 2019;4(2):95-104 [FREE Full text] [doi: 10.1080/23808993.2019.1599685] [Medline: 32775691]
134. Clayton EW, Evans BJ, Hazel J, Rothstein MA. The law of genetic privacy: applications, implications, and limitations. J Law Biosci 2019:1-36. [doi: 10.2139/ssrn.3384321]

Abbreviations

CNV: copy number variation
CODIS: Combined DNA Index System
DTC-GT: direct-to-consumer genetic testing
SGMF: Sorenson Molecular Genealogy Foundation
SNP: single nucleotide polymorphism
SNV: single nucleotide variant
STR: short tandem repeat
SV: structural variant
Y-STR: short tandem repeat on the Y chromosome

Edited by Z Yue; submitted 06.11.23; peer-reviewed by L Guo, J Lai; comments to author 01.01.24; revised version received 26.03.24; accepted 29.03.24; published 27.05.24.

Please cite as:
Thomas M, Mackes N, Preuss-Dodhy A, Wieland T, Bundschus M
Assessing Privacy Vulnerabilities in Genetic Data Sets: Scoping Review
JMIR Bioinform Biotech 2024;5:e54332
URL: https://bioinform.jmir.org/2024/1/e54332
doi:10.2196/54332
PMID:38935957

Original Paper

Comparison of the Neutralization Power of Sotrovimab Against SARS-CoV-2 Variants: Development of a Rapid Computational Method

Dana Ashoor[1], MSc, PhD; Maryam Marzouq[1], MSc, PhD; M-Dahmani Fathallah[1], MSc, MBA, DPhil, PhD

Department of Life Sciences, Health Biotechnology Program - King Fahad Chair for Health Biotechnology, College of Graduate Studies, Arabian Gulf University, Manama, Bahrain

Corresponding Author:
Dana Ashoor, MSc, PhD
Department of Life Sciences, Health Biotechnology Program - King Fahad Chair for Health Biotechnology
College of Graduate Studies
Arabian Gulf University
Road 2904 Building 293
Manama, 329
Bahrain
Phone: 973 17239875
Email: danana@agu.edu.bh

Abstract

Background: The rapid evolution of SARS-CoV-2 imposed a huge challenge on disease control. Immune evasion caused by genetic variations of the SARS-CoV-2 spike protein's immunogenic epitopes affects the efficiency of monoclonal antibody–based therapy of COVID-19. Therefore, a rapid method is needed to evaluate the efficacy of the available monoclonal antibodies against the new emerging variants or potential novel variants.

Objective: The aim of this study is to develop a rapid computational method to evaluate the neutralization power of anti–SARS-CoV-2 monoclonal antibodies against new SARS-CoV-2 variants and other potential new mutations.

Methods: The amino acid sequence of the extracellular domain of the spike proteins of the severe acute respiratory syndrome coronavirus (GenBank accession number YP_009825051.1) and SARS-CoV-2 (GenBank accession number YP_009724390.1) were used to create computational 3D models for the native spike proteins. Specific mutations were introduced to the curated sequence to generate the different variant spike models. The neutralization potential of sotrovimab (S309) against these variants was evaluated based on its molecular interactions and Gibbs free energy in comparison to a reference model after molecular replacement of the reference receptor-binding domain with the variant's receptor-binding domain.

Results: Our results show a loss in the binding affinity of the neutralizing antibody S309 with both SARS-CoV and SARS-CoV-2. The binding affinity of S309 was greater to the Alpha, Beta, Gamma, and Kappa variants than to the original Wuhan strain of SARS-CoV-2. However, S309 showed a substantially decreased binding affinity to the Delta and Omicron variants. Based on the mutational profile of Omicron subvariants, our data describe the effect of the G339H and G339D mutations and their role in escaping antibody neutralization, which is in line with published clinical reports.

Conclusions: This method is rapid, applicable, and of interest to adapt the use of therapeutic antibodies to the treatment of emerging variants. It could be applied to antibody-based treatment of other viral infections.

(JMIR Bioinform Biotech 2024;5:e58018) doi:10.2196/58018

KEYWORDS

in silico; anti–SARS-CoV-2; neutralizing antibody; Sotrovimab; S309; variants; SARS-CoV-2; Omicron; subvariants; computational method; monoclonal; amino acid; protein; mutation

Introduction

While the world has entered its fourth year of the COVID-19 pandemic caused by the newly emergent SARS-CoV-2, this persistent virus is still lingering away. This is mainly due to the virus' relatively high mutational rate, with specific mutations occurring on the spike protein affecting its immunogenicity [1,2]. The battle against this virus covers several aspects ranging

from prevention, mitigation, and treatment. One promising approach that is still developing with proven efficiency consists of using anti–SARS-CoV-2 monoclonal neutralizing antibodies (NAbs). However, selective pressure caused by infection and/or vaccination is accelerating the emergence of new variants and subvariants, which poses a challenge on not only antibody-mediated therapy but also vaccine use and development. Anti–SARS-CoV-2 monoclonal antibodies recognize specific epitopes mainly on the spike protein and prevent target cell binding and/or fusion, and accumulation of mutations in these specific epitopes increases the fitness of the virus. Additionally, the efficacy of the available anti–SARS-CoV-2 NAb therapies varies drastically, and it is difficult to foresee how useful would it be for new circulating variants [3]. Therefore, there is an urgent need for the rapid assessment of anti–SARS-CoV-2 monoclonal antibodies' potential efficiency to treat emergent variants. Toward this end, computational methods aimed at the rapid estimation of the binding affinity and molecular interactions between new variants and a given monoclonal antibody can be used.

Currently, the Food and Drug Administration and the European Medicines Agency have issued emergency use authorization for several anti–SARS-CoV-2 NAbs including Evusheld, Ronapreve and Regkirona, sotrovimab (S309), casirivimab and imdevimab, and bamlanivimab [4,5] and many more are still under evaluation. Based on their binding site, these NAbs are classified into different groups. There are currently 2 classification methods [6]. One of these methods is based on a high-throughput surface plasmon resonance technique combined with negative-stain electron microscopy to identify specific epitopes on the receptor-binding domain (RBD). This method groups the NAbs into 7 distinct communities: RBD-1 through RBD-3, which bind to the receptor-binding motif; RBD-4 and RBD-5, which bind to the outer face of the RBD; and RBD-6 and RBD-7, which bind to the inner face of the RBD. The other method is based on considerations such as the overlap between the NAb with the angiotensin-converting enzyme 2 (ACE2) receptor-binding site and whether it recognizes activated (up) or baseline (down) states of RBD. Four different classes (I-IV) were described: class I competes on the ACE2 binding site and can bind with the RBD in its up position, while class II binds with the RBD in both states (up and down); class III NAb binds at an interface that is outside the RBD domain and hence does not compete with the ACE2 receptor, and binds with both forms of the RBD (up and down); while class IV binds only with RBDs in the up state [7,8].

The computational method we describe in this paper was developed to evaluate the interaction between a given NAb of a specific SARS-CoV-2 variant, compare the interaction of the same antibody with different SARS-CoV-2 variants, and thus predict a possible immune evasion. It is used to describe a model of the interaction between the neutralizing monoclonal antibody S309 and the original SARS-CoV-2 Wuhan variant. This monoclonal antibody was first isolated from the memory B lymphocytes of a SARS-CoV survivor [9,10] and is reported to have neutralization potencies toward the severe acute respiratory syndrome (SARS) coronavirus (SARS-CoV), SARS-CoV-2, and SARS-like coronaviruses. Currently, it is one of only 2 approved therapeutic monoclonal antibodies for newly emerged Omicron subvariants [7,11,12]. S309 is a recombinant human monoclonal antibody used under the generic name Xevudy. In May 2021, it was first granted for emergency use for early treatment of COVID-19 [13]. S309 belongs to class III antibodies that are characterized by their binding site on the spike protein, as they do not compete with the ACE2 receptor [7]. While ACE2 binds to the SARS-CoV-2 spike residues between residues K417 and Y505 [14], S309 recognizes a distinct proteoglycan epitope opposite the ACE2 binding site involving residues N334, E340, N343, T345, R346, K356, and a structural loop (443-450) that can be accessed on both states of the RBD (up and down). These key glycan residues are not affected by mutations of the new omicron subvariants [7,15]. However, other mutations found on the structural loop seem to have a significant effect on the neutralization capacity of S309. Since S309 does not compete with the ACE2 receptor binding site, its neutralization mechanism does not depend on direct blocking of the RBD. Nonetheless, binding of S309 to the SARS-CoV-2 spike protein's RBD induces antibody-dependent cell cytotoxicity and antibody-dependent cellular phagocytosis [16].

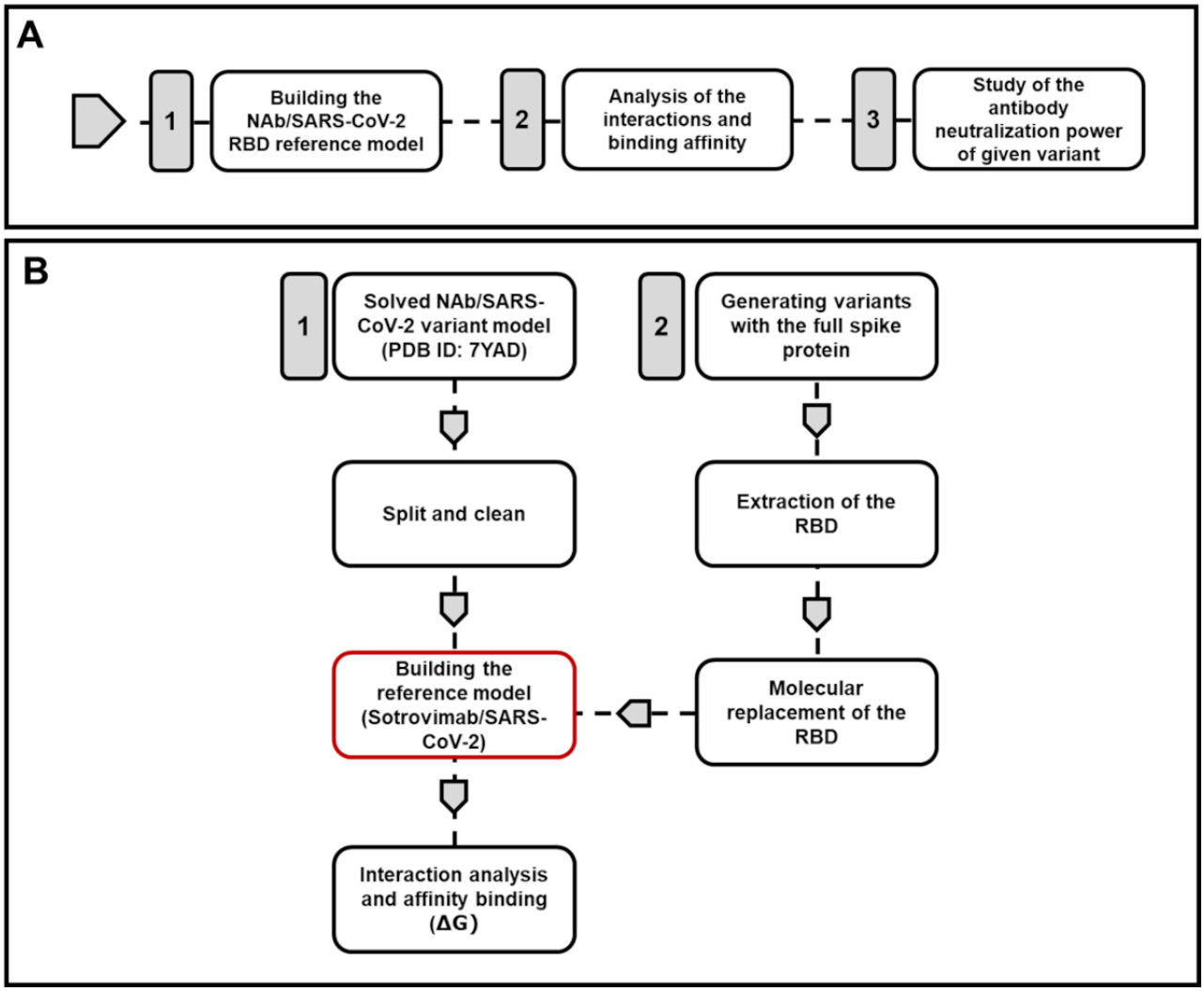
Several experimental and clinical reports have described the neutralizing effect of monoclonal antibody S309 with the original SARS-CoV-2 Wuhan strain and its effect in reducing disease progression [10,17,18]. Therefore, in the computational method we report in this paper, the estimated interaction affinity of the monoclonal antibody S309 to the original SARS-CoV-2 Wuhan strain is assigned a value of 100%. Comparison of the estimated affinities of S309 to each SARS-CoV-2 variant to this reference value facilitates the evaluation of the neutralization efficiency of S309 and the prediction of possible immune evasion for each existing or newly emerging variant. This straightforward computational method can rapidly provide valuable insights on the eventual efficiency of existing neutralizing therapeutic antibodies in treating newly emergent variants prior to the experimental methods. Since immune evasion is a major criterion listed by the World Health Organization and the Centers for Disease Control and Prevention in their labeling systems of new variants, particularly the variants of concern [19], this method can also be considered to label new variants early after their emergence.

Methods

Overview

This work describes a computational method to evaluate the effect of different SARS-CoV-2 mutations on the binding affinity of available NAbs and on the stability of the complex. As a working pattern, we developed a reference complex model between the NAb S309 and the original SARS-CoV-2 Wuhan strain. We evaluated the other variants and subvariants based on the differences of their specific molecular interactions and Gibbs free energy (ΔG) with S309. Figure 1 outlines the methods used to determine the anti–SARS-CoV-2 antibody neutralization potential of S309.

Figure 1. Method outline. (A) Outline of the 3 steps in the method. (B) Workflow of the in silico method for the evaluation of the neutralization power of a SARS-CoV-2 monoclonal antibody. NAb: neutralizing antibody; PDB: Protein Data Bank; RBD: receptor-binding domain.

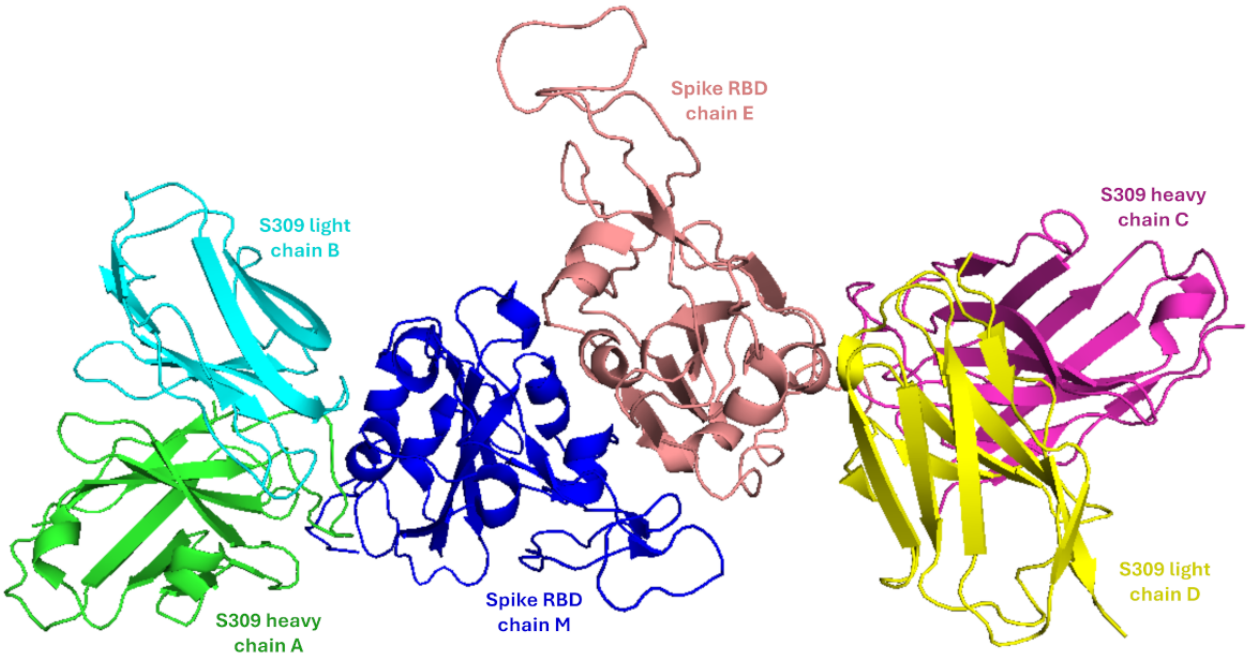


Construction of the Models and Complexes

Building the NAb/SARS-CoV-2 RBD Reference Model

We used a model (Protein Data Bank ID 7YAD) downloaded from Research Collaboratory for Structural Bioinformatics Protein Data Bank [20] to generate our reference model representing the interaction of S309's variable domain (Fv) with the spike protein of the SARS-CoV-2 Omicron variant. The Protein Data Bank model (7YAD) represents the interaction of the SARS-CoV-2 Omicron RBD (residues P330-K529) with the Fv domain of S309. The model shows 6 chains (2 RBDs, 2 heavy chains, and 2 light chains) forming 2 subunits of the RBD-S309 Fv (Figure 2). The selection criteria of the 7YAD model [15] are the generation of a 3D structure via electron microscopy, a high resolution of 2.66 Å, and a relatively good validation report. In addition, it represents the interaction with the SARS-CoV-2 RBD in its open state. Upon downloading the structure, only 1 unit was selected to represent 1 S309 Fv (1 heavy chain and 1 light chain) binding to 1 spike RBD, chains A, B, and M. The complex was extracted, cleaned from any heteroatoms, and used as a reference model to generate the different variant complexes via RBD replacement.

Figure 2. 3D structure of the Protein Data Bank model 7YAD showing 2 subunits of the Sotrovimab (S309) variable domain (Fv; heavy and light chains) binding to the spike protein's receptor-binding domain in Omicron variants.



Retrieval of SARS-CoV and SARS-CoV-2 Variants' Sequences, Modifications, and Modeling

The amino acid sequences of the extracellular domains of SARS-CoV and SARS-CoV-2 spike protein were acquired from the National Center for Biotechnology Information (NCBI) GenBank database (IDs YP_009825051.1 and ID: YP_009724390.1, respectively). SARS-CoV-2 variant–specific mutations were introduced to the curated sequence to generate the different variant sequences based on published mutations in databases such as CoVariants [21] and the Stanford University SARS-CoV-2 Variants database [22]. The sequences corresponding to the spike protein of SARS-CoV and 25 variants of SARS-CoV-2 (including Alpha, Beta, Gamma, Delta-21J, and Kappa strains), in addition to the Omicron strain's subvariants (BA.1, BA.2, BA.4/BA.5, BA.2.12.1, BA.2.75, BQ1, XBB, and XBB.1) were used to build 3D monomer models of the spike protein. The monomers were modeled in an open state using SWISS-MODEL server's User Template Mode [23]. The template for each monomer was selected and extracted from Protein Data Bank. Selection criteria were based on resolution, chain quality, sequence gaps, furin site and proline modifications, and validation report. The templates used for each model are listed in Table 1. The monomer chain representing the open-state RBD was extracted from each model, cleaned from any heteroatoms, and saved using PyMol software [24] into a new Pdb file. Each monomer was introduced in the SWISS-MODEL server's User Template Mode to generate an open-state monomer spike protein for SARS-CoV, SARS-CoV-2 variants, and Omicron subvariants.

Table 1. List of templates and chains (with their PDB[a] IDs) used to build the extracellular domains of the spike protein of SARS-CoV[b] and the different SARS-CoV-2 variants.

Virus	PDB model ID	Resolution (Å)	Selected chain	Reference
SARS-CoV	6ACD	3.9	C	Song et al [25]
SARS-CoV-2–Wuhan[c]	7ND9	2.80	B	Dejnirattisai et al [26]
Alpha	8DLI	2.56	A	Mannar et al [27]
Beta	8DLL	2.56	A	Mannar et al [27]
Delta-21J	7W92	3.1	C	Wang et al [28]
Gamma	8DLO	2.25	A	Mannar et al [27]
Kappa	7TF0	3.02	B	Saville et al [29]
Omicron	7XCO	2.5	C	Zhao et al [15]

[a]PDB: Protein Data Bank.

[b]SARS-CoV: severe acute respiratory syndrome coronavirus.

[c]This refers to the original SARS-CoV-2 Wuhan strain.

Construction of RBD/S309 Complexes

The RBDs of the SARS-CoV, SARS-CoV-2 variants, and Omicron subvariants were extracted from the generated models, and the complexes with S309 were constructed via molecular replacement. The reference crystalized RBD chain M of 7YAD was replaced with the modeled RBD. The complex was saved and energy minimized. Energy minimization was carried out in vacuo, without a reaction field, using the GROMOS 43B1 force field [30] and the Swiss-pdb Viewer (version 4.1.0) [31]. This was applied to all the generated models.

Interactions and Complex Binding Affinity Analysis

The interactions between the RBD of the spike protein of SARS-CoV, SARS-CoV-2 variants, and Omicron subvariants with NAb S309 were analyzed based on polar and hydrophobic interactions using the LigPlot+ software [32]. Stability and affinity were assessed based on thermodynamic measure of the formed complex's energy, Gibbs free energy (ΔG), using a web-based antibody-antigen binding affinity tool CSM-AB [33]. Binding affinity percentage was calculated in reference to that of the original SARS-CoV-2 Wuhan strain/S309 complex.

Testing the Generated Method by Analyzing Newly Reported Omicron Subvariants and Some Experimentally Tested Mutations

Several reports have discussed the neutralizing effect of NAbs and possible antibody escape of some new Omicrons subvariants [34-39]. Here we used our developed method to evaluate the binding affinity of several of these new subvariants including AY.1, XBB.1.5, BF.7, BQ.1.1, BA.1.1, BA.3, BA.2.3.20, BM.1.1.1, BA.5.6.2, BA.2.75.2, and CH.1.1 (Orthrus), with the NAb S309. Additionally, the effect of several amino acid substitutions in the NAb epitope have been tested experimentally using the enzyme-linked immunosorbent assay and/or pseudovirus neutralization assays. Several mutations are reportedly resistant to inhibition by S309 leading to an antibody escape. These key residues include R346, P337, G339, N440, and S371 [40,41]. Therefore, we applied our method to computationally test the effect of some mutations on these residues. As we already generated parent RBD sequences, newly emerged mutations were introduced, new models and complexes were built, and the mutation's effect on binding energy with the NAb was predicted by recalculating complex's ΔG in reference to that of the parent complex and binding affinity with the original SARS-CoV-2 Wuhan strain.

Ethical Considerations

This study was exempt from ethical review since it was conducted in silico and no human subjects were involved.

Results

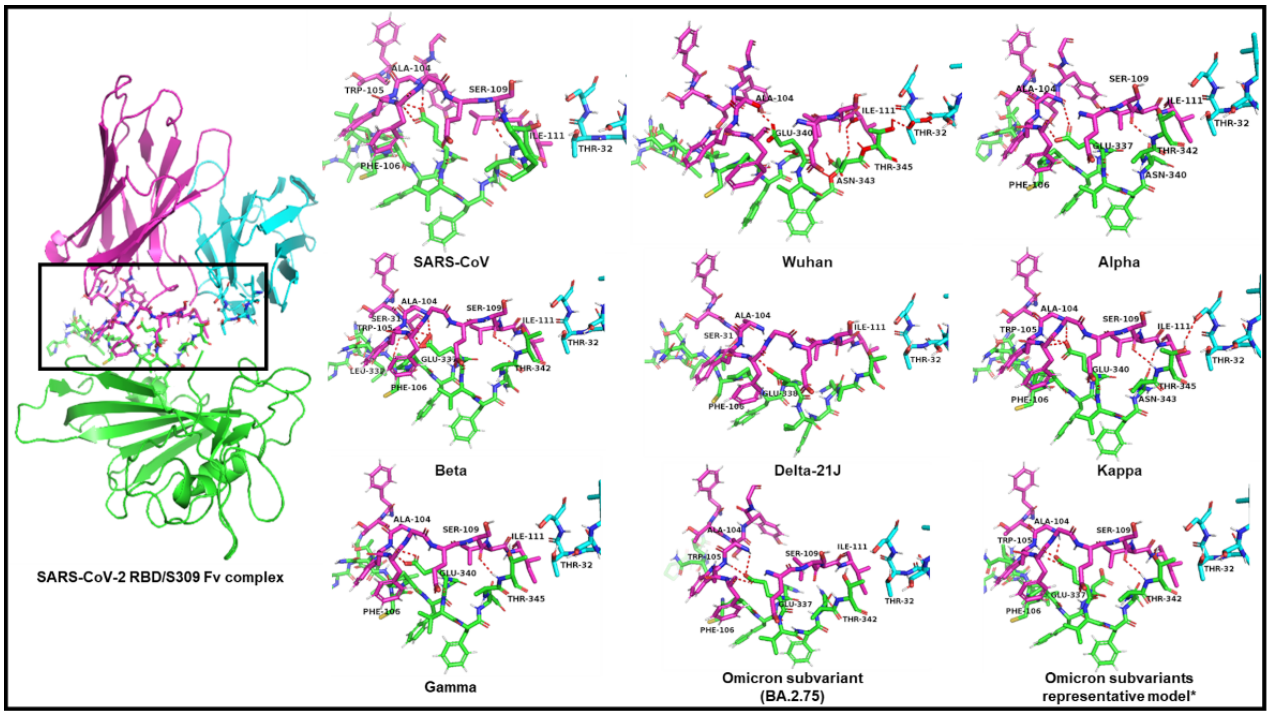
Method Development Workflow

Figure 1 outlines the methods for assessing the anti–SARS-CoV-2 neutralization potential of S309. The blueprint of the method we developed using monoclonal antibody S309—an experimentally proven neutralizing monoclonal antibody for SARS-CoV-2 and its variants—is described in Figure 1A. We proceeded by modifying the available model 7YAD to generate a reference model that can be used to measure neutralization potential in terms of binding affinity ΔG (Figure 1B). Several in silico 3D models representing spike monomer chain of each variant were generated. The quality of the generated 3D model was evaluated based on the homology modeling report and SWISS-MODEL structural assessment. The generated models showed a QMEAN *z* score between –1.0 and –3.2 indicating a good-quality model where *z* scores of around 0.0 are ideal and any value below –4.0 indicates a low-quality model [42]. The QMEANDisCo global score represents the combined scoring of global (for the entire structure) and local (per residue) absolute quality estimates of a single model [43]. Our models' QMEANDisCo global scores ranged from 0.64 to 0.76 (SD 0.05). These values reflect a good-quality model (any value below 0.6 represents a low-quality model). Each complex was built by molecular replacement of chain M of the reference model with the extracted RBD, followed by binding affinity and interaction analyses.

Analysis of the Molecular Interaction Pattern of S309 With 9 Main SARS-CoV-2 Variants

The generated complexes were energy-minimized and polar and hydrophobic interactions were analyzed. Several interactions were identified between the S309 Fv domain and spike RBD with more interactions toward the heavy chain. Interacting residues of the spike protein include residue 321-428 in SARS-CoV and 334-441 in SARS-CoV-2 and its variants. SARS-CoV showed 4 polar interactions compared to the original SARS-CoV-2 Wuhan strain that shares a total of 3 polar interactions with S309. Interestingly, variant Kappa showed the highest number of polar interactions (n=6), while variant Delta-21J showed the lowest (n=1) number of polar interactions. Variant Kappa showed 2 unique salt bridges between residues R346 and K356 with the S309 heavy chain residue E108. All the variants share the same polar interaction between E340 and S309 heavy chain A104 except for variant Delta-21J. All Omicron subvariants showed the same interaction pattern except for BA.2.75 with 1 missing polar interaction between T343 and S309 heavy chain S109. Variant Gamma showed more hydrophobic interactions with the light chain of S309. All polar interactions are represented in Figure 3 and detailed interactions are listed in Multimedia Appendix 1.

Figure 3. Variations of the polar interactions between the monoclonal antibody sotrovimab (S309) and different SARS-CoV-2 variants and subvariants. The monoclonal antibody's heavy chain (magenta), light chain (cyan), SARS-CoV-2 S spike protein–receptor-binding domain (RBD; green). *Residue numbering: BA.1, BA.2, BA.2.12.1 (D337 and T342)/BA.4, BA.5, BQ.1 (D335 and T340)/XBB, and XBB.1.



Evaluation of the Binding Affinity of S309 With 9 SARS-CoV-2 Variants by Comparing Their Binding Affinity With the Original SARS-CoV-2 Wuhan Reference Strain

The thermodynamic stability of the generated complexes was measured via computational prediction of ΔG using the CSM-AB tool. ΔG reflects energy differences between coupled and decoupled antibody-antigen complexes. This difference in energy indicates complex stability where a negative normalized energy (ΔG<0) indicates spontaneous and exergonic reactions and hence more stable complexes and more efficient protein–ligand interactions. Thus, the lower the value of ΔG, the more stable the (antibody-antigen) complex. In our model, we found that the NAb S309 has a binding affinity of –8.26 kcal/mol with SARS-CoV and –7.13.26 kcal/mol with SARS-CoV-2, indicating a loss in binding affinity. However, comparing SARS-CoV-2 variants to the binding affinity of the first Wuhan strain showed an improvement in the binding affinity of S309 with variants Alpha, Beta, Gamma, and Kappa. This improvement in affinity, when compared to the interaction profile, can be related to the increased number of polar and hydrophobic interactions and more similar interaction profiles with SARS-CoV than with the original SARS-CoV-2 Wuhan strain. In contrast, variant Delta showed a substantial decrease in binding affinity as it exhibited only 1 polar interaction. All Omicron subvariants shared similar interaction profiles; however, they exhibited different binding affinities. Although they showed a significant decrease in binding affinity compared to original SARS-CoV-2 Wuhan strain, they can be clustered in 2 groups: those with a G339H mutation (BA.2.75, XBB, and XBB.1) and those with a G339D mutation (BA.1, BA.2, BA.4/5, BQ.1, and BA.2.12.1; Figure 4 and Multimedia Appendix 1). The data show that the H339 residue slightly enhanced binding affinity compared to the D residue substitution. This residue is located in the middle of the interaction loop and hence plays a marked role in maintaining the complex's stability and binding affinity. In addition, our results are in line with the reported effect of the G339D mutation and its role in escaping antibody neutralization [41,44,45].

Furthermore, to test the impact of a mutation in residue G339, we analyzed the effect of reverse mutagenesis. We used the generated models and in silico tools to test the effect of reverse mutation at residue G339 on complex stability in subvariants BA2.75, XBB, and XBB.1. They have an aspartic acid residue at position 339. By reversing this residue to either glycine or histidine (G339 or H339), we calculated the effect in the form of the ΔG value. Our results showed an increase in the stability of the SARS-CoV-2/S309 complex and hence enhanced binding affinity with the glycine residue. However, reverse mutagenesis to histidine has no to a very low effect, except for subvariant BA.2.12.1 where there was a slight increase in binding affinity (Table 2).

Figure 4. Binding energy (ΔG) of the severe acute respiratory syndrome coronavirus (SARS-CoV) and different SARS-CoV-2 variants (represented in affinity percentage in comparison to SARS-CoV-2).

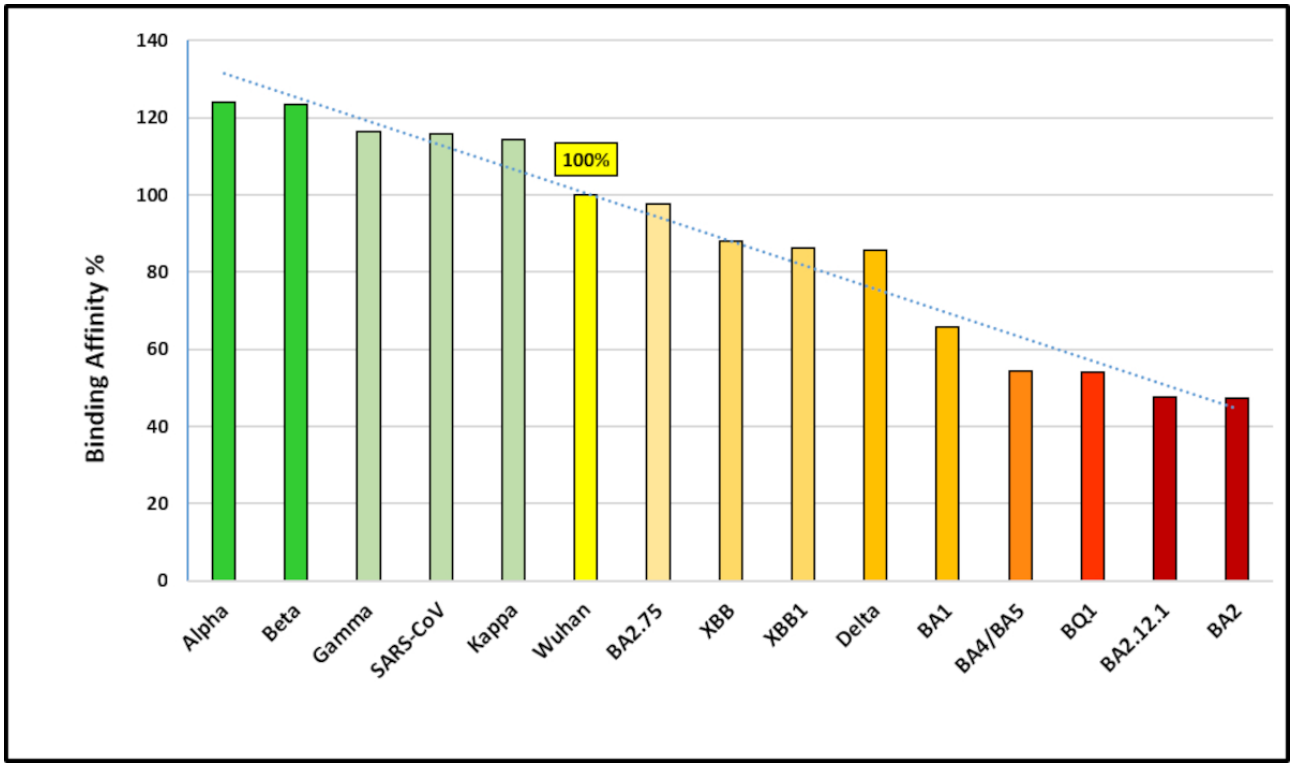


Table 2. Gibbs free energy (ΔG) analysis of the effect of the D339 reverse mutation on the binding affinity of SARS-CoV-2 Omicron subvariants with the neutralizing antibody sotrovimab.

SARS-CoV-2 Omicron subvariant	D339 ΔG (kcal/mol)	D339G substitution ΔG (kcal/mol)	D339H substitution ΔG (kcal/mol)	Effect on binding affinity
BA.1	–4.7	–6.83	–6.92	Increase
BA.2	–3.38	–6.18	–6.59	Increase
BA.4/BA.5	–3.87	–6.96	–7.15	Increase
BA.2.12.1	–3.39	–6.19	–6.59	Increase
BQ.1	–3.86	–6.92	–7.29	Increase

Evaluation of S309's Binding Affinity to Experimentally Tested SARS-CoV-2 Variants and Some Hypothetical SARS-CoV-2 Variants

The effect of several amino acid substitutions in the NAb S309 epitope have been tested experimentally using the enzyme-linked immunosorbent assay and/or pseudovirus neutralization assays. These mutations resulted in resistance to neutralization by S309, leading to antibody escape. These key substitutions include R346S and P337L, G339D, N440K, and S371L [40,41]. Here we used our developed method to evaluate this effect computationally. By generating models with the newly reported mutations and CSM-AB tool, we predicted the effect of the reported mutations on the binding affinity of the complex and hence on neutralizing effect of S309. Interestingly, our computational results are comparable with the experimentally reported effect of these mutations on the S309 evasion from monoclonal antibodies. Additionally, we predicted a possible effect of hypothetical mutations on some of the proteoglycan epitopes (Table 3).

Table 3. Prediction of the effect of the newly reported SARS-CoV-2 subvariants and some experimentally tested spike mutations on the binding affinity with sotrovimab.

Parent model, variants, and subvariants	Gibbs free energy (kcal/mol)	Mutations	References	New subvariants	Gibbs free energy (Kcal/mol)	Effect on binding affinity	Binding affinity in reference to the original SARS-CoV-2 Wuhan strain (%)
Newly reported SARS-CoV-2 variants and subvariants							
BA.4/5	–3.87	R346T	Qu et al [39]	BF.7	–2.85	Decrease	39.97
BQ.1	–3.86	R346T	Qu et al [39]	BQ.1.1	–2.82	Decrease	39.55
BA.1	–4.7	R346K	Manjunath et al [34], Liu et al [35], and Martins et al [36]	BA.1.1	–4.99	Increase	69.99
BA.1	–4.7	L371F and D405N	Stanford University [22]	BA.3	–4.99	Increase	69.99
BA.2	–3.38	K444R, N450D, L452M, N460K, A484R, and R493Q	Stanford University [22]	BA.2.3.20	–4.13	Increase	47.41
BA.2	–3.38	D339H, R346T, G446S, N460K, F486S, F490S, and R493Q	Stanford University [22]	BM.1.1.1	–5.83	Increase	81.77
BA.4/5	–3.87	K444T	Stanford University [22]	BA.5.6.2	–3.86	Decrease	54.14
DELTA-21J	–6.12	K417N	Stanford University [22]	AY.1	–6.12	No effect	85.83
BA.2.75	–6.96	R346T and F486S	Qu et al [39]	BA.2.75.2	–6.27	Decrease	90.74
BA.2.75	–6.96	R346T, K444T, L452R, and F486S	Neher [38]	CH.1.1 (Orthrus)	–5.73	Decrease	80.36
XBB.1	–6.15	S486P	Yue et al [37]	XBB.1.5 (Kraken)	–6.15	No effect	86.26
Experimental							
Wuhan	–7.13	R346K	Magnus et al [40]	—[a]	–7.23	Increase	101.4
Wuhan	–7.13	R346S	Magnus et al [40]	—	–6.21	Decrease	87.1
Wuhan	–7.13	R346T	Magnus et al [40]	—	–6.97	Decrease	97.75
Wuhan	–7.13	P337L	Magnus et al [40]	—	–6.73	Decrease	94.39
Wuhan	–7.13	P337L and R346K	Magnus et al [40]	—	–5.45	Decrease	76.45
Omicron BA.2.75	–6.96	H339D	Cao et al [41]	—	–3.34	Decrease	46.84
Omicron BA.2.75	–6.96	R346K	Cao et al [41]	—	–6.52	Decrease	91.44
Omicron BA.2.75	–6.96	S371L	Cao et al [41]	—	–6.53	Decrease	91.58
Omicron BA.2.75	–6.96	Q493R	Cao et al [41]	—	–6.81	Decrease	95.51

[a]Not applicable.

Discussion

Principal Findings

Antibody-based therapies have proven effective against SARS-CoV-2 infection and appear to be the most promising approach to control the COVID-19 pandemic. A number of neutralizing monoclonal antibodies used in the clinical setting have shown highly favorable results, particularly in stopping disease progression [46,47]. However, the constant emergence of new virus variants has hindered the potency of available anti–SARS-CoV-2 antibodies and urged the continuous development of improved, more effective NAbs. In this study, we describe an in silico rapid method that we developed to predict a possible effect of newly emerged mutations on the efficacy of available neutralizing anti–SARS-CoV-2 antibodies. We used the monoclonal antibody S309 as an example. S309 recognizes a proteoglycan epitope embedded in a structural loop located on the outer side the SARS-CoV-2 spike protein and encompasses residues 334-441 (Multimedia Appendix 1). This specific epitope location permits the binding to RBD in both the up and down configurations without affecting binding to the ACE2 receptor. Indeed, this epitope does not overlap with the ACE2 binding site. However, several newly emerged RBD mutations were reported to have an impact on the neutralizing effect of S309. To further explore this, we developed this computational method to evaluate and compare the neutralization potential of S309 against different SARS-CoV-2 variants and possible new emerging mutations (Figure 1).

Using bioinformatics tools, we developed spike models for several new SARS-CoV-2 variants and evaluated the effect of several emerged mutations on the interaction with the neutralizing monoclonal antibody S309 used for the treatment of mild-to-moderate COVID-19. In addition, by applying this method, we foresee the effect of some predicted or not yet observed mutations. Interestingly, the predicted significantly decreased computational neutralization values of the monoclonal antibody S309 (from 10% to 50%) for some new Omicron subvariants are confirmed by the newly published clinical results indicating a reduction in its effectiveness against these same new Omicron subvariants and possible immune evasion [39,48-51]. Early on, S309 was clinically considered one of the most effective monoclonal antibodies against all SARS-CoV-2 variants [7]. However, this statement has been proven wrong as recent convergent evolution of Omicron and its subvariants has led to a new set of spike mutations within the S309 epitope, and, consequently, the new subvariants became increasingly resistant [52]. Several mutations were identified to be critical, and others are yet to be investigated. For example, a substitution in the nonpolar G339 residue located at the center of the antibody epitope to the acidic charged aspartic acid residue (G339D) has been shown to have a remarkable impact on the binding affinity of Omicron's subvariants [44,53], with a predicted reduction in neutralization power of 30% for BA.1; 45% for BA.4, BA.5, and BQ.1; 50% for BA.2.12.1 and BA.2; and 60% for BF.7 and BQ.1.1. We reported a similar effect in our proposed computational method and we found that the impact was less intense with the G339H mutation (Table 2 and Multimedia Appendix 1). However, the combination of multiple mutations in Omicron subvariants has a more profound effect on binding affinity, indicating increased antibody resistance. This effect was clearly detected in the subsequent, potentially dominant new subvariants BM.1.1.1, BA.2.3.20, and CH.1.1 (Orthrus) [54] (Table 3). Furthermore, to test our method, we examined some experimentally evaluated mutations in residues P337, R346, G339, and S371 that are located in the S309 epitope, and once more, our computational method was compatible with the experimental results (Table 3). This reduced susceptibility of S309 with mutations in residues P337, R346, and other residues has been experimentally recognized [13,40,41]. Considering the clinical observations of the efficiency of Sotrovimab in neutralizing SARS-CoV, SARS-CoV-2 variants, and Omicron subvariants, a 50% reduction in binding affinity, compared to that in the reference model, may be considered the cutoff for determining whether a monoclonal antibody will neutralize a new variant, using the method described in this paper. Comparison of the predicted values of the evaluation of neutralizing power with a larger number of clinical observations about the efficiency of a neutralizing monoclonal antibody would help refine this theoretical cutoff value and further validates the method. Ultimately, molecular dynamics simulations can be performed to more accurately define the most stable conformation of monoclonal antibody/spike protein–RBD complexes.

Conclusions

This in silico method provides significant insights into possible antibody escape following the emergence of new SARS-CoV-2 mutants and helps evaluate the usefulness of existing NAbs in combating new emerging variants and subvariants. This method is straightforward, rapid, and applicable ahead of obtaining statistically significant clinical observations. In addition, this method highlights the advantages of computational approaches in viral the rapid surveillance and for the development of novel monoclonal antibody therapies.

Acknowledgments

All authors declared that they had insufficient or no funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided APF support for the publication of this article.

Data Availability

All data generated or analyzed in this study are included in this published article and in Multimedia Appendix 1.

Authors' Contributions

DA carried out the in silico analysis, designed the methodology, curated the data, and drafted and edited the manuscript. MM designed the illustrations and figures. M-DF conceptualized the study, analyzed the data, drafted and edited the manuscript, and supervised the study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional information.
[DOCX File , 304 KB - bioinform_v5i1e58018_app1.docx]

References

1. van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect Genet Evol 2020 Sep;83:104351 [FREE Full text] [doi: 10.1016/j.meegid.2020.104351] [Medline: 32387564]
2. Amicone M, Borges V, Alves MJ, Isidro J, Zé-Zé L, Duarte S, et al. Mutation rate of SARS-CoV-2 and emergence of mutators during experimental evolution. Evol Med Public Health 2022;10(1):142-155 [FREE Full text] [doi: 10.1093/emph/eoac010] [Medline: 35419205]
3. Focosi D, Quiroga R, McConnell S, Johnson MC, Casadevall A. Convergent evolution in SARS-CoV-2 spike creates a variant soup from which new COVID-19 waves emerge. Int J Mol Sci 2023 Jan 23;24(3):2264 [FREE Full text] [doi: 10.3390/ijms24032264] [Medline: 36768588]
4. Emergency Use Authorization. Food and Drug Administration. URL: https://www.fda.gov/emergency-preparedness-and-response/mcm-legal-regulatory-and-policy-framework/emergency-use-authorization#coviddrugs [accessed 2024-09-17]
5. COVID-19 medicines. European Medicines Agency. URL: https://www.ema.europa.eu/en/human-regulatory/overview/public-health-threats/coronavirus-disease-covid-19/treatments-vaccines/covid-19-treatments [accessed 2024-09-17]
6. Hastie KM, Li H, Bedinger D, Schendel SL, Dennison SM, Li K, CoVIC-DB team1, et al. Defining variant-resistant epitopes targeted by SARS-CoV-2 antibodies: a global consortium study. Science 2021 Oct 22;374(6566):472-478 [FREE Full text] [doi: 10.1126/science.abh2315] [Medline: 34554826]
7. Mittal A, Khattri A, Verma V. Structural and antigenic variations in the spike protein of emerging SARS-CoV-2 variants. PLoS Pathog 2022 Feb 17;18(2):e1010260 [FREE Full text] [doi: 10.1371/journal.ppat.1010260] [Medline: 35176090]
8. Barnes CO, Jette CA, Abernathy ME, Dam KA, Esswein SR, Gristick HB, et al. SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. Nature 2020 Dec 12;588(7839):682-687 [FREE Full text] [doi: 10.1038/s41586-020-2852-1] [Medline: 33045718]
9. Piccoli L, Park YJ, Tortorici MA, Czudnochowski N, Walls AC, Beltramello M, et al. Mapping neutralizing and immunodominant sites on the SARS-CoV-2 spike receptor-binding domain by structure-guided high-resolution serology. Cell 2020 Nov 12;183(4):1024-1042.e21 [FREE Full text] [doi: 10.1016/j.cell.2020.09.037] [Medline: 32991844]
10. Pinto D, Park Y, Beltramello M, Walls AC, Tortorici MA, Bianchi S, et al. Cross-neutralization of SARS-CoV-2 by a human monoclonal SARS-CoV antibody. Nature 2020 Jul 18;583(7815):290-295. [doi: 10.1038/s41586-020-2349-y] [Medline: 32422645]
11. Kumar S, Karuppanan K, Subramaniam G. Omicron (BA.1) and sub-variants (BA.1.1, BA.2, and BA.3) of SARS-CoV-2 spike infectivity and pathogenicity: a comparative sequence and structural-based computational assessment. J Med Virol 2022 Oct;94(10):4780-4791 [FREE Full text] [doi: 10.1002/jmv.27927] [Medline: 35680610]
12. McCallum M, Czudnochowski N, Rosen LE, Zepeda SK, Bowen JE, Walls AC, et al. Structural basis of SARS-CoV-2 Omicron immune evasion and receptor engagement. Science 2022 Feb 25;375(6583):864-868 [FREE Full text] [doi: 10.1126/science.abn8652] [Medline: 35076256]
13. Heo Y. Sotrovimab: first approval. Drugs 2022 Mar 14;82(4):477-484 [FREE Full text] [doi: 10.1007/s40265-022-01690-7] [Medline: 35286623]
14. Ashoor D, Ben Khalaf N, Marzouq M, Jarjanazi H, Chlif S, Fathallah MD. A computational approach to evaluate the combined effect of SARS-CoV-2 RBD mutations and ACE2 receptor genetic variants on infectivity: the COVID-19 host-pathogen nexus. Front Cell Infect Microbiol 2021 Aug 9;11:707194 [FREE Full text] [doi: 10.3389/fcimb.2021.707194] [Medline: 34434902]
15. Zhao Z, Zhou J, Tian M, Huang M, Liu S, Xie Y, et al. Omicron SARS-CoV-2 mutations stabilize spike up-RBD conformation and lead to a non-RBM-binding monoclonal antibody escape. Nat Commun 2022 Aug 24;13(1):4958 [FREE Full text] [doi: 10.1038/s41467-022-32665-7] [Medline: 36002453]

16. Lv Z, Deng Y, Ye Q, Cao L, Sun C, Fan C, et al. Structural basis for neutralization of SARS-CoV-2 and SARS-CoV by a potent therapeutic antibody. Science 2020 Sep 18;369(6510):1505-1509 [FREE Full text] [doi: 10.1126/science.abc5881] [Medline: 32703908]
17. Gupta A, Gonzalez-Rojas Y, Juarez E, Crespo Casal M, Moya J, Falci DR, et al. Early treatment for Covid-19 with SARS-CoV-2 neutralizing antibody sotrovimab. N Engl J Med 2021 Nov 18;385(21):1941-1950. [doi: 10.1056/nejmoa2107934]
18. Gupta A, Gonzalez-Rojas Y, Juarez E, Crespo Casal M, Moya J, Rodrigues Falci D, COMET-ICE Investigators. Effect of sotrovimab on hospitalization or death among high-risk patients with mild to moderate COVID-19: a randomized clinical trial. JAMA 2022 Apr 05;327(13):1236-1246 [FREE Full text] [doi: 10.1001/jama.2022.2832] [Medline: 35285853]
19. Ashoor D, Marzouq M, Trabelsi K, Chlif S, Abotalib N, Khalaf NB, et al. How concerning is a SARS-CoV-2 variant of concern? Computational predictions and the variants labeling system. Front Cell Infect Microbiol 2022 Aug 10;12:868205 [FREE Full text] [doi: 10.3389/fcimb.2022.868205] [Medline: 36034694]
20. Protein Data Bank. RCSB. URL: https://www.rcsb.org/ [accessed 2024-09-17]
21. Hodcroft E. CoVariants. CoVariants. URL: https://covariants.org/ [accessed 2024-09-17]
22. SARS-CoV-2 Variants. Stanford University: Coronavirus Antiviral & Resistance Database. URL: https://covdb.stanford.edu/variants/omicron_ba_1_3/ [accessed 2024-09-17]
23. 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]
24. DeLano W. The PyMOL Molecular Graphics System. PyMOL. 2002. URL: http://www.pymol.org/ [accessed 2024-09-17]
25. Song W, Gui M, Wang X, Xiang Y. Cryo-EM structure of the SARS coronavirus spike glycoprotein in complex with its host cell receptor ACE2. PLoS Pathog 2018 Aug 13;14(8):e1007236 [FREE Full text] [doi: 10.1371/journal.ppat.1007236] [Medline: 30102747]
26. Dejnirattisai W, Zhou D, Ginn HM, Duyvesteyn HME, Supasa P, Case JB, et al. The antigenic anatomy of SARS-CoV-2 receptor binding domain. Cell 2021 Apr 15;184(8):2183-2200.e22 [FREE Full text] [doi: 10.1016/j.cell.2021.02.032] [Medline: 33756110]
27. Mannar D, Saville JW, Sun Z, Zhu X, Marti MM, Srivastava SS, 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]
28. Wang Y, Liu C, Zhang C, Wang Y, Hong Q, Xu S, et al. Structural basis for SARS-CoV-2 Delta variant recognition of ACE2 receptor and broadly neutralizing antibodies. Nat Commun 2022 Feb 15;13(1):871 [FREE Full text] [doi: 10.1038/s41467-022-28528-w] [Medline: 35169135]
29. Saville JW, Mannar D, Zhu X, Srivastava SS, Berezuk AM, Demers J, et al. Structural and biochemical rationale for enhanced spike protein fitness in delta and kappa SARS-CoV-2 variants. Nat Commun 2022 Feb 08;13(1):742 [FREE Full text] [doi: 10.1038/s41467-022-28324-6] [Medline: 35136050]
30. The GROMOS Software for (Bio)Molecular Simulation. Volume 1: About the GROMOS package: Overview. The GROMOS Software for (Bio)Molecular Simulation. Volume 1: About the GROMOS package: Overview. 2023. URL: https://www.gromos.net/gromos11_pdf_manuals/vol1.pdf [accessed 2024-09-17]
31. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 1997 Dec 14;18(15):2714-2723. [doi: 10.1002/elps.1150181505] [Medline: 9504803]
32. Laskowski RA, Swindells MB. LigPlot+: multiple ligand-protein interaction diagrams for drug discovery. J Chem Inf Model 2011 Oct 24;51(10):2778-2786. [doi: 10.1021/ci200227u] [Medline: 21919503]
33. Myung Y, Pires DEV, Ascher DB. CSM-AB: graph-based antibody-antigen binding affinity prediction and docking scoring function. Bioinformatics 2022 Jan 27;38(4):1141-1143. [doi: 10.1093/bioinformatics/btab762] [Medline: 34734992]
34. Manjunath R, Gaonkar SL, Saleh EAM, Husain K. A comprehensive review on Covid-19 Omicron (B.1.1.529) variant. Saudi J Biol Sci 2022 Sep;29(9):103372 [FREE Full text] [doi: 10.1016/j.sjbs.2022.103372] [Medline: 35855306]
35. Liu X, Xiong J, Sun Z, Hu J, Thilakavathy K, Chen M, et al. Omicron: a chimera of two early SARS-CoV-2 lineages. Signal Transduct Target Ther 2022 Mar 17;7(1):90 [FREE Full text] [doi: 10.1038/s41392-022-00949-5] [Medline: 35301279]
36. Martins M, do Nascimento GM, Nooruzzaman M, Yuan F, Chen C, Caserta LC, et al. The omicron variant BA.1.1 presents a lower pathogenicity than B.1 D614G and delta variants in a feline model of SARS-CoV-2 infection. J Virol 2022 Sep 14;96(17):e0096122 [FREE Full text] [doi: 10.1128/jvi.00961-22] [Medline: 36000850]
37. Yue C, Song W, Wang L, Jian F, Chen X, Gao F, et al. Enhanced transmissibility of XBB.1.5 is contributed by both strong ACE2 binding and antibody evasion. bioRxiv Preprint posted online January 5, 2023. [doi: 10.1101/2023.01.03.522427]
38. Neher R. Variant report 2022-12-22. GitHub. URL: https://github.com/neherlab/SARS-CoV-2_variant-reports/blob/d2d531c6deb12e52e5a6fde9af25f2cce023302b/reports/variant_report_2022-12-22.md [accessed 2024-09-17]
39. Qu P, Evans JP, Faraone JN, Zheng Y, Carlin C, Anghelina M, et al. Enhanced neutralization resistance of SARS-CoV-2 Omicron subvariants BQ.1, BQ.1.1, BA.4.6, BF.7, and BA.2.75.2. Cell Host Microbe 2023 Jan 11;31(1):9-17.e3 [FREE Full text] [doi: 10.1016/j.chom.2022.11.012] [Medline: 36476380]

40. Magnus CL, Hiergeist A, Schuster P, Rohrhofer A, Medenbach J, Gessner A, et al. Targeted escape of SARS-CoV-2 from monoclonal antibody S309, the precursor of sotrovimab. Front Immunol 2022 Aug 24;13:966236 [FREE Full text] [doi: 10.3389/fimmu.2022.966236] [Medline: 36090991]
41. Cao Y, Wang J, Jian F, Xiao T, Song W, Yisimayi A, et al. Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. Nature 2022 Feb 23;602(7898):657-663. [doi: 10.1038/s41586-021-04385-3] [Medline: 35016194]
42. Benkert P, Biasini M, Schwede T. Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics 2011 Feb 01;27(3):343-350 [FREE Full text] [doi: 10.1093/bioinformatics/btq662] [Medline: 21134891]
43. Studer G, Rempfer C, Waterhouse AM, Gumienny R, Haas J, Schwede T. QMEANDisCo-distance constraints applied on model quality estimation. Bioinformatics 2020 Mar 01;36(6):1765-1771 [FREE Full text] [doi: 10.1093/bioinformatics/btz828] [Medline: 31697312]
44. Asif A, Ilyas I, Abdullah M, Sarfraz S, Mustafa M, Mahmood A. The comparison of mutational progression in SARS-CoV-2: a short updated overview. JMP 2022 Oct 06;3(4):201-218. [doi: 10.3390/jmp3040018]
45. Huang M, Wu L, Zheng A, Xie Y, He Q, Rong X, et al. Atlas of currently available human neutralizing antibodies against SARS-CoV-2 and escape by Omicron sub-variants BA.1/BA.1.1/BA.2/BA.3. Immunity 2022 Aug 09;55(8):1501-1514.e3 [FREE Full text] [doi: 10.1016/j.immuni.2022.06.005] [Medline: 35777362]
46. van de Veerdonk FL, Giamarellos-Bourboulis E, Pickkers P, Derde L, Leavis H, van Crevel R, et al. A guide to immunotherapy for COVID-19. Nat Med 2022 Jan 21;28(1):39-50. [doi: 10.1038/s41591-021-01643-9] [Medline: 35064248]
47. Cheng ZJ, Li B, Zhan Z, Zhao Z, Xue M, Zheng P, et al. Clinical application of antibody immunity against SARS-CoV-2: comprehensive review on immunoassay and immunotherapy. Clin Rev Allergy Immunol 2023 Feb 15;64(1):17-32 [FREE Full text] [doi: 10.1007/s12016-021-08912-y] [Medline: 35031959]
48. Aggarwal NR, Beaty LE, Bennett TD, Carlson NE, Mayer DA, Molina KC, et al. Change in effectiveness of sotrovimab for preventing hospitalization and mortality for at-risk COVID-19 outpatients during an Omicron BA.1 and BA.1.1-predominant phase. Int J Infect Dis 2023 Mar;128:310-317 [FREE Full text] [doi: 10.1016/j.ijid.2022.10.002] [Medline: 36229005]
49. Arora P, Kempf A, Nehlmeier I, Schulz SR, Cossmann A, Stankov MV, et al. Augmented neutralisation resistance of emerging omicron subvariants BA.2.12.1, BA.4, and BA.5. Lancet Infect Dis 2022 Aug;22(8):1117-1118. [doi: 10.1016/s1473-3099(22)00422-4]
50. Imai M, Ito M, Kiso M, Yamayoshi S, Uraki R, Fukushi S, et al. Efficacy of antiviral agents against omicron subvariants BQ.1.1 and XBB. N Engl J Med 2023 Jan 05;388(1):89-91. [doi: 10.1056/nejmc2214302]
51. Cox M, Peacock TP, Harvey WT, Hughes J, Wright DW, COVID-19 Genomics UK (COG-UK) Consortium, et al. SARS-CoV-2 variant evasion of monoclonal antibodies based on in vitro studies. Nat Rev Microbiol 2023 Feb 28;21(2):112-124 [FREE Full text] [doi: 10.1038/s41579-022-00809-7] [Medline: 36307535]
52. COVID-19 weekly epidemiological update, edition 119, 23 November 2022. World Health Organization. 2022. URL: https://iris.who.int/handle/10665/364724 [accessed 2024-09-17]
53. Willett BJ, Grove J, MacLean OA, Wilkie C, De Lorenzo G, Furnon W, PITCH Consortium, COVID-19 Genomics UK (COG-UK) Consortium, et al. SARS-CoV-2 Omicron is an immune escape variant with an altered cell entry pathway. Nat Microbiol 2022 Aug;7(8):1161-1179. [doi: 10.1038/s41564-022-01143-7] [Medline: 35798890]
54. Mohapatra RK, Mahal A, Kutikuppala LS, Pal M, Kandi V, Sarangi AK, et al. Renewed global threat by the novel SARS-CoV-2 variants 'XBB, BF.7, BQ.1, BA.2.75, BA.4.6': a discussion. Front Virol 2022 Dec 23;2:104. [doi: 10.3389/fviro.2022.1077155]

Abbreviations

ΔG: Gibbs free energy
ACE2: angiotensin-converting enzyme 2
Fv: variable domain
NAb: neutralizing antibody
NCBI: National Center for Biotechnology Information
RBD: receptor-binding domain
S309: sotrovimab
SARS: severe acute respiratory syndrome
SARS-CoV: severe acute respiratory syndrome coronavirus

Edited by H Yan; submitted 04.03.24; peer-reviewed by A Hosny, V Nagesh; comments to author 02.04.24; revised version received 07.04.24; accepted 19.04.24; published 10.10.24.

Please cite as:
Ashoor D, Marzouq M, Fathallah MD
Comparison of the Neutralization Power of Sotrovimab Against SARS-CoV-2 Variants: Development of a Rapid Computational Method
JMIR Bioinform Biotech 2024;5:e58018
URL: https://bioinform.jmir.org/2024/1/e58018
doi:10.2196/58018
PMID:39388246

Original Paper

Machine Learning Models for Prediction of Maternal Hemorrhage and Transfusion: Model Development Study

Homa Khorrami Ahmadzia[1,2*], MD, MPH; Alexa C Dzienny[3], BS; Mike Bopf[4*], MS; Jaclyn M Phillips[1], MD; Jerome Jeffrey Federspiel[5*], MD, PhD; Richard Amdur[6], PhD; Madeline Murguia Rice[7], PhD; Laritza Rodriguez[4], MD, PhD

[1]Division of Maternal-Fetal Medicine, Department of Obstetrics and Gynecology, George Washington University, Washington, DC, United States

[2]Division of Maternal-Fetal Medicine, Department of Obstetrics and Gynecology, Inova Health System, Falls Church, VA, United States

[3]The George Washington University School of Medicine and Health Sciences,, Washington DC, DC, United States

[4]Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, Bethesda, MD, United States

[5]Division of Maternal-Fetal Medicine, Department of Obstetrics and Gynecology, Duke University, Durham, NC, United States

[6]Medical Faculty Associates, The George Washington University School of Medicine and Health Sciences, Washington, DC, United States

[7]George Washington University Biostatistics Center, Washington, DC, United States

[*]these authors contributed equally

Corresponding Author:
Homa Khorrami Ahmadzia, MD, MPH
Division of Maternal-Fetal Medicine
Department of Obstetrics and Gynecology
Inova Health System
3300 Gallows Road
Falls Church, VA, 22042
United States
Phone: 1 571 472 0920
Email: homa.ahmadzia@inova.org

Abstract

Background: Current postpartum hemorrhage (PPH) risk stratification is based on traditional statistical models or expert opinion. Machine learning could optimize PPH prediction by allowing for more complex modeling.

Objective: We sought to improve PPH prediction and compare machine learning and traditional statistical methods.

Methods: We developed models using the Consortium for Safe Labor data set (2002-2008) from 12 US hospitals. The primary outcome was a transfusion of blood products or PPH (estimated blood loss of ≥1000 mL). The secondary outcome was a transfusion of any blood product. Fifty antepartum and intrapartum characteristics and hospital characteristics were included. Logistic regression, support vector machines, multilayer perceptron, random forest, and gradient boosting (GB) were used to generate prediction models. The area under the receiver operating characteristic curve (ROC-AUC) and area under the precision/recall curve (PR-AUC) were used to compare performance.

Results: Among 228,438 births, 5760 (3.1%) women had a postpartum hemorrhage, 5170 (2.8%) had a transfusion, and 10,344 (5.6%) met the criteria for the transfusion-PPH composite. Models predicting the transfusion-PPH composite using antepartum and intrapartum features had the best positive predictive values, with the GB machine learning model performing best overall (ROC-AUC=0.833, 95% CI 0.828-0.838; PR-AUC=0.210, 95% CI 0.201-0.220). The most predictive features in the GB model predicting the transfusion-PPH composite were the mode of delivery, oxytocin incremental dose for labor (mU/minute), intrapartum tocolytic use, presence of anesthesia nurse, and hospital type.

Conclusions: Machine learning offers higher discriminability than logistic regression in predicting PPH. The Consortium for Safe Labor data set may not be optimal for analyzing risk due to strong subgroup effects, which decreases accuracy and limits generalizability.

(JMIR Bioinform Biotech 2024;5:e52059) doi:10.2196/52059

KEYWORDS

postpartum hemorrhage; machine learning; prediction; maternal; predict; predictive; bleeding; hemorrhage; hemorrhaging; birth; postnatal; blood; transfusion; antepartum; obstetric; obstetrics; women's health; gynecology; gynecological

Introduction

Maternal morbidity and mortality have been regarded as a reflection of health care quality nationwide. Among lower-income countries, postpartum hemorrhage (PPH) is typically the most common cause of maternal mortality and remains among the top causes in higher-income countries. In the United States, hemorrhage accounted for 11.0% of deaths between 2011 and 2016 [1-4]. To address maternal hemorrhage, maternal hemorrhage protocols have been implemented, which incorporate prospective PPH risk assessment to tailor PPH prophylactic and management approaches for patients' individual risk profiles. However, these protocols are often based on observational studies that approximated the strength of associations with hemorrhage via logistic regression (LR) models and combined the results of multiple studies together in a linear fashion [5-7]. However, "standard" LR assumes that (1) there is a linear relationship between predictors and the log odds of outcomes and (2) there are independent relationships between predictors. Additionally, LR and related models often perform poorly with large numbers of included variables [8,9]. Consequently, current risk stratification models fail to accurately ascertain pregnant patients' risk of hemorrhage [10]. Studies attempting to validate existing LR and related models have instead identified gaps in the efficacy of these models, as the majority of patients with PPH and transfusions were stratified in low or moderate risk groups [11,12].

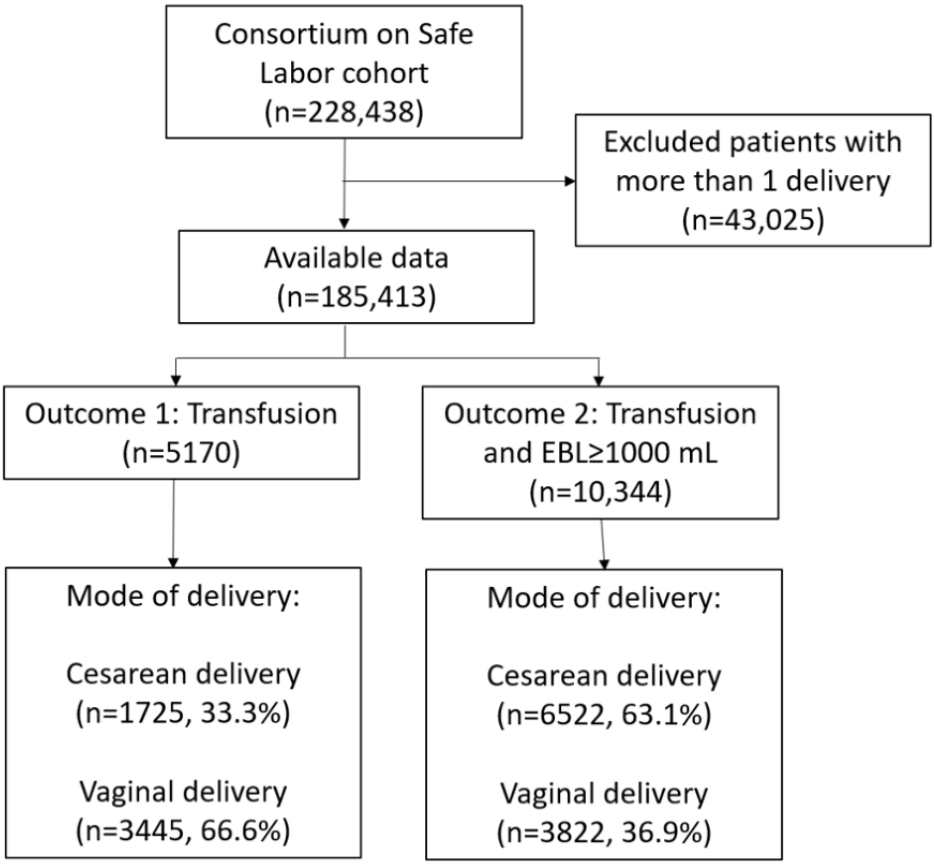
Machine learning offers an advantage to current risk assessment methods through its ability to create a robust model based on larger numbers of predictors, with nonlinear relationships and interactions between variables included in analyses [13]. Our objective in this analysis was to create a validated prediction model using machine learning for postpartum hemorrhage and transfusion to optimize risk-based triage and inform policy makers and stakeholders who aim to further reduce maternal morbidity and mortality associated with hemorrhage.

Methods

Data Collection

Data for this analysis were extracted from the Consortium for Safe Labor (CSL) data set created by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD). It includes antepartum, intrapartum, and postpartum medical histories of 224,438 women from 12 hospitals in the United States (Figure 1). Variables in this data set include maternal demographics, reproductive history, medical history, prenatal history of current pregnancy, labor admission assessment, labor progression, labor and delivery summary, maternal postpartum condition, and newborn information. For this database, data were extracted retrospectively from existing records for deliveries most recently occurring at each site. Data were extracted electronically using a method suitable to each hospital's unique data systems. Data transfer and integrity were managed by a data coordinating center that created a central database. The data were deidentified and are available for research under request from the NICHD. Women with only 1 recorded pregnancy in the data set were included for data analysis; if women had more than 1 pregnancy during the study period, only the first one was used in the analysis. We selected maternal, fetal, and pregnancy variables as candidates to build the prediction model for transfusion risk.

Figure 1. Flowchart of inclusion of women with transfusion or postpartum hemorrhage (or both).



Missing Data

Machine learning methods are known to generate errors in the presence of missing values [14]. To avoid this, we imputed values as follows: categorical variables with missing and unknown values were assigned to an "unknown" category; continuous variables with missing and unknown values were coded to the median value. Continuous variables for maternal age and BMI were coded into ordinal categories (age of <20, between ≥20 and <40, between ≥40 and <45, and ≥45 years; BMI of ≤20, between >20 and ≤40, between >40 and ≤50, and >50 kg/m^2). Imputing estimated blood loss (EBL) as the median value (350 mL) meant that missing values were assumed to be <1000 mL.

Feature Selection

We used the Cramér V index of nominal association for variable selection [15]. Features were classified into antepartum and intrapartum variables. Two different prediction models were constructed: (1) an antenatal-only model intended to be used in the clinic setting to inform appropriate patient referral and (2) an intrapartum model that included both antepartum and intrapartum characteristics. Individual antepartum and intrapartum maternal variables included for model development are shown in the Multimedia Appendix 1.

Outcomes

Separate models were constructed to predict 2 target outcomes. The primary outcome was a composite including all patients who received a transfusion of any blood product or had a PPH defined by documented blood loss of ≥1000 mL during or after delivery. Our secondary outcome was all patients who received transfusion of any blood product. Both blood loss of ≥1000 mL and blood transfusion are clinically significant metrics in obstetric care. Transfusion alone represents patients who are at risk for high maternal morbidity and mortality and is a clinically important metric to evaluate in isolation; hence, it was evaluated independently in a model as a secondary outcome.

Data Analysis

For each of the 4 combinations of predictors and outcomes (for predictors, antepartum vs antepartum and intrapartum; for outcomes, transfusion and blood loss greater than a liter versus transfusion alone), the data were split so that 70% of the observations were used for training and 30% were used for testing, with both sets having the same outcome rate. We applied a number of methods, including LR, support vector machines (SVMs), multilayer perceptron (MLP), random forest (RF), and gradient boosting (GB), as well as deep learning algorithms including TensorFlow imbalanced (TFIM) and learned embedding (Emb). Hyperparameters were tuned for each algorithm using a customized grid search technique. The model performance for each combination of outcome and algorithm was measured using the Matthews correlation coefficient (MCC), area under the receiver operating characteristic curve (ROC-AUC), area under the precision/recall curve (PR-AUC), and modified F-score skewed toward recall (F2). A modified F2 score was chosen to minimize false negatives and thus maximize the identification of patients at high risk for bleeding and transfusion. Existing LR models and risk classification

schemes perform poorly, and the majority of patients with hemorrhage or transfusion are misclassified as low risk. Misclassification of a "high risk" patient as "low risk" may have important clinical implications. Additionally, interventions can be implemented to minimize risk and enhance patient safety (eg, type and cross, multiple intravenous access sites, provider awareness, medications, etc). Models will then be evaluated for those with the highest positive predictive value (PPV) given these parameters. A model with the highest PPV will be clinically useful to identify a high-risk patient population without increasing the clinical burden on the hospital system or patient with the abovementioned interventions. Algorithms were processed and results were analyzed using Python (version 3.6; Python Software Foundation), Pandas (version 1.2; The Pandas Development Team), scikit-learn (version 0.24; scikit-learn Developers), and TensorFlow (version 2.2; Python Software Foundation).

The primary study objective was to identify the strongest set of pre- and intraoperative predictors of hemorrhage or transfusion and the strongest modeling technique. Secondary objectives included determining the level of agreement between metrics for model evaluation and the extent to which any technique produced results that are clinically useful. Given the heterogeneity of this data set derived from multiple institutions, a site-specific sensitivity analysis was performed.

Ethical Considerations

This analysis was exempt from review by the George Washington University's institutional review board (NCR202746).

Results

Of 228,438 births included in the CSL cohort, we included 185,413 patients (Figure 1), having excluded patients with more than 1 delivery (n=43,025). Maternal age ranged from 11 to 58 (median 27) years; 32% (n=60,193) of the participants were publicly insured, 49% (n=90,466) were white non-Hispanic, 22% (n=41,780) were Black, and 17% (n=32,727) were Hispanic. Of the 185,413 women included in the analysis, 71% (n=131,130) had a vaginal delivery, and 29% (n=54,283) had a cesarean delivery. In total, 5170 (3%) women experienced the primary outcome of transfusion of any blood product, 5760 (3.11%) had a PPH defined by an estimated blood loss of ≥1000 mL, and 10,344 (6%) experienced the secondary composite outcome of transfusion or estimated blood loss of loss of ≥1000 mL. Additional demographic data are summarized in Multimedia Appendix 2.

After building the models in an iterative process, their performance in predicting both the primary and secondary outcomes was compared using a variety of metrics. The metrics ROC-AUC, PR-AUC, MCC, and F2, as well as sensitivity and specificity at a probability cut point of 50% are shown in Tables 1 and 2.

Table 1. Performance of machine learning and statistical models based on antepartum and intrapartum maternal variables at predicting transfusion or postpartum hemorrhage (or both). Primary outcome: blood transfusion or blood loss of ≥1 L.

Algorithm	True positives[a], n	True negatives[a], n	False positives[a], n	False negatives[a], n	Positive predictive value	Sensitivity	Specificity	ROC-AUC[b]	PR-AUC[c]	MCC[d]	F2[e]
GB[f]	50	6	318	626	0.135	0.889	0.663	0.833	0.210	0.260	0.419
RF[g]	50	6	339	605	0.138	0.857	0.641	0.830	0.204	0.261	0.409
Emb[h]	46	10	296	649	0.134	0.821	0.687	0.813	0.181	0.246	0.406
MLP[i]	49	7	335	609	0.127	0.875	0.645	0.808	0.149	0.245	0.402
TFIM[j]	48	8	323	619	0.129	0.861	0.655	0.822	0.194	0.245	0.403
SVM[k]	49	6	349	595	0.124	0.886	0.630	0.804	0.159	0.242	0.397
LR[l]	46	10	314	631	0.129	0.830	0.668	0.813	0.177	0.238	0.393

[a]Values are normalized per 1000, so they are easier to compare across different models; the actual N value is 55,624.

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

[c]PR-AUC: area under the precision-recall curve.

[d]MCC: Matthews correlation coefficient.

[e]F2: modified F-score skewed toward recall.

[f]GB: gradient boosting.

[g]RF: random forest.

[h]Emb: learned embedding.

[i]MLP: multilayer perceptron.

[j]TFIM: TensorFlow imbalanced.

[k]SVM: support vector machine.

[l]LR: logistic regression.

Table 2. Performance of machine learning and statistical models based on antepartum and intrapartum maternal variables in predicting transfusion or postpartum hemorrhage (or both). Secondary outcome: blood transfusion.

Algorithm	True positives[a], n	True negatives[a], n	False positives[a], n	False negatives[a], n	Positive predictive value	Sensitivity	Specificity	ROC-AUC[b]	PR-AUC[c]	MCC[d]	F2[e]
GB[f]	24	4	235	737	0.093	0.866	0.758	0.860	0.111	0.234	0.325
RF[g]	25	3	251	721	0.090	0.887	0.742	0.862	0.107	0.232	0.319
Emb[h]	22	6	223	750	0.090	0.789	0.771	0.837	0.096	0.215	0.309
MLP[i]	24	4	237	735	0.091	0.849	0.756	0.845	0.095	0.227	0.318
TFIM[j]	24	4	240	732	0.091	0.859	0.753	0.855	0.111	0.229	0.319
SVM[k]	24	4	244	728	0.091	0.871	0.749	0.852	0.116	0.230	0.320
LR[l]	24	3	250	722	0.089	0.876	0.743	0.853	0.111	0.228	0.317

[a]Values are normalized per 1000, so they are easier to compare across different models; the actual N value is 55,624.

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

[c]PR-AUC: area under the precision-recall curve.

[d]MCC: Matthews correlation coefficient.

[e]F2: modified F-score skewed toward recall.

[f]GB: gradient boosting.

[g]RF: random forest.

[h]Emb: learned embedding.

[i]MLP: multilayer perceptron.

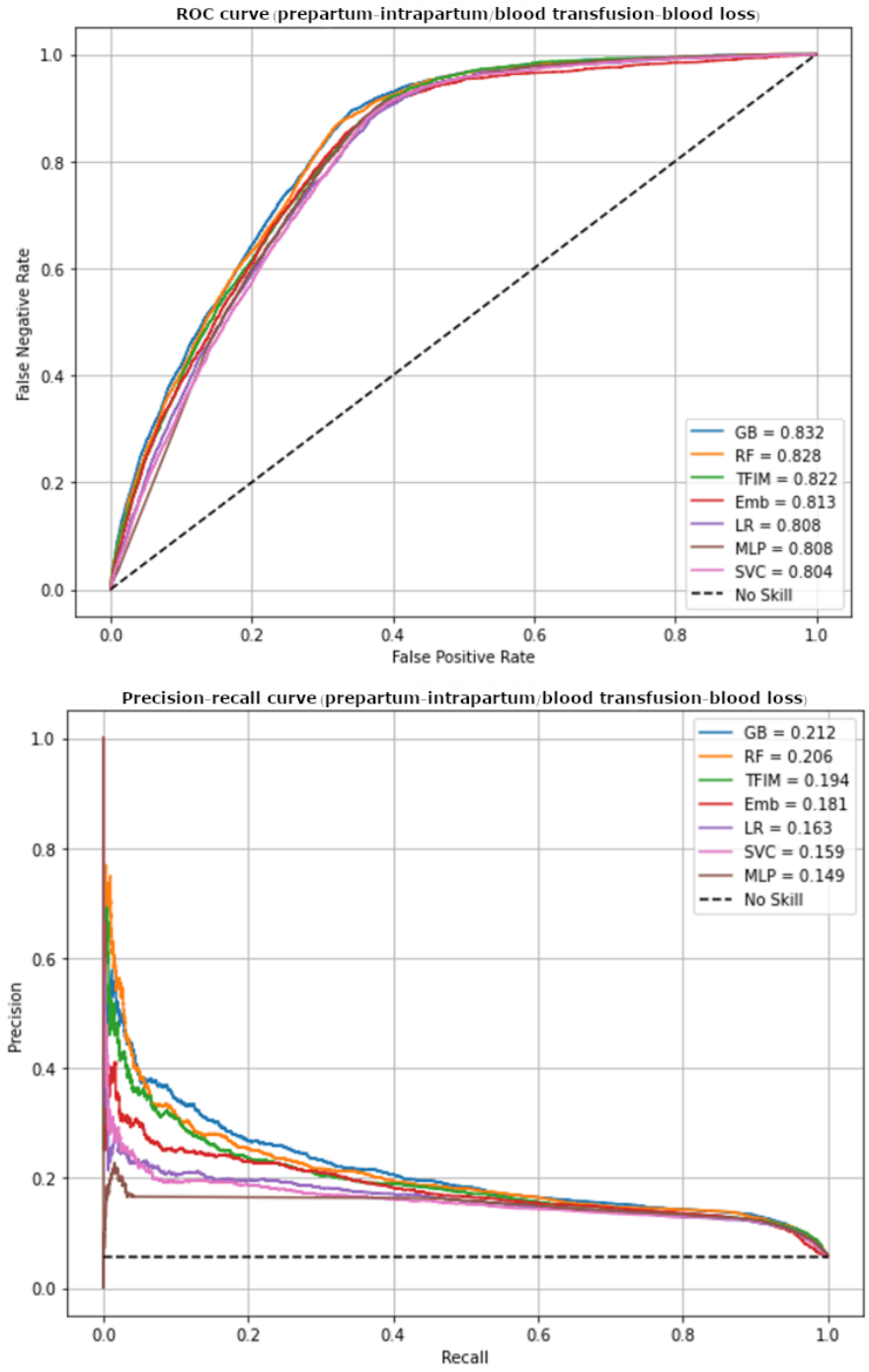
[j]TFIM: TensorFlow imbalanced.

[k]SVM: support vector machine.

[l]LR: logistic regression.

For both the primary and secondary outcomes, models developed using antepartum and intrapartum maternal variables (see Multimedia Appendix 1 for a list of variables) to predict the primary outcome performed better with higher PPVs than those solely using antepartum maternal variables (Multimedia Appendices 3 and 4). For the primary composite outcome, the machine learning technique GB using intrapartum maternal variables had the highest PPV (PR-AUC=0.21, 95% CI 0.20-0.22; ROC-AUC=0.83, 95% CI 0.828-0.838; Figure 2). For the secondary outcome of transfusion alone, there was little difference in model performance when comparing several performance metrics.

Figure 2. Receiver operating characteristic and precision/recall curves for different models using intrapartum maternal variables predicting transfusion or postpartum hemorrhage.



The remainder of our results focus on the model with the highest PPV: the intrapartum model (containing both antepartum and intrapartum variables) evaluating our primary outcome of a composite of blood loss of more than 1000 mL or transfusion. Both RF and GB had significantly higher PPVs for predicting the composite transfusion or PPH when compared with LR (PR-AUC=0.18, 95% CI 0.17-0.19; ROC-AUC=0.81, 95% CI 0.808-0.818).

Figure 3 reveals the calibration curves for the models constructed with intrapartum maternal variables and predicting the transfusion-PPH composite. Calibration curves portray the predicted PPH risk versus the observed PPH rate across a range of predicted PPH values. There was better agreement between the models with a lower fraction of positives, and none of the models were able to reach the standard curve—for all models, the predicted PPH risk overestimated the observed PPH rate across the range of predicted values.

Figure 3. Calibration curves for models using intrapartum maternal variables to predict transfusion or postpartum hemorrhage (or both). Emb: learned embedding; GB: gradient boosting; LR: logistic regression; MLP: multilayer perceptron; RF: random forest; SVC: support vector machine; TFIM: TensorFlow imbalanced.

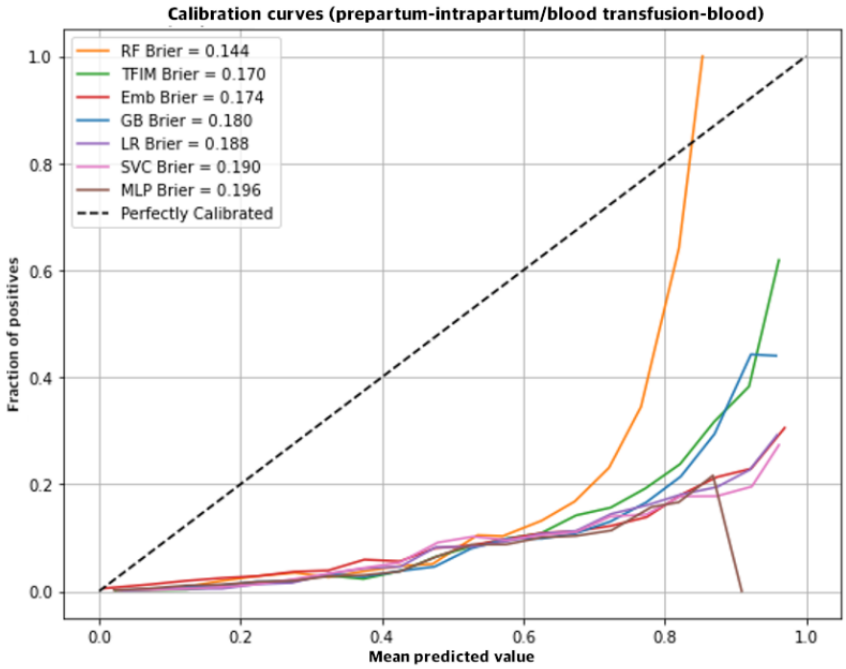
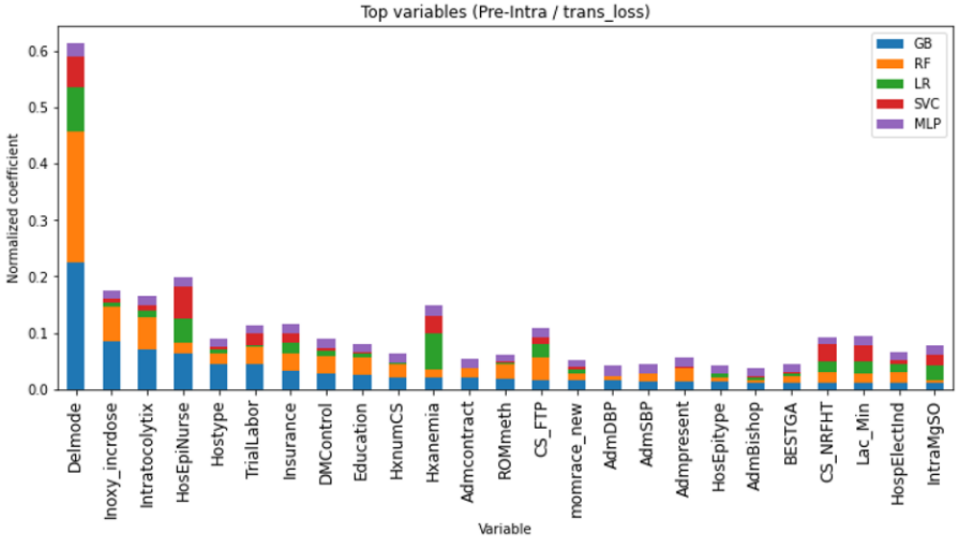


Figure 4 displays the top 25 predictive variables included for model development using antepartum and intrapartum features for the prediction of the transfusion-PPH composite. As the machine learning GB model was the best performing model overall, the variables in Figure 4 are in order of variable importance within the GB model. The top 10 variables from most predictive rate to least predictive rate for intrapartum prediction of the transfusion-PPH composite using the GB model are mode of delivery, oxytocin incremental dose for labor (mU/minute), intrapartum tocolytic use, use of anesthesia nurse, hospital type, a trial of labor after prior cesarean delivery, insurance, most serious diabetes control, education, and history of prior cesarean sections. The results of the models for antepartum-only models are listed in Multimedia Appendix 3. The ROC-AUC and PR-AUC did not perform as well for the models using antepartum-only variables, though this was less obvious for the models predicting transfusion only (Multimedia Appendix 4). Of note, upon further sensitivity analysis, we also determined that some of the top variables in the model were site-specific (ie, oxytocin incremental dose for labor, intrapartum tocolytic use, use of anesthesia nurse, and hospital type) for transfusion outcomes specifically (data not included).

Figure 4. Top 25 predictors based on each model using intrapartum maternal factors predicting transfusion or postpartum hemorrhage (or both). GB: gradient boosting; LR: logistic regression; MLP: multilayer perceptron; RF: random forest; SVC: support vector machine.



Discussion

Principal Findings

In this study, LR and machine learning techniques were analyzed and compared to develop prediction models for PPH and transfusions. We found that the machine learning techniques, particularly GB, performed best to predict PPH when PPH was defined as blood transfusion or blood loss of greater than 1 L. However, all prediction models had difficulties with calibration when predicting the rare outcome of transfusion alone.

Clinical Implications

Risk assessment for PPH has been shown in a pre-post study to reduce rates of blood transfusion and PPH [16]. However, the risk stratification approaches most commonly used for PPH in the United States were developed and implemented on the basis of expert opinion, and subsequent validation studies have revealed the limitations of these tools [17,18]. Validation studies using the California Maternal Quality Care Collaborative (CMQCC) risk assessment tool found that while the tool generated populations with different rates of hemorrhage among those stratified to low, medium, and high-risk groups, the rate of PPH among women stratified in the high-risk group for PPH was only 22% [19]. Others have found that the AUC-ROC for the CMQCC and Association of Women's Health, Obstetric and Neonatal Nurses' (AWHONN's) tools for predicting severe PPH, defined by transfusion of at least 4 units packed red blood cells during postpartum period, were relatively modest at 0.77 and 0.69, respectively [20]. Furthermore, parameters that are included in PPH risk models based on univariate association with PPH risk may not be independent predictors when incorporated into multivariate models [20]. For these reasons, improvements in PPH risk models are a promising target for improving PPH care.

A previously published risk assessment for PPH using the CSL data set demonstrated exceptional model performance, but model performance was drastically lower in an external validation cohort [21,22]. This study augments the findings of these prior studies via incorporation of antepartum and intrapartum risk factors. Nonetheless, additional work is needed before such a model can be implemented in clinical practice. In particular, it will be important to develop prediction models that are implementable either through straightforward bedside data entry or can be automated via real-time data capture from electronic medical records, which are well validated in a variety of hospital settings, and ideally, which are paired with recommended risk-based interventions to reduce hemorrhage risk and mitigate the occurrence of hemorrhage. In our study, among the top predictors were variables that reflect patients' access to care and resources, such as hospital type and insurance. This highlights the possible need for a layered prediction model, which may help stratify patients who may need to be transferred to a tertiary care center with more resources (using an antepartum model focusing on patient factors along with hospital factors to designate risk).

Research Implications

For all the intrapartum methods that we tested for predicting transfusion or hemorrhage, the ROC-AUC values were greater than 0.80, which is often cited as a threshold indicating adequate discrimination. However, this conclusion is misleading because in a situation where incidence of the outcome is low (here, it was ~3% for transfusion or hemorrhage alone), the PPV, also known as "precision," is likely to be quite low. Our precision for the best-performing model was ~13%, meaning that of those predicted to be positive for the outcome, 13% were positive and 87% were negative. This may be satisfactory for clinical uses where preventive interventions have very low cost (in terms of both financial cost and added risk to the patient) but would not be acceptable when the intervention is of higher risk or is more expensive. In this situation, the PR-AUC provided a more realistic measure of model quality. Precision/recall plots show PPV (aka precision) as a function of sensitivity (aka recall); thus, they account for true positives in positive predictions. In contrast, the ROC-AUC emphasizes specificity, which is likely to be very high when true positives are rare [23,24]. The metric with the largest difference between the best and worst-performing models is PR-AUC (0.16 vs 0.21). This metric could be used more frequently in modeling studies when the occurrence of the outcome of interest is ≤6%.

Strengths and Limitations

The strengths of this study include the use of a large, national multicenter data set to develop a data-driven model that can predict PPH using antepartum and intrapartum factors using cutting-edge machine learning techniques. Furthermore, we considered both commonly used end points such as estimated blood loss greater than 1 L and clinically relevant end points such as transfusion; this led us to conclude that due to a less frequent occurrence and transfusion practice, variation made it more challenging to develop a reliable model for transfusion only.

Limitations of the study include the low reported precision of algorithms. Sensitivity is prioritized for prediction, as clinically missing PPH has more consequences than a false positive. Therefore, the algorithms are trained to be biased toward predicting positives resulting in lower false negative rates at the risk of higher false positive rates and decreased precision. As a result, as shown in the calibration plots, the models systematically overstate hemorrhage risk. In this study, the outcomes of interest were either a composite of transfusion or blood loss of ≥1 L or transfusion only. Our PPH definition was based on the American College of Obstetricians and Gynecologists' reVITALize program's definition of PPH as blood loss of ≥1 L or loss of blood with clinical signs of hypovolemia within 24 hours of delivery. This definition deviates from older traditional definitions that defined PPH as ≥500 mL for vaginal delivery and 1000 mL for cesarean delivery [25]. Therefore, clinical care could have been guided by older definitions, as the CSL data set was collected between 2002 and 2008 [21]. However, a strength of our study is the use of EBL rather than a clinical designation of PPH so that we only include patients who were designated to have an EBL above the current threshold for PPH, that is, 1000 mL. Beyond that,

measures of EBL have been shown to be imprecise with low volumes overestimated and high volumes of blood loss underestimated [26]. Furthermore, transfusion was used as a proxy for PPH, and transfusion thresholds vary depending on the institution and provider. In addition, the machine learning algorithms are limited by the variables measured and accurately recorded in the data set.

Conclusions

In conclusion, machine learning and data-driven statistical modeling may offer more objective and discriminative prediction of PPH based on individual antepartum and intrapartum patient features, compared to expert opinion, and may improve upon traditional regression models. This can increase the opportunity for precision medicine and improved clinical care to reduce the burden of PPH as a leading cause of maternal morbidity and mortality.

Acknowledgments

HKA's effort was supported by the National Heart Lung and Blood Institute of the National Institutes of Health (award K23HL141640) and JJF's effort was supported by the National Center for Advancing Translation Sciences (award TL1TR002555). LR's and MB's effort was supported by the Intramural Research Program at the National Institutes of Health, National Library of Medicine.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors would like to acknowledge Dr Christian Macedonia and Dr Chad Grotegut for their insights on the initial model design and Dr Mina Felfeli for helping to submit the manuscript. HKA and JJF were supported by grants (K23HL141640 and TL1TR002555, respectively).

Conflicts of Interest

RA has stock ownership in Abbvie, Bristol Myers Squibb, and Pfizer. This is not related to this study.

Multimedia Appendix 1
All antepartum and intrapartum variables were included for analysis for feature selection.
[DOCX File , 21 KB - bioinform_v5i1e52059_app1.docx]

Multimedia Appendix 2
Overall Patient Characteristics.
[DOCX File , 14 KB - bioinform_v5i1e52059_app2.docx]

Multimedia Appendix 3
Performance of machine learning and statistical models. The model included antepartum maternal features predicting transfusion and/or postpartum hemorrhage. Pre/Trans Loss. Footnote for table: aAlg=algorithm, bNTP=normalized true positive, cNFN=normalized false negative, dNFP=normalized false positive, eNTN=normalized true negative, fROC_AUC (receiver operator curve_area under the curve; 0.5 was considered no better than chance, greater than 0.5 to less than 0.7 poor, 0.7 to less than 0.8 acceptable, 0.8 to less than 0.9 excellent, 0.9 or greater outstanding), gPR_AUC (precision recall_area under the curve), hMCC=Matthews correlation coefficient, iF2= modified F-score skewed towards recall), jGradient boosting, kRandom forests, llearned embedding, mMulti-layer percepton, nTensorflow imbalanced, oSupport vector machines, plogistic regression.
[DOCX File , 14 KB - bioinform_v5i1e52059_app3.docx]

Multimedia Appendix 4
Performance of machine learning and statistical models. The model included antepartum maternal features predicting transfusion of any blood products only. Pre/ Trans_yes Footnote: aAlg=algorithm, bNTP=normalized true positive, cNFN=normalized false negative, dNFP=normalized false positive, eNTN=normalized true negative, fROC_AUC (receiver operator curve_area under the curve; 0.5 was considered no better than chance, greater than 0.5 to less than 0.7 poor, 0.7 to less than 0.8 acceptable, 0.8 to less than 0.9 excellent, 0.9 or greater outstanding), gPR_AUC (precision recall_area under the curve), hMCC=Matthews correlation coefficient, iF2= modified F-score skewed towards recall), jGradient boosting, kRandom forests, llearned embedding, mMulti-layer percepton, nTensorflow imbalanced, oSupport vector machines, plogistic regression.
[DOCX File , 14 KB - bioinform_v5i1e52059_app4.docx]

References

1. MacDorman MF, Declercq E, Cabral H, Morton C. Recent increases in the U.S. maternal mortality rate: disentangling trends from measurement issues. Obstet Gynecol 2016 Sep;128(3):447-455 [FREE Full text] [doi: 10.1097/AOG.0000000000001556] [Medline: 27500333]

2. Neggers YH. Trends in maternal mortality in the United States. Reprod Toxicol 2016 Sep;64:72-76. [doi: 10.1016/j.reprotox.2016.04.001] [Medline: 27063184]
3. Pregnancy Mortality Surveillance System. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/reproductivehealth/maternal-mortality/pregnancy-mortality-surveillance-system.htm [accessed 2021-01-09]
4. Collier AY, Molina RL. Maternal mortality in the United States: updates on trends, causes, and solutions. Neoreviews 2019 Oct;20(10):e561-e574 [FREE Full text] [doi: 10.1542/neo.20-10-e561] [Medline: 31575778]
5. Shields LE, Wiesner S, Fulton J, Pelletreau B. Comprehensive maternal hemorrhage protocols reduce the use of blood products and improve patient safety. Am J Obstet Gynecol 2015 Mar;212(3):272-280. [doi: 10.1016/j.ajog.2014.07.012] [Medline: 25025944]
6. Al-Zirqi I, Vangen S, Forsen L, Stray-Pedersen B. Prevalence and risk factors of severe obstetric haemorrhage. BJOG 2008 Sep;115(10):1265-1272. [doi: 10.1111/j.1471-0528.2008.01859.x] [Medline: 18715412]
7. Bateman BT, Berman MF, Riley LE, Leffert LR. The epidemiology of postpartum hemorrhage in a large, nationwide sample of deliveries. Anesth Analg 2010 May 01;110(5):1368-1373. [doi: 10.1213/ANE.0b013e3181d74898] [Medline: 20237047]
8. Lee H, Yoon SB, Yang S, Kim WH, Ryu H, Jung C, et al. Prediction of acute kidney injury after liver transplantation: machine learning approaches vs. logistic regression model. J Clin Med 2018 Nov 08;7(11) [FREE Full text] [doi: 10.3390/jcm7110428] [Medline: 30413107]
9. Deo RC. Machine learning in medicine. Circulation 2015 Nov 17;132(20):1920-1930 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.115.001593] [Medline: 26572668]
10. Pressly MA, Parker RS, Waters JH, Beck SL, Jeyabalan A, Clermont G. Improvements and limitations in developing multivariate models of hemorrhage and transfusion risk for the obstetric population. Transfusion 2021 Feb;61(2):423-434. [doi: 10.1111/trf.16216] [Medline: 33305364]
11. Ruppel H, Liu VX, Gupta NR, Soltesz L, Escobar GJ. Validation of postpartum hemorrhage admission risk factor stratification in a large obstetrics population. Am J Perinatol 2021 Sep;38(11):1192-1200 [FREE Full text] [doi: 10.1055/s-0040-1712166] [Medline: 32455467]
12. Hussain S, Guarini C, Blosser C, Poole A. Obstetric hemorrhage outcomes by intrapartum risk stratification at a single tertiary care center. Cureus 2019 Dec 24;11(12):e6456 [FREE Full text] [doi: 10.7759/cureus.6456] [Medline: 32025387]
13. Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. N Engl J Med 2016 Sep 29;375(13):1216-1219 [FREE Full text] [doi: 10.1056/NEJMp1606181] [Medline: 27682033]
14. Rosati G. Dealing with Missing Data. Medium. 2019. URL: https://towardsdatascience.com/dealing-with-missing-data-17f8b5827664 [accessed 2023-12-11]
15. Cramér H. Mathematical Methods of Statistics (PMS-9). Princeton, NJ: Princeton University Press; 1999.
16. Ahmadzia HK, Phillips JM, Kleiman R, Gimovsky AC, Bathgate S, Luban NLC, et al. Hemorrhage risk assessment on admission: utility for prediction of maternal morbidity. Am J Perinatol 2021 Sep;38(11):1126-1133 [FREE Full text] [doi: 10.1055/s-0040-1710501] [Medline: 32446252]
17. Bingham D, Melsop K, Main E. CMQCC Obstetric Hemorrhage Hospital Level Implementation Guide. The California Maternal Quality Care Collaborative (CMQCC). 2010. URL: https://www.cmqcc.org/resource/1489/download [accessed 2023-12-11]
18. Postpartum hemorrhage (PPH) risk assessment table 1. The AWHONN postpartum hemorrhage project. URL: https://cdn-links.lww.com/permalink/aog/b/aog_134_6_2019_10_06_kawakita_19-1065_sdc2.pdf [accessed 2021-01-07]
19. Dilla A, Waters J, Yazer M. Clinical validation of risk stratification criteria for peripartum hemorrhage. Obstet Gynecol 2013 Jul;122(1):120-126. [doi: 10.1097/AOG.0b013e3182941c78] [Medline: 23743452]
20. Kawakita T, Mokhtari N, Huang JC, Landy HJ. Evaluation of risk-assessment tools for severe postpartum hemorrhage in women undergoing cesarean delivery. Obstet Gynecol 2019 Dec;134(6):1308-1316. [doi: 10.1097/AOG.0000000000003574] [Medline: 31764744]
21. Venkatesh K, Strauss R, Grotegut C, Heine RP, Chescheir NC, Stringer JSA, et al. Machine learning and statistical models to predict postpartum hemorrhage. Obstet Gynecol 2020 Apr;135(4):935-944 [FREE Full text] [doi: 10.1097/AOG.0000000000003759] [Medline: 32168227]
22. Meyer SR, Carver A, Joo H, Venkatesh KK, Jelovsek JE, Klumpner TT, et al. External validation of postpartum hemorrhage prediction models using electronic health record data. Am J Perinatol 2022 Mar 02. [doi: 10.1055/a-1745-1348] [Medline: 35045573]
23. Leonard S, Kennedy C, Carmichael S, Lyell D, Main E. An expanded obstetric comorbidity scoring system for predicting severe maternal morbidity. Obstet Gynecol 2020 Sep;136(3):440-449 [FREE Full text] [doi: 10.1097/AOG.0000000000004022] [Medline: 32769656]
24. 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]
25. Postpartum Hemorrhage. American College of Obstetricians and Gynecologists. 2017. URL: https://www.acog.org/en/clinical/clinical-guidance/practice [accessed 2021-07-11]

26. Consortium on Safe Labor (CSL) - 3.13 GB. NIH Data and Specimen Hub. 2016. URL: https://dash.nichd.nih.gov/study/2331 [accessed 2021-07-11]

Abbreviations

CSL: Consortium for Safe Labor
EBL: estimated blood loss
Emb: learned embedding
GB: gradient boosting
LR: logistic regression
MCC: Matthews correlation coefficient
MLP: multilayer perceptron
NICHD: Eunice Kennedy Shriver National Institute of Child Health and Human Development
PPH: postpartum hemorrhage
PPV: positive predictive value
PR-AUC: precision/recall area under the curve
RF: random forest
ROC-AUC: receiver operating characteristic area under the curve
SVM: support vector machine
TFIM: TensorFlow imbalanced

Edited by S Hacking; submitted 22.08.23; peer-reviewed by M Wong, Y Zhang; comments to author 27.09.23; revised version received 10.11.23; accepted 03.12.23; published 05.02.24.

Please cite as:
Ahmadzia HK, Dzienny AC, Bopf M, Phillips JM, Federspiel JJ, Amdur R, Rice MM, Rodriguez L
Machine Learning Models for Prediction of Maternal Hemorrhage and Transfusion: Model Development Study
JMIR Bioinform Biotech 2024;5:e52059
URL: https://bioinform.jmir.org/2024/1/e52059
doi:10.2196/52059
PMID:38935950

Original Paper

Deep Learning–Based Identification of Tissue of Origin for Carcinomas of Unknown Primary Using MicroRNA Expression: Algorithm Development and Validation

Ananya Raghu[1*]; Anisha Raghu[1*]; Jillian F Wise[2,3,4], PhD

[1]Quarry Lane School, San Ramon, CA, United States

[2]Department of Biology and Biomedical Sciences, Salve Regina University, Newport, RI, United States

[3]Broad Institute of MIT and Harvard, Cambridge, MA, United States

[4]Pre-College Programs, Tufts University, Medford, MA, United States

[*]these authors contributed equally

Corresponding Author:
Jillian F Wise, PhD
Department of Biology and Biomedical Sciences
Salve Regina University
100 Ochre Point Avenue
Newport, RI, 02840
United States
Phone: 1 401 847 6650 ext 2822
Email: jillian.wise@salve.edu

Abstract

Background: Carcinoma of unknown primary (CUP) is a subset of metastatic cancers in which the primary tissue source of the cancer cells remains unidentified. CUP is the eighth most common malignancy worldwide, accounting for up to 5% of all malignancies. Representing an exceptionally aggressive metastatic cancer, the median survival is approximately 3 to 6 months. The tissue in which cancer arises plays a key role in our understanding of sensitivities to various forms of cell death. Thus, the lack of knowledge on the tissue of origin (TOO) makes it difficult to devise tailored and effective treatments for patients with CUP. Developing quick and clinically implementable methods to identify the TOO of the primary site is crucial in treating patients with CUP. Noncoding RNAs may hold potential for origin identification and provide a robust route to clinical implementation due to their resistance against chemical degradation.

Objective: This study aims to investigate the potential of microRNAs, a subset of noncoding RNAs, as highly accurate biomarkers for detecting the TOO through data-driven, machine learning approaches for metastatic cancers.

Methods: We used microRNA expression data from The Cancer Genome Atlas data set and assessed various machine learning approaches, from simple classifiers to deep learning approaches. As a test of our classifiers, we evaluated the accuracy on a separate set of 194 primary tumor samples from the Sequence Read Archive. We used permutation feature importance to determine the potential microRNA biomarkers and assessed them with principal component analysis and t-distributed stochastic neighbor embedding visualizations.

Results: Our results show that it is possible to design robust classifiers to detect the TOO for metastatic samples on The Cancer Genome Atlas data set, with an accuracy of up to 97% (351/362), which may be used in situations of CUP. Our findings show that deep learning techniques enhance prediction accuracy. We progressed from an initial accuracy prediction of 62.5% (226/362) with decision trees to 93.2% (337/362) with logistic regression, finally achieving 97% (351/362) accuracy using deep learning on metastatic samples. On the Sequence Read Archive validation set, a lower accuracy of 41.2% (77/188) was achieved by the decision tree, while deep learning achieved a higher accuracy of 80.4% (151/188). Notably, our feature importance analysis showed the top 3 most important features for predicting TOO to be microRNA-10b, microRNA-205, and microRNA-196b, which aligns with previous work.

Conclusions: Our findings highlight the potential of using machine learning techniques to devise accurate tests for detecting TOO for CUP. Since microRNAs are carried throughout the body via extracellular vesicles secreted from cells, they may serve as key biomarkers for liquid biopsy due to their presence in blood plasma. Our work serves as a foundation toward developing blood-based cancer detection tests based on the presence of microRNA.

(JMIR Bioinform Biotech 2024;5:e56538) doi:10.2196/56538

KEYWORDS

cancer genomics; machine learning algorithms; deep learning; gene expression; RNA; RNAs; cancer; oncology; tumor; tumors; tissue; tissues; metastatic; microRNA; microRNAs; gene; genes; genomic; genomics; machine learning; algorithm; algorithms; carcinoma; genetics; genome; detection; bioinformatics

Introduction

Carcinoma of unknown primary (CUP) originates when a patient presents at diagnosis with malignant disease across the body; yet, the cancer cells tissue of origin (TOO) remains unidentifiable. Thus, CUP is a unique subset of metastasized cancer representing an advanced stage in which cancer has gained the ability to thrive in new tissue sites and has spread from the primary tumor site. In the United States, an estimated 31,490 people were diagnosed with cases of cancer of unknown TOO in 2008. This accounts for nearly 3%-5% of all cancer cases [1] and given the lack of knowledge on tissue response to current therapeutics the median survival of patients remains only 3-9 months [2]. In many cases of CUP, the primary site is never identified, preventing the use of treatment that can be effective for the true TOO [3,4]. It has been demonstrated that pinpointing the primary site can significantly increase survival rates by enabling precise and targeted treatment [5].

Unfortunately, primary tumor identification poses various challenges. Techniques such as serum tumor markers and imaging tests are used to identify the TOO, although only 30% of these tests are successful. Moreover, some positive findings can be misleading [6] and CUP diagnostic workups are often time-consuming, expensive, and unsuccessful [7]. These difficulties have spurred interest in using genetic expression data, such as microRNA, to identify the TOO.

MicroRNAs belong to a class of noncoding regulatory RNAs, small single-stranded RNA molecules that are between 19 and 25 nucleotides long and are involved in the regulation of gene expression of mRNAs. MicroRNAs hold promise as informative biomarkers for cancer due to their significant involvement in cellular processes such as cell division, apoptosis, proliferation, and oncogenesis [8]. Beyond their intracellular role in gene regulation, microRNAs may be carried throughout the body via extracellular vesicles secreted from cells and have been identified in the blood. Additionally, microRNA, unlike mRNA, is characterized by resistance to extreme temperatures and pH. This makes microRNAs far more stable biomarkers [9,10].

Previous work [11] demonstrates that microRNA expression is more informative in classifying tumor samples by their origin in comparison to mRNA. Specifically, microRNAs are better at classifying poorly differentiated tumors [12]. Moreover, microRNAs have shown great potential for identifying TOO for cancers of unknown primary origin [13]. MicroRNAs have been investigated as prognostic and diagnostic biomarkers extensively in the research community and have even been found to be deregulated in numerous cancers [14].

With the wide availability of large data sets containing gene expression data, computational techniques such as machine learning have emerged as promising tools for improving TOO detection. Machine learning implementations have increased accuracy in predicting cancer and have the potential to improve the diagnosis, prognosis, and therapy selection for patients with cancer [15]. The 3 traditional machine learning models are decision trees, random forests, and logistic regression. Decision trees [16] attempt to partition the training set into subsets that contain samples of only one class, thereby predicting the class of interest. Random forests are ensemble classifiers, combining multiple trees for higher accuracy [17]. In contrast, logistic regression is a predictive algorithm to find a model that can predict categorical output [18]. Deep learning is a subset of machine learning designed to mimic the human brain through the use of artificial neural networks by using many layers and larger data sets. Generally, deep learning techniques are well suited for discovering and recognizing complex patterns in data that traditional machine learning methods can often miss. The increasing incorporation of deep learning in health care along with the availability of highly characterized cancer data sets has further accelerated research into the applications of deep learning in the analysis of the biology of cancer [19].

Given the complexities of diagnosing a TOO from a cancer that has spread throughout the body, previous investigators have applied machine learning methods to determine TOO for metastasized cancers [20,21]. Longstanding techniques of microarrays and polymerase chain reaction have been used for the generation of machine learning models for CUP detection, including support vector machines with 89% accuracy [22] and the k-nearest neighbor algorithm with 82% accuracy [23,24]. LoCUP, a TOO classifier, was the first machine learning model using a multinomial logistic regression classifier with ridge penalties to incorporate tumor purity and reached a 95.8% accuracy [25]. Cup AI Dx [20] used mRNA gene expression data from The Cancer Genome Atlas (TCGA) data set to train a network based on the popular inception model [22] to identify the TOO, achieving an accuracy of 96.7% on a validation set of 354 TCGA metastatic samples. The TOD-CUP method [21] addressed the variation in mRNA platforms and used a gene expression rank–based majority vote algorithm to achieve an overall accuracy of 94%. Early work with microRNAs and nondeep learning machine learning algorithms showed 84% accuracy with k-nearest neighbor models [26] and binary decision trees at 85% [27]. However, the investigation of deep learning machine learning models may improve these accuracies with TOO detection by microRNA. MicroRNAs are also at the forefront of extracellular vesicle liquid biopsy development and may be better suited for the noninvasive classification of TOO [28].

This study sets out to explore the possibility of developing a model for using microRNA profiles from metastatic tissues to determine the TOO through the application of deep learning techniques. Successful TOO detection from microRNAs will

provide a route for cancer detection without requiring samples from the primary tumor site in cases of CUP malignancies. We hypothesize that we would be able to predict the origin of metastatic tumors with higher accuracy than previous reports by leveraging larger data sets of microRNA profiles from both normal and primary site tissues to train the model.

The data for this project were collected from TCGA data set [29] and the Sequence Read Archive (SRA) [30] from microRNA tissue expression database. The TCGA data set contains samples from 18 different cancer types representing 9648 samples, of which 365 were metastatic, 633 were solid normal, and 8650 were from the primary tumor site. Each sample consisted of microRNA expression data, available as RPM (reads per million mapped reads), as well as metadata including age and gender. We split TCGA data set into a combined primary tumor or solid normal samples training set and a metastatic sample test set. We then further split the primary tumor and solid normal samples into a training and validation set with a 9:1 ratio. The training set consisted of 8355 samples and the validation set consisted of 928 samples.

We use 2 data sets for evaluating the performance of our models. The SRA test data set consisted of 194 samples from 5 different cancer types, all of which were from the primary tumor. We also used the metastatic samples from TCGA data set as our final test data set, which contained samples from 6 cancer types. We developed 4 machine learning models, a decision tree classifier, random forest, logistic regression, and finally, a deep learning model. Our deep learning model performed with the highest accuracy, achieving an accuracy of 97% in detecting TOO for metastatic samples and 80.4% on the nonmetastatic SRA cohort. Feature importance analysis revealed the top 3 differentiating microRNA targets as microRNA-10b, microRNA-196b, and microRNA-205, which confirms prior investigations on microRNAs associated with metastatic cancer [31-33].

Methods

Data Sets

In Figure 1, we outline the data preprocessing pipeline. Our study analyzed published data and did not generate any new sequencing data. TCGA data were obtained [29]. Data were further filtered by querying the Genomics Data Commons via the Application Programming Interfaces specified [34]. We restricted the tissue type to be one of the primary tumors, solid tissue normal, or metastatic. We further restricted the data to microRNA transcriptome profiling and picked data corresponding to 18 types of cancer each containing a sufficient number of samples, obtaining 9648 files (Figure 2 and Table S1 in Multimedia Appendix 1).

To obtain the SRA data, we used the microRNA tissue expression database portal and restricted the cancer types to 6 types of cancer, seen in further detail in Figure 2. We obtained 207 samples, each containing expression data for 2656 microRNAs. After removing samples with missing features, 194 samples were remaining.

We selected microRNA features that were expressed in at least 50% (4824/9648) of the samples, which reduced the number of features in the TCGA data set from 1889 to 562. We then picked the common features between the SRA data set and the TCGA data set, reducing this number to 497. On both data sets, we normalized the RPM of the selected features per sample to sum to a million. We then transformed the RPM values using the transformation log(RPM + 1) to restrict the range of the input.

Figure 1. Overview of our data processing pipeline. Data from the TCGA GDC portal and SRA miTED portal was obtained. Underexpressed microRNA and samples containing missing features from the miTED data were filtered. Common features were selected between both data sets, reducing the number of microRNA to 497. Features were normalized as reads per million per sample and log-transformed. TCGA data set was split into (1) the primary tissue and solid normal set and (2) the metastatic test set. The first, combined, set was further split into a training and validation set. GDC: Genomics Data Commons; miTED: microRNA tissue expression database; SRA: Sequence Read Archive; TCGA: The Cancer Genome Atlas.

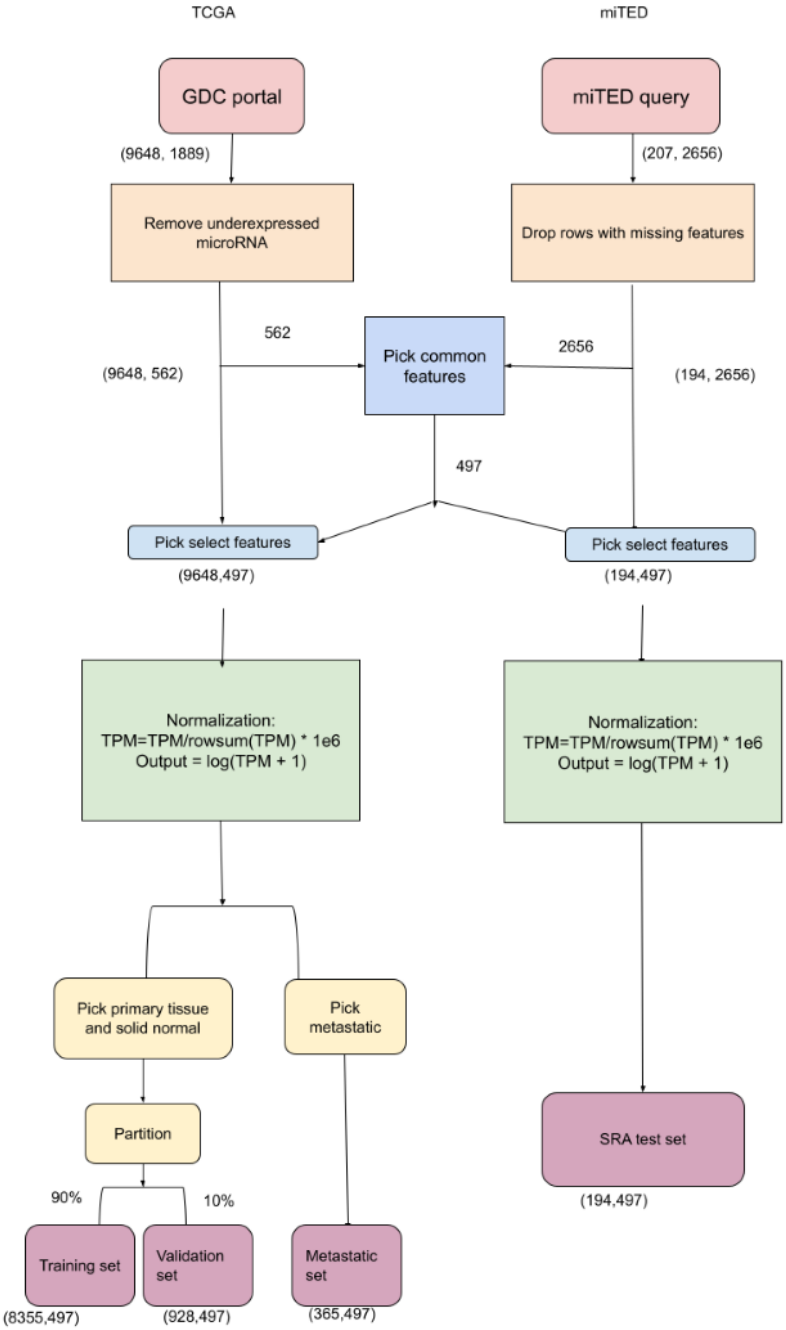
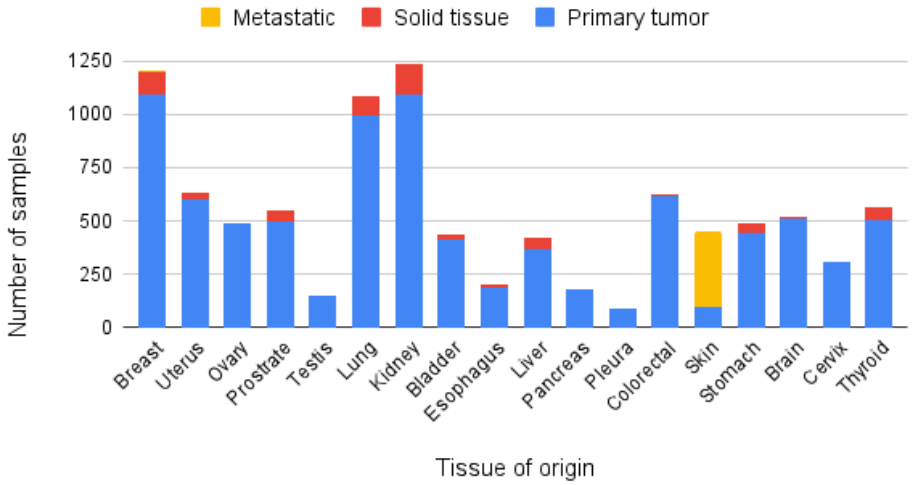


Figure 2. The Cancer Genome Atlas (TCGA) data set distribution across tissue of origin (TOO). Distribution of the different cancer samples in the TCGA data set that are from the primary tumor site, solid tissue, or metastatic. Note that metastatic samples primarily corresponded to the skin as the TOO.



Training Procedure

For the implementation of decision tree, random forest, and logistic regression classifiers, the sklearn package was used [35]. We used classification accuracy as the primary metric to evaluate our models. Deep learning models were created with PyTorch (Meta AI) [36]. To optimize and train our neural network, we used Adam optimizer and trained for 50 epochs. Since our objective was classification, we used softmax with cross-entropy loss [37] to optimize the model. We used the validation set to determine the hyperparameters of the models and picked the best-performing model for further evaluation on the test set. Feature importance was calculated with sklearn's permutation feature importance function.

Ethical Considerations

This study was conducted in accordance with the ethical standards of the Salve Regina University ethical standards. The research study was reviewed by the institutional review board of Salve Regina University and was determined to be exempt from further review as per criteria contained in Title 45 CFR §46.104(d) section 4ii of federal regulations. As such, the study used only publicly available deidentified or anonymized data, and the project was reviewed (Exemption #Wise.2024.6.11).

Results

In order to develop a model to detect TOO, we set out to find the best-performing machine learning model for determining the TOO from the TCGA primary tumor and solid normal tissue cohorts. The models were then tested on the validation set, and we could accurately determine the TOO based on primary or normal microRNA profiles, with an accuracy of over 90% for 15 of 18 different tissue types using deep learning (Table 1 and Table S2 in Multimedia Appendix 1).

We can note that the deep learning model performs consistently the highest on the validation set, with logistic regression and random forest classifiers providing comparable performance.

We then set out to apply our best-performing deep learning model and evaluate its performance on the SRA test set that contains microRNA expression data from primary tumors (Table 2). We accurately determined the TOO with an accuracy of over 90% (90/100) for 3 of the 5 cancer types but saw a decrease in accuracy for bladder and colorectal cancer.

Finally, we analyzed our deep learning model on microRNA expression data from metastatic tissue samples in the TCGA data set (Table 3). We accurately determined the TOO with an accuracy of over 85% (308/362) for all cancer types with an average of 97% (351/362).

Since random forest and logistic regression classifiers provided comparable performance on the primary or normal validation set, we compared the classifier accuracy on both test sets for all created models (Table 4).

The input features of our models consist of microRNA expression data common to TCGA and SRA data sets. Figure 3 describes the overall architecture of the model, which consists of 2 linear layers. The second layer has 18 outputs, corresponding to each cancer type. The cancer type corresponds to the output with the maximum value.

We used dropout for the input layer [38] as it is a common technique to improve model accuracy and reduce overfitting. We also augmented our input data with noise.

To evaluate the performance of our models, we computed confusion matrices for performance on metastatic samples (Figure S2A and S2B in Multimedia Appendix 1) and plotted the receiver operating characteristic curves for performance on metastatic skin cancer (Figure S2C and S2D in Multimedia Appendix 1), as the majority of the metastatic samples were obtained from skin cancer cases. We observed that the deep learning model performed significantly better than our decision tree model, which was consistent when evaluated on the SRA validation cohort (Figure S3 in Multimedia Appendix 1). To illustrate the effectiveness of our models, we created Sankey plots representing the deep learning model performance on metastatic samples from the TCGA data set and primary tissue sites from the SRA data set (Figure 4).

Table 1. Model accuracies on the validation test set. Performance of 4 models for the identification of tissue of origin. The validation set consists of both primary tumor and solid normal tissue samples from The Cancer Genome Atlas data set.

Cancer type	Decision tree (%)	Random forest (%)	Logistic regression (%)	Deep learning (%)
Breast (n=131)	91.6	99.2	96.9	99.2
Uterus (n=73)	76.7	100	90.4	94.5
Ovary (n=48)	89.6	91.6	93.8	100
Prostrate (n=54)	94.5	100	100	100
Testis (n=18)	61.1	94.5	94.4	88.9
Lung (n=117)	81.1	95.7	82.9	98.2
Kidney (n=116)	94.8	100	99.1	100
Bladder (n=35)	71.4	95.7	88.5	88.5
Esophagus (n=24)	33.3	29.2	54.1	83.3
Liver (n=42)	97.6	100	97.6	100
Pancreas (n=20)	55.0	95	95.2	100
Pleura (n=7)	42.8	85.7	100	100
Colorectal (n=57)	85.6	98.2	94.7	100
Skin (n=6)	66.6	100	100	100
Stomach (n =45)	82.2	97.8	75.5	91.1
Brain (n=47)	100	100	100	100
Cervix (n=32)	62.5	78.1	78.1	93.7
Thyroid (n=55)	98.1	100	100	100
Overall—across cancer types	84.6	95.3	96.4	97.2

Table 2. Performance of our deep learning model for the identification of tissue of origin on the primary tissue site cohorts from the SRA[a].

Cancer type	SRA test accuracy—deep learning (%)
Breast (n=44)	91.6
Prostrate (n=37)	100
Lung (n=19)	100
Bladder (n=10)	80
Colorectal (n=78)	58.9
Skin (n=0)	N/A[b]
Overall—across cancer types	80.4

[a]SRA: Sequence Read Archive.

[b]N/A: not applicable.

Table 3. Performance of our deep learning model for the identification of tissue of origin in metastatic tumor tissue.

Cancer type	TCGA[a] metastatic test accuracy—deep learning (%)
Breast (n=7)	85.7
Prostrate (n=1)	100
Lung (n=0)	N/A[b]
Bladder (n=1)	100
Colorectal (n=1)	100
Skin (n=352)	97.4
Overall—across cancer types	97

[a]TCGA: The Cancer Genome Atlas.

[b]N/A: not applicable.

Table 4. Accuracy of developed models on metastatic and SRA[a] test sets. The accuracy for all 4 models is presented on the TCGA[b] metastatic and SRA cohorts. The decision tree classifier had a depth of 14 and the random forest had a depth of 19.

Classifier	Accuracy on TCGA metastatic test set (%)	Accuracy on SRA test set (%)
Decision tree	62.5	41.2
Random forest	94.2	74.2
Logistic regression	93.2	71.6
Deep learning	97	80.4

[a]SRA: Sequence Read Archive.

[b]TCGA: The Cancer Genome Atlas.

Figure 3. A schematic of the machine learning model architecture. MiRNA: microRNA.

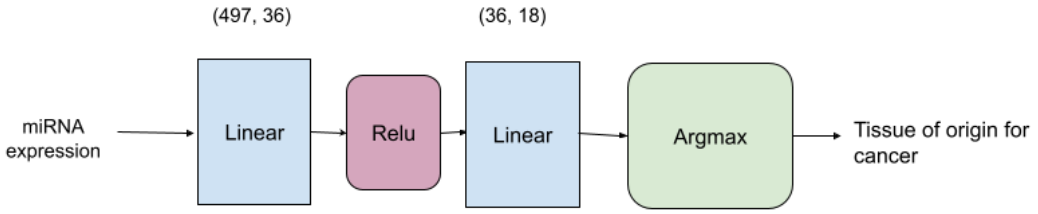
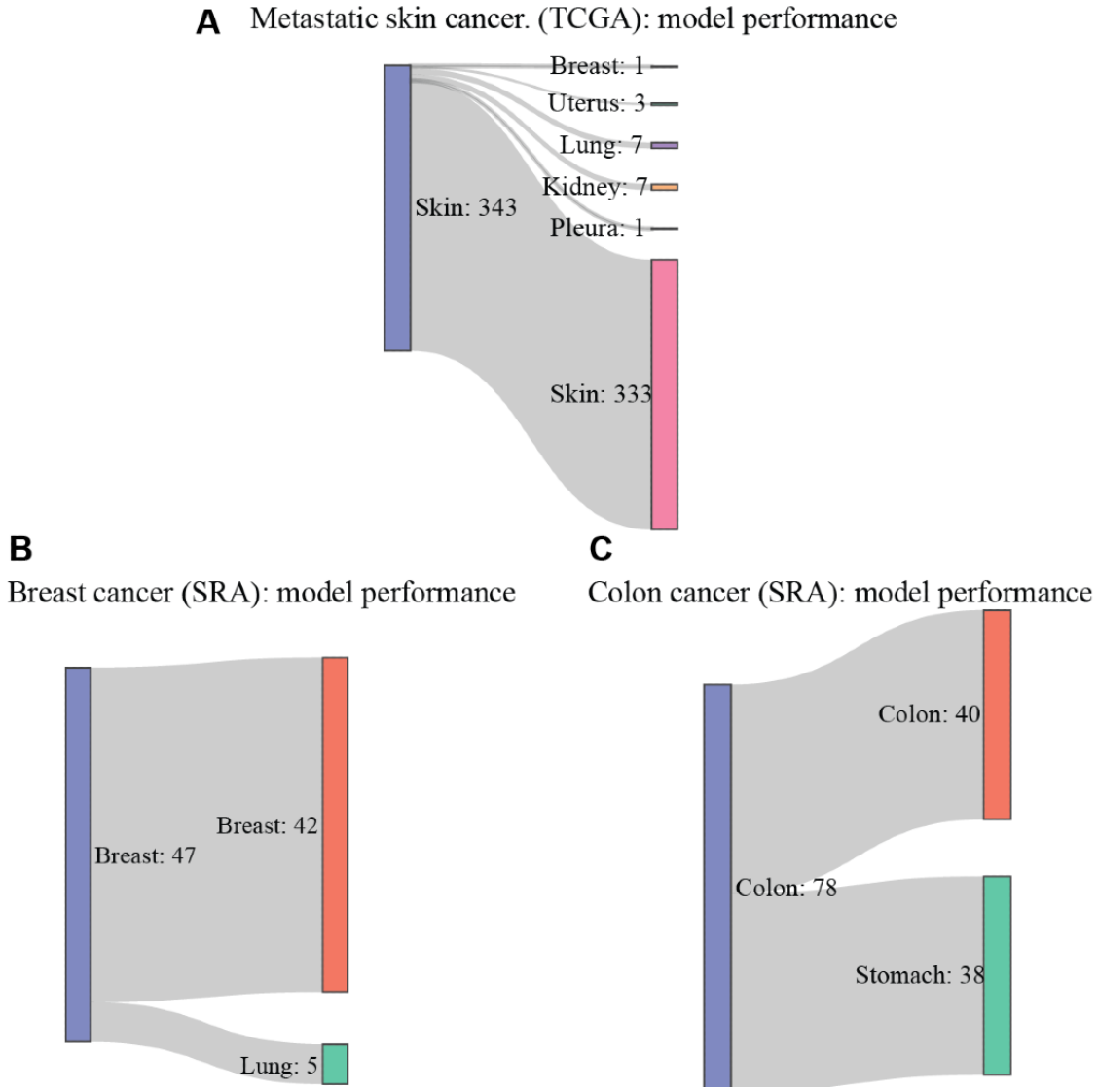


Figure 4. Sankey plot for deep learning model on Sequence Read Archive (SRA) and The Cancer Genome Atlas (TCGA) test data sets. (A) On the TCGA data set, our deep learning model is able to correctly classify 333 out of 343 metastatic skin cancer samples, demonstrating high accuracy. (B) On the SRA test data set, we show representative plots for breast and colon cancers, showing high accuracy for breast cancer tissue of origin identification. (C) The model performance on colon cancer is less accurate due to microRNA expression consistently overlapping for colon and stomach cancers [40].

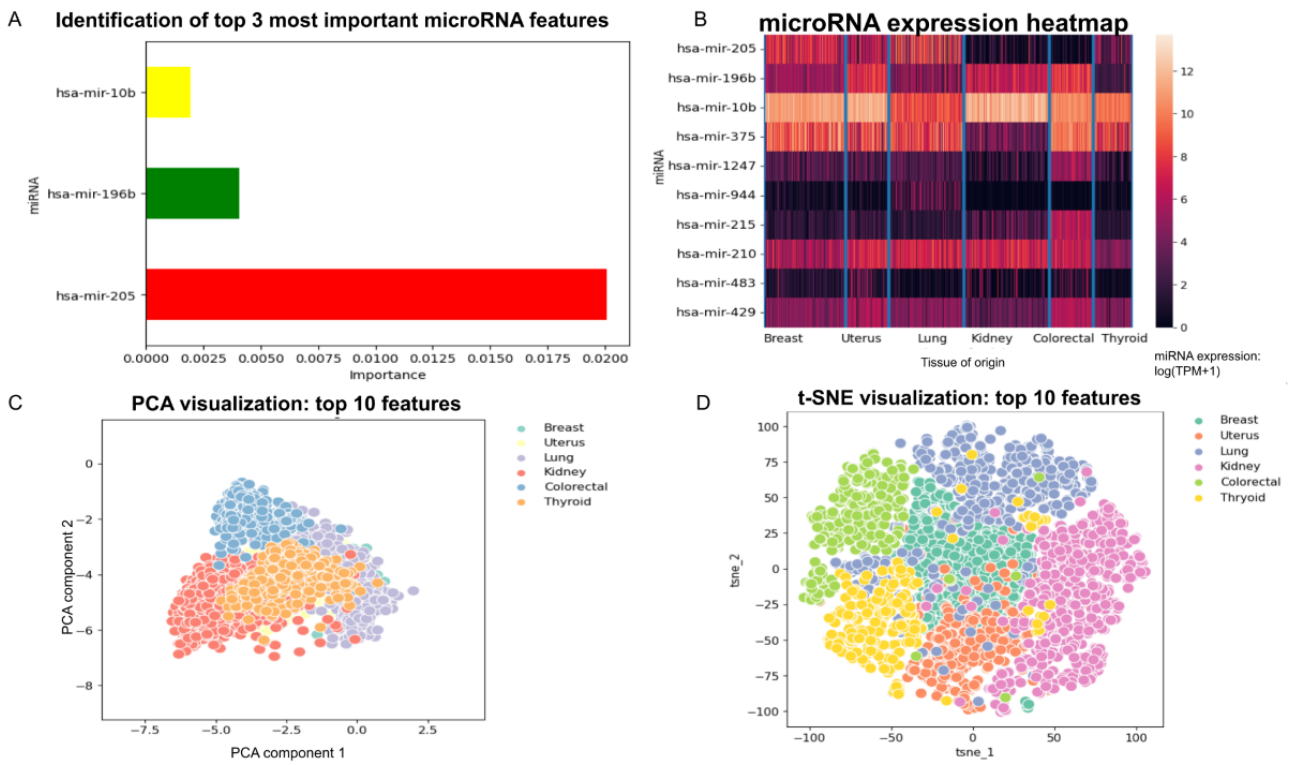


These results confirm our hypotheses and show that we were able to predict the TOO with high accuracy using deep learning. Furthermore, our findings demonstrated that deep learning techniques significantly increase the accuracy in comparison to decision tree, logistic regression, and random forest models.

To reveal the significance of individual features, we performed feature importance analysis using the permutation feature importance method (Figure 5A). The top 3 microRNAs contributing to our deep learning model based on our combined normal and primary site training set are microRNA-10b, microRNA-196, and microRNA-205. MicroRNA-10b has been shown to function as a metastasis-promoting factor in many cancer types. In fact, it was one of the first microRNAs to have been discovered with aberrant expression in cancer cells [31]. MicroRNA-196 has been linked to the progression of many cancers, notably metastatic colorectal cancer [32], while microRNA-205 expression is downregulated in metastatic breast and prostate cancer [33].

To further understand the significance of the identified important features, we compute a heat map (Figure 5B) showing the microRNA expression values for the top 10 microRNA features for samples in the training data set. Visually, it is apparent that the microRNA features can be used to distinguish the cancer type. To further validate this, we perform principal component analysis and t-SNE analysis using only the top 10 features (Figures 5C and 5D). We note that the t-SNE plot shows a clear separation of features into distinct clusters corresponding to each cancer type, showing the significance of the features for detecting the TOO.

Figure 5. MicroRNA feature importance visualizations. (A) Permutation feature importance for the top 3 microRNA candidates. A bar graph of the importance values for the 3 top microRNA candidates for the logistic regression model. (B) MicroRNA expression heatmap. MicroRNA expression values for the top 10 most important features (as determined by permutation feature importance) for a subset of samples. The top 10 microRNA features can cluster cancer type. Low mir-205 and mir-944 and a high mir-10b are indicative of colorectal cancer. Similarly, low expressions for microRNA-429, microRNA-483, microRNA-215, microRNA-944, microRNA-1247, microRNA-375, and microRNA-205 are indicative of kidney cancer. (C) PCA visualization. (D) t-SNE visualization. PCA and t-SNE visualization of data corresponding to the 6 cancer types with the most samples in our data set, using only the top 10 microRNA features. In the PCA plot, note that there is significant overlap between the cancer types, while in the t-SNE plot, the cancer types are well separated, suggesting that with 10 microRNA features, machine learning models may correctly identify patterns and predict tissue of origin. PCA: principal component analysis; t-SNE: t-distributed stochastic neighbor embedding.



Discussion

Principal Findings

In these investigations, while using successively more powerful classifiers, we were able to detect the TOO on solely metastatic cancer samples with accuracies ranging from 62.5% (226/362) with a decision tree to 97% (351/362) with a deep learning model. Our methods show that one can leverage larger amounts of gene expression data for primary and solid tissue normal tumor samples (~10,000 samples) to come up with accurate classifiers to determine TOO for metastatic cancer (currently limited to ~300 samples). In order to verify the robustness of our model, we assessed its performance on primary tumor data from the SRA and obtained accuracies ranging from 41.2% (77/188) with decision tree to 80.4% (151/188) when using deep learning. Our methods have also identified promising microRNA candidates, reaffirming prior research in this field and demonstrating the potential of machine learning.

The predominant failure of our model on the SRA test cohort was within colorectal cancer as can be seen in Figure 4C. Many colorectal samples were incorrectly classified as stomach or gastric cancer. This is consistent with previous research in this area as microRNA expression profiles for gastrointestinal cancers show significant overlap [39]. In addition, colorectal and stomach cancer are often synchronous with probabilities ranging from 20.1% to 37.2% [40].

We used permutation feature importance, a model-agnostic metric that permutes features across samples in the test set to assess the change in model accuracy. The results are in line with existing research in this area and serve as a good indicator of the feasibility of machine learning techniques to identify promising biomarkers.

Limitations

To effectively use our model in clinical care, accuracy must be improved further. Our model currently performs with an accuracy of 97% (351/362). While this may seem impressive, clinical classifiers should be highly accurate so that there are a negligible number of cases with errors in identifying TOO. To improve the accuracy, the accumulation of larger data sets is necessary, and as the noncoding genome continues to reveal significant contributions to cancer, we predict that available data sets will expand. A further limitation to our study is that the available microRNA metastatic data sets are predominantly skin cancer. Thus, access to a larger, more varied, data set would improve our assessment of model performance. Furthermore, in order to develop a truly noninvasive method of TOO identification relevant to all cancers, it would be ideal to extend our method to microRNA expression data from blood samples. Detecting the TOO through blood-based microRNA biomarkers

would significantly impact the diagnosis and treatment of patients with CUP. Additionally, our model cannot differentiate between tumor and solid tissue normal samples, as it was designed to identify the TOO specifically.

Conclusions

To summarize, our developed machine learning models can accurately identify the TOO with high accuracy from microRNA expression data when trained on primary tumor and solid tissue samples. Importantly, our results identified key microRNA differentiators of tissue type. Our models are robust and perform well across different data sets (TCGA and the SRA data set). We look forward to developing further deep learning models that can accurately detect TOO as microRNA data sets expand, with the goal of having a noninvasive test for diagnosing the presence of cancer and determining cancer TOO with high accuracy.

Acknowledgments

We are grateful to The Cancer Genome Atlas Project and patients for providing the data used in this research. We are thankful to Soroush Hajizadeh for reviewing our code and providing insightful feedback. The results are in part based on data generated by The Cancer Genome Atlas Project Research Network [29].

Data Availability

The data sets generated during and/or analyzed during this study are available in the github repository [41].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed analyses of feature importance and confusion matrices to support the study's findings on tissue of origin classification using microRNA features.
[DOCX File , 619 KB - bioinform_v5i1e56538_app1.docx]

References

1. Qaseem A, Usman N, Jayaraj J, Janapala R, Kashif T. Cancer of unknown primary: a review on clinical guidelines in the development and targeted management of patients with the unknown primary site. Cureus 2019;11(9):e5552 [FREE Full text] [doi: 10.7759/cureus.5552] [Medline: 31695975]
2. Monzon FA, Medeiros F, Lyons-Weiler M, Henner WD. Identification of tissue of origin in carcinoma of unknown primary with a microarray-based gene expression test. Diagn Pathol 2010;5:3 [FREE Full text] [doi: 10.1186/1746-1596-5-3] [Medline: 20205775]
3. Laprovitera N, Riefolo M, Ambrosini E, Klec C, Pichler M, Ferracin M. Cancer of unknown primary: challenges and progress in clinical management. Cancers (Basel) 2021;13(3):451 [FREE Full text] [doi: 10.3390/cancers13030451] [Medline: 33504059]
4. Pu X, Yang S, Xu Y, Chen B, Wang Q, Gong Q, et al. Case report: tissue origin identification for cancer of unknown primary: gene expression profiling approach. Front Oncol 2021;11:702887 [FREE Full text] [doi: 10.3389/fonc.2021.702887] [Medline: 34858803]
5. Tang W, Wan S, Yang Z, Teschendorff AE, Zou Q. Tumor origin detection with tissue-specific miRNA and DNA methylation markers. Bioinformatics 2018;34(3):398-406. [doi: 10.1093/bioinformatics/btx622] [Medline: 29028927]
6. Greco FA, Burris HA, Erland JB, Gray JR, Kalman LA, Schreeder MT, et al. Carcinoma of unknown primary site. Cancer 2000;89(12):2655-2660. [Medline: 11135228]
7. Schapira DV, Jarrett AR. The need to consider survival, outcome, and expense when evaluating and treating patients with unknown primary carcinoma. Arch Intern Med 1995;155(19):2050-2054. [Medline: 7575063]
8. Zhang B, Pan X, Cobb GP, Anderson TA. microRNAs as oncogenes and tumor suppressors. Dev Biol 2007;302(1):1-12 [FREE Full text] [doi: 10.1016/j.ydbio.2006.08.028] [Medline: 16989803]
9. Chen B, Dragomir MP, Yang C, Li Q, Horst D, Calin GA. Targeting non-coding RNAs to overcome cancer therapy resistance. Signal Transduct Target Ther 2022;7(1):121 [FREE Full text] [doi: 10.1038/s41392-022-00975-3] [Medline: 35418578]
10. Smolarz B, Durczyński A, Romanowicz H, Szyłło K, Hogendorf P. miRNAs in cancer (review of literature). Int J Mol Sci 2022;23(5):2805 [FREE Full text] [doi: 10.3390/ijms23052805] [Medline: 35269947]
11. Chakrabortty A, Patton DJ, Smith BF, Agarwal P. miRNAs: potential as biomarkers and therapeutic targets for cancer. Genes (Basel) 2023;14(7):1375 [FREE Full text] [doi: 10.3390/genes14071375] [Medline: 37510280]
12. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, et al. MicroRNA expression profiles classify human cancers. Nature 2005;435(7043):834-838. [doi: 10.1038/nature03702] [Medline: 15944708]

13. Rosenfeld N, Aharonov R, Meiri E, Rosenwald S, Spector Y, Zepeniuk M, et al. MicroRNAs accurately identify cancer tissue origin. Nat Biotechnol 2008;26(4):462-469. [doi: 10.1038/nbt1392] [Medline: 18362881]
14. Tucci P. The role of microRNAs in cancer: functions, biomarkers and therapeutics. Cancers (Basel) 2022;14(4):872 [FREE Full text] [doi: 10.3390/cancers14040872] [Medline: 35205620]
15. Zhang B, Shi H, Wang H. Machine learning and AI in cancer prognosis, prediction, and treatment selection: a critical approach. J Multidiscip Healthc 2023;16:1779-1791 [FREE Full text] [doi: 10.2147/JMDH.S410301] [Medline: 37398894]
16. Song YY, Lu Y. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry 2015;27(2):130-135 [FREE Full text] [doi: 10.11919/j.issn.1002-0829.215044] [Medline: 26120265]
17. Breiman L. Random forests. Mach Learn 2001;45:5-32 [FREE Full text]
18. Chung M. Introduction to Logistic Regression. Ithaca, NY: Cornell University; 2020.
19. Tran KA, Kondrashova O, Bradley A, Williams ED, Pearson JV, Waddell N. Deep learning in cancer diagnosis, prognosis and treatment selection. Genome Med 2021;13(1):152 [FREE Full text] [doi: 10.1186/s13073-021-00968-x] [Medline: 34579788]
20. Zhao Y, Pan Z, Namburi S, Pattison A, Posner A, Balachander S, et al. CUP-AI-Dx: a tool for inferring cancer tissue of origin and molecular subtype using RNA gene-expression data and artificial intelligence. EBioMedicine 2020;61:103030 [FREE Full text] [doi: 10.1016/j.ebiom.2020.103030] [Medline: 33039710]
21. Shen Y, Chu Q, Yin X, He Y, Bai P, Wang Y, et al. TOD-CUP: a gene expression rank-based majority vote algorithm for tissue origin diagnosis of cancers of unknown primary. Brief Bioinform 2021;22(2):2106-2118. [doi: 10.1093/bib/bbaa031] [Medline: 32266390]
22. Tothill RW, Kowalczyk A, Rischin D, Bousioutas A, Haviv I, van Laar RK, et al. An expression-based site of origin diagnostic method designed for clinical application to cancer of unknown origin. Cancer Res 2005;65(10):4031-4040. [doi: 10.1158/0008-5472.CAN-04-3617] [Medline: 15899792]
23. van Laar RK, Ma XJ, de Jong D, Wehkamp D, Floore AN, Warmoes MO, et al. Implementation of a novel microarray-based diagnostic test for cancer of unknown primary. Int J Cancer 2009;125(6):1390-1397 [FREE Full text] [doi: 10.1002/ijc.24504] [Medline: 19536816]
24. Ma XJ, Patel R, Wang X, Salunga R, Murage J, Desai R, et al. Molecular classification of human cancers using a 92-gene real-time quantitative polymerase chain reaction assay. Arch Pathol Lab Med 2006;130(4):465-473 [FREE Full text] [doi: 10.5858/2006-130-465-MCOHCU] [Medline: 16594740]
25. Søndergaard D, Nielsen S, Pedersen CNS, Besenbacher S. Prediction of primary tumors in cancers of unknown primary. J Integr Bioinform 2017;14(2):20170013 [FREE Full text] [doi: 10.1515/jib-2017-0013] [Medline: 28686574]
26. Varadhachary GR, Spector Y, Abbruzzese JL, Rosenwald S, Wang H, Aharonov R, et al. Prospective gene signature study using microRNA to identify the tissue of origin in patients with carcinoma of unknown primary. Clin Cancer Res 2011;17(12):4063-4070. [doi: 10.1158/1078-0432.CCR-10-2599] [Medline: 21531815]
27. Rosenwald S, Gilad S, Benjamin S, Lebanony D, Dromi N, Faerman A, et al. Validation of a microRNA-based qRT-PCR test for accurate identification of tumor tissue origin. Mod Pathol 2010;23(6):814-823 [FREE Full text] [doi: 10.1038/modpathol.2010.57] [Medline: 20348879]
28. de Miguel Pérez D, Rodriguez Martínez A, Ortigosa Palomo A, Delgado Ureña M, Garcia Puche JL, Robles Remacho A, et al. Extracellular vesicle-miRNAs as liquid biopsy biomarkers for disease identification and prognosis in metastatic colorectal cancer patients. Sci Rep 2020;10(1):3974 [FREE Full text] [doi: 10.1038/s41598-020-60212-1] [Medline: 32132553]
29. The Cancer Genome Atlas Program (TCGA). National Cancer Institute. URL: https://www.cancer.gov/ccg/research/genome-sequencing/tcga [accessed 2024-05-31]
30. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database Collaboration. The sequence read archive. Nucleic Acids Res 2011;39(Database issue):D19-D21 [FREE Full text] [doi: 10.1093/nar/gkq1019] [Medline: 21062823]
31. Sheedy P, Medarova Z. The fundamental role of miR-10b in metastatic cancer. Am J Cancer Res 2018;8(9):1674-1688 [FREE Full text] [Medline: 30323962]
32. Chauhan N, Dhasmana A, Jaggi M, Chauhan SC, Yallapu MM. miR-205: a potential biomedicine for cancer therapy. Cells 2020;9(9):1957 [FREE Full text] [doi: 10.3390/cells9091957] [Medline: 32854238]
33. Lu Y, Chang JT, Chan E, Chao Y, Yeh T, Chen J, et al. miR-196, an emerging cancer biomarker for digestive tract cancers. J Cancer 2016;7(6):650-655 [FREE Full text] [doi: 10.7150/jca.13460] [Medline: 27076845]
34. NCI Genomic Data Commons. GDC Application Programming Interface (API). URL: https://gdc.cancer.gov/developers/gdc-application-programming-interface-api [accessed 2024-05-31]
35. Pedregosa, Fabian, Varquaux, Gaël, Gramfort, Alexandre, Michel, Vincent, Thirion, Bertrand, Grisel, Olivier, Blondel, Mathieu, Pretternhofer, Peter, Weiss, Ron, Dubourg, Vincent, Vanderplas, Jake, Passos, Alexandre, Cornapeau, David, Brucher, Matthieu, Perrot, Mathhieu. Scikit-learn: machine learning in Python. J Mach Learn Res 2011:2825-2830 [FREE Full text]
36. Paszke A, Gross S, Massa F. Pytorch: an imperative style, high-performance deep learning library. 2019 Presented at: arXiv.org; Dec 3 2019; NA URL: https://doi.org/10.48550/arXiv.1912.01703 [doi: https://doi.org/10.48550/arXiv.1912.01703]

37. Mao A, Mohri M, Zhong Y. Cross-entropy loss functions: theoretical analysis and applications. : PMLR; 2023 Presented at: Proceedings of the 40th International Conference on Machine Learning; July 2023; Honolulu, Hawaii p. 23803-23828.
38. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Toronto: JMLR; 2014.
39. Laprovitera N, Riefolo M, Porcellini E, Durante G, Garajova I, Vasuri F, et al. MicroRNA expression profiling with a droplet digital PCR assay enables molecular diagnosis and prognosis of cancers of unknown primary. Mol Oncol 2021;15(10):2732-2751 [FREE Full text] [doi: 10.1002/1878-0261.13026] [Medline: 34075699]
40. Suh BJ. Synchronous and metachronous colon cancers in patients with gastric cancer: report of 2 cases. Case Rep Oncol 2016;9(3):752-759 [FREE Full text] [doi: 10.1159/000452831] [Medline: 27990112]
41. ANIHA234/identifying-tissue-of-origin-from-mirna. GitHub. 2023. URL: https://github.com/Anisha234/miRNA [accessed 2024-05-31]

Abbreviations

CUP: carcinoma of unknown primary
RPM: reads per million
SRA: Sequence Read Archive
TCGA: The Cancer Genome Atlas
TOO: tissue of origin

Edited by A Uzun; submitted 18.01.24; peer-reviewed by A Robertson, VA Timmaraju; comments to author 06.02.24; revised version received 02.04.24; accepted 25.04.24; published 24.07.24.

Please cite as:
Raghu A, Raghu A, Wise JF
Deep Learning–Based Identification of Tissue of Origin for Carcinomas of Unknown Primary Using MicroRNA Expression: Algorithm Development and Validation
JMIR Bioinform Biotech 2024;5:e56538
URL: https://bioinform.jmir.org/2024/1/e56538
doi:10.2196/56538
PMID:39046787

Original Paper

Enhancing Suicide Risk Prediction With Polygenic Scores in Psychiatric Emergency Settings: Prospective Study

Younga Heather Lee[1,2,3*], PhD; Yingzhe Zhang[1,2,4*], MS; Chris J Kennedy[3,5], MPA, PhD; Travis T Mallard[1,2,3], PhD; Zhaowen Liu[1,2,3,6], PhD; Phuong Linh Vu[7]; Yen-Chen Anne Feng[1,2,3,8], MS, ScD; Tian Ge[1,2,3,5], PhD; Maria V Petukhova[9], PhD; Ronald C Kessler[9], PhD; Matthew K Nock[10,11,12], PhD; Jordan W Smoller[1,2,3,5], MD, ScD

[1]Psychiatric & Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, United States
[2]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, United States
[3]Department of Psychiatry, Harvard Medical School, Boston, MA, United States
[4]Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA, United States
[5]Center for Precision Psychiatry, Massachusetts General Hospital, Boston, MA, United States
[6]School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[7]Harvard College, Harvard University, Cambridge, MA, United States
[8]Institute of Health Data Analytics and Statistics, College of Public Health, National Taiwan University, Taipei, Taiwan
[9]Department of Health Care Policy, Harvard Medical School, Boston, MA, United States
[10]Department of Psychology, Harvard University, Cambridge, MA, United States
[11]Mental Health Research Program, Franciscan Children's, Brighton, MA, United States
[12]Department of Psychiatry, Massachusetts General Hospital, Boston, MA, United States
[*]these authors contributed equally

Corresponding Author:
Jordan W Smoller, MD, ScD
Psychiatric & Neurodevelopmental Genetics Unit
Center for Genomic Medicine
Massachusetts General Hospital
185 Cambridge St
6th Floor
Boston, MA, 02114
United States
Phone: 1 617 724 0835
Fax: 1 617 726 0830
Email: jsmoller@mgh.harvard.edu

Abstract

Background: Despite growing interest in the clinical translation of polygenic risk scores (PRSs), it remains uncertain to what extent genomic information can enhance the prediction of psychiatric outcomes beyond the data collected during clinical visits alone.

Objective: This study aimed to assess the clinical utility of incorporating PRSs into a suicide risk prediction model trained on electronic health records (EHRs) and patient-reported surveys among patients admitted to the emergency department.

Methods: Study participants were recruited from the psychiatric emergency department at Massachusetts General Hospital. There were 333 adult patients of European ancestry who had high-quality genotype data available through their participation in the Mass General Brigham Biobank. Multiple neuropsychiatric PRSs were added to a previously validated suicide prediction model in a prospective cohort enrolled between February 4, 2015, and March 13, 2017. Data analysis was performed from July 11, 2022, to August 31, 2023. Suicide attempt was defined using diagnostic codes from longitudinal EHRs combined with 6-month follow-up surveys. The clinical risk score for suicide attempt was calculated from an ensemble model trained using an EHR-based suicide risk score and a brief survey, and it was subsequently used to define the baseline model. We generated PRSs for depression, bipolar disorder, schizophrenia, suicide attempt, and externalizing traits using a Bayesian polygenic scoring method for European ancestry participants. Model performance was evaluated using area under the receiver operator curve (AUC), area under the precision-recall curve, and positive predictive values.

Results: Of the 333 patients (n=178, 53.5% male; mean age 36.8, SD 13.6 years; n=333, 100% non-Hispanic and n=324, 97.3% self-reported White), 28 (8.4%) had a suicide attempt within 6 months. Adding either the schizophrenia PRS or all PRSs to the baseline model resulted in the numerically highest discrimination (AUC 0.86, 95% CI 0.73-0.99) compared to the baseline model (AUC 0.84, 95% Cl 0.70-0.98). However, the improvement in model performance was not statistically significant.

Conclusions: In this study, incorporating genomic information into clinical prediction models for suicide attempt did not improve patient risk stratification. Larger studies that include more diverse participants are required to validate whether the inclusion of psychiatric PRSs in clinical prediction models can enhance the stratification of patients at risk of suicide attempts.

(JMIR Bioinform Biotech 2024;5:e58357) doi:10.2196/58357

KEYWORDS

polygenic risk score; suicide risk prediction; suicide attempt; predictive algorithms; genomics; genotypes; electronic health record; machine learning

Introduction

Between 2000 and 2018, suicide rates increased by 37%, making suicide one of the leading causes of death in the United States [1]. Data from US health care systems show that most individuals who die by suicide in the United States had health care visits in the month preceding their death, highlighting opportunities for health care providers to identify and intervene with those at risk for suicide-related behavior [2].

We previously developed and validated a prognostic model combining electronic health records (EHRs) and a brief patient-reported survey that was able to prospectively predict short-term risk for suicide attempts after an emergency department (ED) visit for psychiatric problems [3]. This study was designed to extend our previous work by evaluating whether adding polygenic risk scores (PRSs) for neuropsychiatric phenotypes can improve the predictive performance of models trained on clinical data (EHR + survey) alone.

The incorporation of PRSs into data-driven prediction models could be justified if PRSs sufficiently improved predictive performance and were paired with evidence-based interventions. Although integrating PRSs into clinical workflows presents implementation challenges, there is increasing momentum toward the broad implementation of genomic information in health care practice [4]. As the cost of genome sequencing continues to decrease, genomic data are expected to ultimately become a standard component of patient health care records. The goal of this paper was to provide a first look at whether such information might in fact provide predictive enhancements that could justify its use.

Methods

Sample

Eligible patients for this study were those who participated in our previous study [3] of adult patients visiting the ED between February 4, 2015, and March 13, 2017; had their blood samples genotyped through their participation in the Mass General Brigham (MGB) Biobank [5] (88% self-reported White); and had nonmissing information on suicide attempt(s) within 6 months following their ED discharge. In total, 333 patients with genetically identified European ancestry met the eligibility criteria and demonstrated a suicide attempt prevalence of 8.4% (n=28) at the 6-month follow-up (n=178, 53.5% self-reported male and n=324, 97.3% self-reported White). Although our previous study [3] also examined suicide attempts at 1 month after ED discharge, the event rate within this window was too low to permit stable estimates. The study sample differed significantly from the original cohort [3] by age (P<.001), self-reported race (P<.001) and ethnicity (P=.06), insurance type (P=.001), and patterns of health care utilization (P<.001; see Multimedia Appendix 1 [3]). Details on recruitment, informed consent process, and data collection can be found in Boutin et al [5] (for the MGB Biobank study) and Nock et al [3] (for the suicide prediction study).

Outcome

The primary outcome was any suicide attempt within 6 months of the ED visit based on either follow-up surveys or a review of linked EHRs [3]. For the latter, we used the *International Classification of Diseases, Ninth Revision* (*ICD-9*) and *International Classification of Diseases, Tenth Revision* (*ICD-10*) to identify qualifying diagnostic codes for suicide attempts that we previously validated [6,7].

Predictors

We extracted the predicted probabilities from the best-performing ensemble model from our previous work [3] for 6-month suicide attempts. This model incorporated patient-reported surveys, a previously developed EHR-based suicide risk score, and sociodemographic characteristics (eg, age, sex, income, education, race and ethnicity, and employment status). In addition, we generated PRSs for depression, bipolar disorder, schizophrenia, suicide attempt, and externalizing traits derived from the largest available European ancestry genome-wide association study of these phenotypes using a Bayesian polygenic risk scoring method called “PRS-CS” (see Multimedia Appendices 2 and 3) [8]. We subsequently residualized individual disorder PRSs for biological sex, age, genomic chip, and the top 20 principal components for population stratification to adjust for potential confounding.

Statistical Analysis

We first established the baseline model by fitting our previously validated suicide risk score and calculated patient risk stratification accuracy (measured using the area under the receiver operating characteristic curve [AUC], area under the precision-recall curve [AUPRC], and positive predictive value

[PPV]). We then added each PRS to the baseline model to evaluate whether adding individual disorder PRSs would improve the AUC, AUPRC, or PPV. Lastly, we incorporated all 5 PRSs to examine whether incorporating multiple neuropsychiatric PRSs would increase the predetermined metrics more than adding individual disorder PRSs to the baseline model alone.

In addition to fitting logistic regression models, we used the SuperLearner stacked generalization approach that combines predictions across a range of algorithms, including those that can capture nonlinear relationships (see Multimedia Appendix 4) [9]. We used 10-fold stratified cross-validation in a 70% training sample (n_{train}=235) to develop the models and evaluated the models in a 30% holdout sample ($n_{holdout}$=98). There were no significant differences in sample characteristics and feature distributions between the train and holdout samples (all P>.05; see Multimedia Appendix 5). All statistical analyses were conducted using R software (version 4.1.2; R Foundation for Statistical Computing).

Ethical Considerations

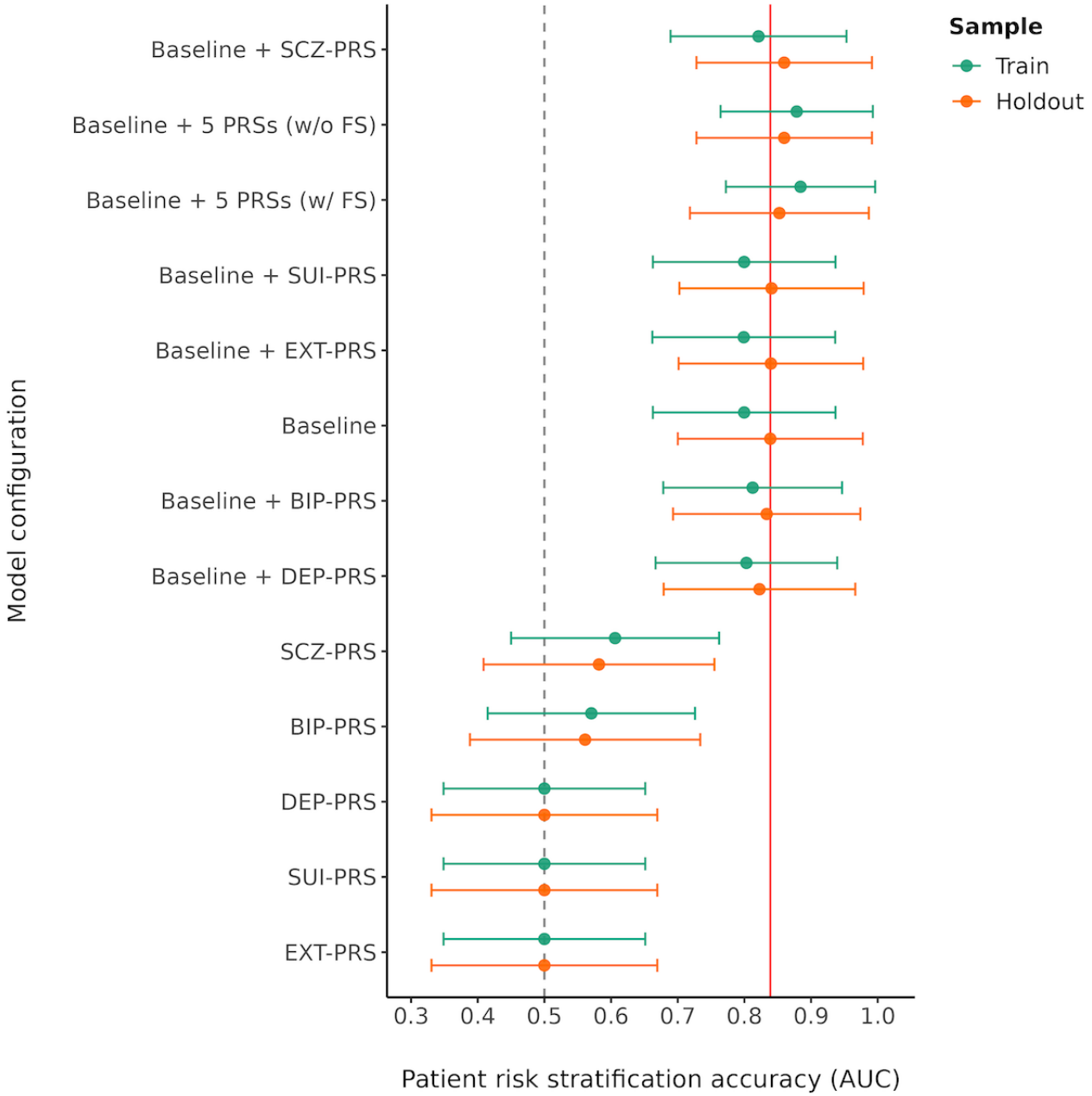
The study procedures were approved by the Institutional Review Boards of Harvard University and MGB (protocol code 2010P000246, approved on February 18, 2010). Additionally, the MGB Biobank study was conducted in accordance with the Declaration of Helsinki and approved by the MGB Institutional Review Board (protocol code 2009P002312, approved on April 29, 2010), with no compensation provided to participants. This study involves secondary analyses using de-identified data from the original studies, which is covered under the initial consent and IRB approval, without requiring additional consent.

Results

Model Discrimination

The baseline model for 6-month suicide attempts had an AUC of 0.84 (95% CI 0.70-0.98; see Figure 1 and Multimedia Appendix 6)**.** Models that included individual disorder PRSs alone had modest or poor AUC, with the schizophrenia PRS having the highest AUC (0.58, 95% CI 0.41-0.76), followed by the bipolar disorder PRS (0.56, 95% CI 0.39-0.73). When individual disorder PRSs were added to the baseline model, the logistic regression and the ensemble models that included the schizophrenia PRS and clinical risk score had the highest AUC (0.86, 95% CI 0.73-0.99), followed by ensemble models each including the suicide PRS and externalizing disorder PRS, but these provided only a modest numerical increase in AUC compared to the baseline model alone (see Figure 1). In general, there was no improvement in AUC when adding the PRS for depression or bipolar disorder to the clinical risk score. However, we observed a numerically higher AUC when the depression PRS was incorporated using an ensemble approach than using logistic regression. The ensemble model that included the clinical risk score and all 5 PRSs had the same AUC (0.86, 95% CI 0.72-0.99) as the ensemble model including the schizophrenia PRS and clinical risk score and had nearly the same AUC as the logistic regression including the same set of features.

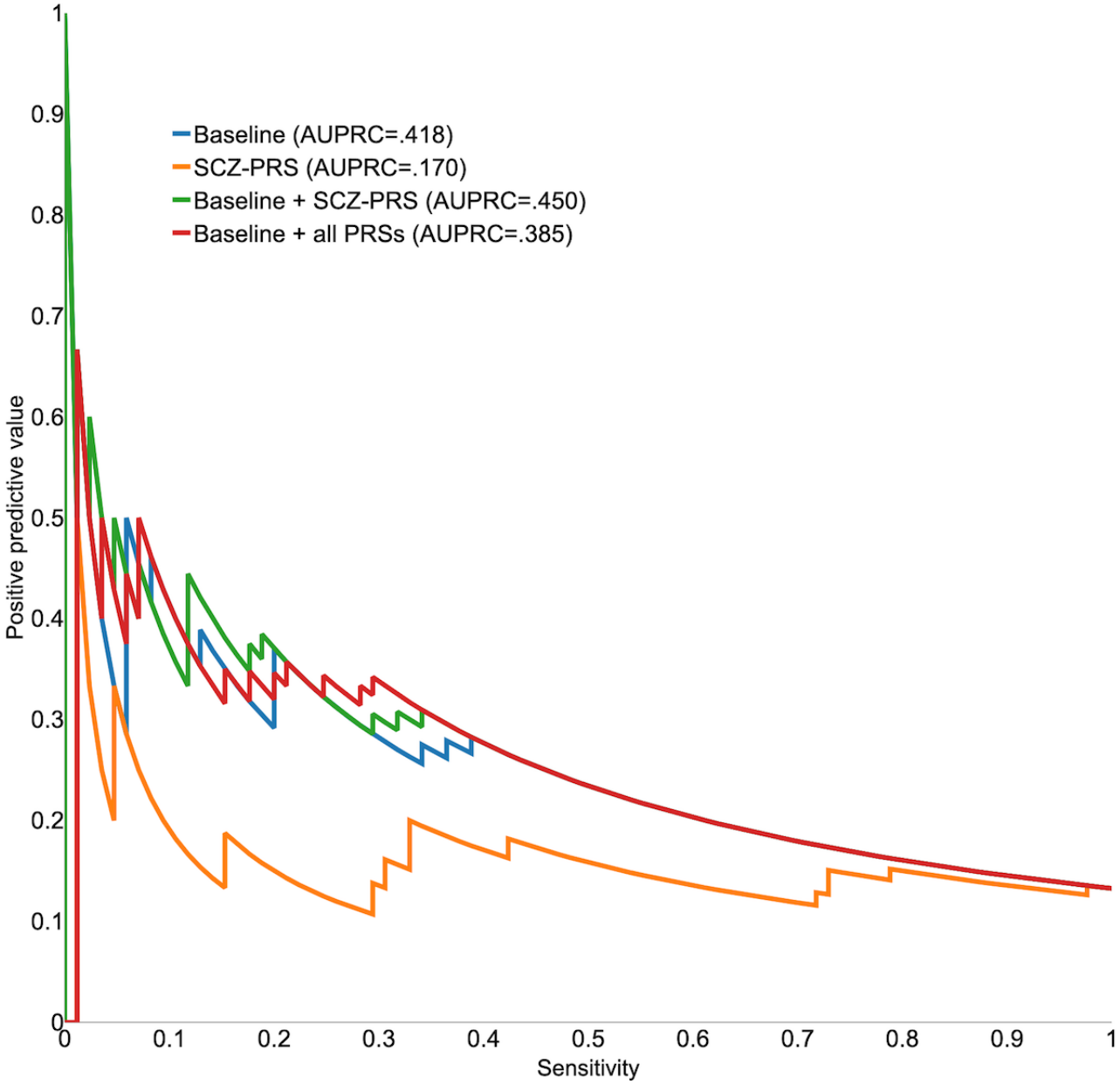
Figure 1. Patient risk stratification accuracy from SuperLearner models estimated using the train (in green) and holdout (in orange) samples. The y-axis is sorted based on the AUC point estimates in the holdout sample. The red line represents the reference AUC point estimate from the baseline model in the holdout sample and is depicted to facilitate visual comparison of AUC estimates across different model configurations. Baseline: baseline clinical risk score for suicide attempt; BIP: bipolar disorder; DEP: depression; EXT: externalizing traits; PRS: polygenic risk score; SCZ: schizophrenia; SUI: suicide attempt; w: with; w/o: without.



Model Performance

We examined precision-recall curves to see how PPV varies across levels of sensitivity with the goal of explaining the best-performing model, which included the clinical risk score and schizophrenia PRS (see Figure 2). All models that included the clinical risk score were comparable in identifying 40% to 50% of suicide attempt cases within 6 months after ED discharge, indicating a higher sensitivity than the models only including individual disorder PRSs (see Multimedia Appendix 7). Specifically, shown in Figure 2, the baseline model had a higher PPV (26%-50%) than the other models when the sensitivity was in the 0.05 to 0.35 range. The models including the clinical risk score with or without PRSs had the same PPV (13%-26%) when the sensitivity was in the 0.4 to 1.0 range, and the model with the schizophrenia PRS alone had a lower PPV (12%-18%). AUPRC was 0.42 for the baseline model but reached 0.45 when the schizophrenia PRS was added, which is consistent with the observed improvement in AUC with the same model configuration.

Figure 2. A precision-recall curve for predicting suicide attempt within 6 months after an ED discharge. AUPRC: area under the precision-recall curve; ED: emergency department; PRS: polygenic risk score; SCZ: schizophrenia.



Discussion

Principal Findings

We found modest evidence suggesting that the integration of the PRS for schizophrenia (but the PRSs for not the other related phenotypes) might enhance the prediction of short-term risk for suicide attempt in patients discharged from the ED; both the AUC and AUPRC were numerically, although not significantly, higher when the schizophrenia PRS was added to the baseline clinical model. The improved predictive performance is likely explained by the higher heritability and statistical power of the schizophrenia PRS compared to the other PRSs examined in this study (see Multimedia Appendix 8). However, while heritability provides a compelling explanation, it does not fully account for the schizophrenia findings, as the predictive power of PRSs is also influenced by factors such as genetic architecture and heterogeneity in phenotype ascertainment. Furthermore, given the high dimensionality of the phenotypic features in the suicide prediction model, the addition of 1 or more PRSs is expected to have only a modest effect on prediction accuracy.

Limitations

Nevertheless, the nonsignificant improvement in performance we observed should be interpreted in light of our limited study sample size and statistical power of neuropsychiatric PRSs. Of the PRSs we examined, only the schizophrenia PRS was well powered (88%) to detect an association with suicide attempt in the holdout sample.

Future Work

Future studies utilizing larger biobank samples will enable a more robust and well-powered evaluation of the potential utility of PRSs in enhancing patient risk stratification in high-risk clinical settings. For instance, larger samples could facilitate the training of separate, context-specific baseline models using

EHR and survey data from patients with schizophrenia or bipolar disorder, followed by the integration of the respective PRSs into each model. Such an approach would provide a more nuanced understanding of the clinical relevance of PRSs and their potential role in improving risk stratification and patient outcomes.

Conclusions

In conclusion, we did not observe a substantial benefit of adding psychiatric PRSs to EHR and survey-based prediction models of suicide attempt in an ED setting. Given the importance of optimizing risk stratification to inform suicide prevention, further studies in large, diverse samples are warranted to clarify the value of incorporating genomic risk factors.

Acknowledgments

This study would not be possible without the contributions of Mass General Brigham patients and Biobank participants. We would also like to thank the research coordinators and the Biobank study for their tremendous effort in participant recruitment and sample collection.

Authors' Contributions

YHL, YZ, and CJK were responsible for study design, execution, all statistical analyses, manuscript drafting, and critical discussions. JWS and RCK were responsible for study design, execution, drafting, and critical discussions and provided overall supervision. MKN collected and provided the data, and MVP contributed to statistical analysis. YCAF and TG were responsible for preprocessing and quality control of genotype data. TTM was responsible for providing the genome-wide association study summary statistics required to train the polygenic risk score for externalizing traits. All authors revised the paper critically for important intellectual content, commented on and approved the final manuscript, are accountable for all aspects of the work, and read and agreed to the published version of the manuscript.

Conflicts of Interest

MKN reports receiving royalties from authoring psychology textbooks from Macmillan and Pearson; receiving consulting fees from Microsoft Corp, the Veterans Health Administration, Cerebral, and for a legal case about suicide; and being an unpaid scientific advisor for Empatica and TalkLife. RCK reports being a consultant for Cambridge Health Alliance; Canandaigua VA Medical Center; Child Mind Institute; Holmusk; Massachusetts General Hospital; Partners Healthcare, Inc.; RallyPoint Networks, Inc.; Sage Therapeutics; and University of North Carolina, and having stock options in Cerebral Inc.; Mirah; PYM (Prepare Your Mind); Roga Sciences; and Verisense Health. JWS reports being a member of the Leon Levy Foundation Neuroscience Advisory Board and the Sensorium Therapeutics Scientific Advisory Board; receiving honoraria for internal seminars at Biogen Inc and Tempus Labs; receiving grants from a Harvard University subcontract during the conduct of the study; and being a principal investigator of a collaborative study of the genetics of depression and bipolar disorder sponsored by 23andMe, for which 23andMe provides analysis time as in-kind support but no payments. No other disclosures are reported.

Multimedia Appendix 1
Comparison of demographic and clinical characteristics of the study population relative to the original population in Nock et al (2022).
[DOCX File , 20 KB - bioinform_v5i1e58357_app1.docx]

Multimedia Appendix 2
Supplemental methods.
[DOCX File , 30 KB - bioinform_v5i1e58357_app2.docx]

Multimedia Appendix 3
A list of genome-wide association study summary statistics used for polygenic risk score calculation.
[DOCX File , 20 KB - bioinform_v5i1e58357_app3.docx]

Multimedia Appendix 4
Ensemble weights and cross-validated risk sorted in descending order of ensemble weights and risk.
[DOCX File , 28 KB - bioinform_v5i1e58357_app4.docx]

Multimedia Appendix 5
Demographic and clinical characteristics of the study population, stratified by train-holdout split.
[DOCX File , 3665 KB - bioinform_v5i1e58357_app5.docx]

Multimedia Appendix 6
Patient risk stratification accuracy from SuperLearner models in the holdout sample.
[DOCX File , 16 KB - bioinform_v5i1e58357_app6.docx]

Multimedia Appendix 7
Sensitivity and positive predictive value of the ensemble models predicting a suicide attempt within 6 months of emergency department discharge in the holdout sample based on the baseline model and the best-performing model.
[DOCX File , 19 KB - bioinform_v5i1e58357_app7.docx]

Multimedia Appendix 8
Power curves for univariate associations of 5 polygenic risk scores with suicide attempt in the holdout sample.
[DOCX File , 179 KB - bioinform_v5i1e58357_app8.docx]

References

1. Suicide data and statistics. Centers for Disease Control and Prevention. 2024 May 13. URL: https://www.cdc.gov/suicide/facts/data.html [accessed 2024-07-12]
2. Ahmedani BK, Westphal J, Autio K, Elsiss F, Peterson EL, Beck A, et al. Variation in patterns of health care before suicide: a population case-control study. Prev Med 2019 Oct;127:105796. [doi: 10.1016/j.ypmed.2019.105796] [Medline: 31400374]
3. Nock MK, Millner AJ, Ross EL, Kennedy CJ, Al-Suwaidi M, Barak-Corren Y, et al. Prediction of suicide attempts using clinician assessment, patient self-report, and electronic health records. JAMA Netw Open 2022 Jan 04;5(1):e2144373. [doi: 10.1001/jamanetworkopen.2021.44373] [Medline: 35084483]
4. Lennon NJ, Kottyan LC, Kachulis C, Abul-Husn NS, Arias J, Belbin G, et al. Selection, optimization and validation of ten chronic disease polygenic risk scores for clinical implementation in diverse US populations. Nat Med 2024 Feb;30(2):480-487 [FREE Full text] [doi: 10.1038/s41591-024-02796-z] [Medline: 38374346]
5. Boutin NT, Schecter SB, Perez EF, Tchamitchian NS, Cerretani XR, Gainer VS, et al. The evolution of a large biobank at Mass General Brigham. J Pers Med 2022 Aug 17;12(8):1323 [FREE Full text] [doi: 10.3390/jpm12081323] [Medline: 36013271]
6. Barak-Corren Y, Castro VM, Javitt S, Hoffnagle AG, Dai Y, Perlis RH, et al. Predicting suicidal behavior from longitudinal electronic health records. Am J Psychiatry 2017 Feb 01;174(2):154-162. [doi: 10.1176/appi.ajp.2016.16010077] [Medline: 27609239]
7. Barak-Corren Y, Castro VM, Nock MK, Mandl KD, Madsen EM, Seiger A, et al. Validation of an electronic health record-based suicide risk prediction modeling approach across multiple health care systems. JAMA Netw Open 2020 Mar 02;3(3):e201262 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.1262] [Medline: 32211868]
8. Ge T, Chen CY, Ni Y, Feng YCA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat Commun 2019 Apr 16;10(1):1776. [doi: 10.1038/s41467-019-09718-5] [Medline: 30992449]
9. van der Laan MJ, Polley EC, Hubbard AE. Super learner. Stat Appl Genet Mol Biol 2007;6:Article25. [doi: 10.2202/1544-6115.1309] [Medline: 17910531]

Abbreviations

AUC: area under the receiver operator curve
AUPRC: area under the precision-recall curve
ED: emergency department
EHR: electronic health record
***ICD-9*:** *International Classification of Diseases, Ninth Revision*
***ICD-10*:** *International Classification of Diseases, Tenth Revision*
MGB: Mass General Brigham
PPV: positive predictive value
PRS: polygenic risk score

Edited by E Uzun; submitted 14.03.24; peer-reviewed by A Bhattacharya, CA Brunette, T Lu; comments to author 05.07.24; revised version received 13.08.24; accepted 22.08.24; published 23.10.24.

Please cite as:
Lee YH, Zhang Y, Kennedy CJ, Mallard TT, Liu Z, Vu PL, Feng YCA, Ge T, Petukhova MV, Kessler RC, Nock MK, Smoller JW
Enhancing Suicide Risk Prediction With Polygenic Scores in Psychiatric Emergency Settings: Prospective Study
JMIR Bioinform Biotech 2024;5:e58357
URL: https://bioinform.jmir.org/2024/1/e58357
doi:10.2196/58357
PMID:39442166

Review

Exploring the Intersection of Schizophrenia, Machine Learning, and Genomics: Scoping Review

Alexandre Hudon[1,2,3], BEng, MD, PhD; Mélissa Beaudoin[4,5], MSc; Kingsada Phraxayavong[6], BSc; Stéphane Potvin[2,4], PhD; Alexandre Dumais[2,4,6,7], MD, PhD

[1]Department of psychiatry and addictology, Faculty of Medicine, Université de Montréal, Montréal, QC, Canada

[2]Centre de recherche de l'Institut universitaire en santé mentale de Montréal, Montréal, QC, Canada

[3]Institut universitaire en santé mentale de Montréal, Montréal, QC, Canada

[4]Department of psychiatry and addictology, Université de Montréal, Montréal, QC, Canada

[5]Faculty of Medicine, McGill University, Montréal, QC, Canada

[6]Services et Recherches Psychiatriques AD, Montréal, QC, Canada

[7]Institut nationale de psychiatrie légale Philippe-Pinel, Montréal, QC, Canada

Corresponding Author:
Alexandre Dumais, MD, PhD
Department of psychiatry and addictology
Université de Montréal
2900 Edouard Montpetit Blvd
Montréal, QC, H3T 1J4
Canada
Phone: 1 514 648 8461
Email: alexandre.dumais@umontreal.ca

Abstract

Background: An increasing body of literature highlights the integration of machine learning with genomic data in psychiatry, particularly for complex mental health disorders such as schizophrenia. These advanced techniques offer promising potential for uncovering various facets of these disorders. A comprehensive review of the current applications of machine learning in conjunction with genomic data within this context can significantly enhance our understanding of the current state of research and its future directions.

Objective: This study aims to conduct a systematic scoping review of the use of machine learning algorithms with genomic data in the field of schizophrenia.

Methods: To conduct a systematic scoping review, a search was performed in the electronic databases MEDLINE, Web of Science, PsycNet (PsycINFO), and Google Scholar from 2013 to 2024. Studies at the intersection of schizophrenia, genomic data, and machine learning were evaluated.

Results: The literature search identified 2437 eligible articles after removing duplicates. Following abstract screening, 143 full-text articles were assessed, and 121 were subsequently excluded. Therefore, 21 studies were thoroughly assessed. Various machine learning algorithms were used in the identified studies, with support vector machines being the most common. The studies notably used genomic data to predict schizophrenia, identify schizophrenia features, discover drugs, classify schizophrenia amongst other mental health disorders, and predict the quality of life of patients.

Conclusions: Several high-quality studies were identified. Yet, the application of machine learning with genomic data in the context of schizophrenia remains limited. Future research is essential to further evaluate the portability of these models and to explore their potential clinical applications.

(JMIR Bioinform Biotech 2024;5:e62752) doi:10.2196/62752

KEYWORDS

schizophrenia; genomic data; machine learning; artificial intelligence; classification techniques; psychiatry; mental health; genomics; predictions; ML; psychiatric; synthesis; review methods; searches; scoping review; prediction models

Introduction

Schizophrenia is a complex mental health disorder that can have a significant negative impact on patients' resilience, quality of life, and self-esteem [1]. Considering the heterogenous nature of schizophrenia, several fields of research, such as genomics, also use the terminology psychotic disorder spectrum to refer to schizophrenia-like disorders [2]. Furthermore, while untreated, this mental health condition can lead to violence and violent offending [3]. A recent review of the literature estimated that schizophrenia has the highest societal cost among all mental health diseases. Indeed, reports from 10 countries estimated schizophrenia-related costs per person per year to be around US $2004-$94,229, with considerable variability amongst countries [4]. Despite several treatments being available, such as antipsychotics (dopamine receptor antagonists and partial agonists), up to 20%-30% of patients will remain treatment-resistant, and further approaches, such as cognitive behavioral therapy, will be used as adjuncts [5-7]. Various studies have explored the diverging clinical presentations of patients with schizophrenia and developed complexity estimators to aid clinicians in understanding the neuropathological processes involved in this complex illness [8,9]. Among recent research, several key factors have been identified as being linked to the development of the disorder, such as the length of the first psychotic episode, hormonal variations, as well as the presence of negative symptoms [10]. Despite the current knowledge that early interventions can help in the prognosis of patients diagnosed with schizophrenia, no prediction model is used in clinical practice as they usually do not account for variance between individuals [11].

To account for this variance and the dimensional aspects of schizophrenia, there have been tremendous efforts to gather genomic data and in-depth knowledge of neurobiological aspects of this disorder [12]. The entirety of the genetic information contained in an organism's DNA is referred to as genomic data [13-15]. This comprises details on gene structure, function, and variation in addition to the nucleotide sequence (adenine, thymine, cytosine, and guanine) found in the genome [16]. Genomic data is used to research the genetic contributions to traits, diseases, and biological processes [17]. It includes a variety of genetic information, such as single nucleotide polymorphisms (SNPs), copy number variations (CNVs), and gene expression patterns [18]. Worldwide collaborations have resulted in genome-wide association studies (GWAS) in over 56,000 schizophrenia cases and 78,000 controls, which identified 270 distinct genetic loci and polygenic risk scores, which can currently explain around 7.7% of the variance in schizophrenia case-control status [19]. Despite over 300 studies on gene expression in schizophrenia conducted over the past 15 years, none has consistently identified specific genes that contribute to schizophrenia risk [20]. Due to the complexity of schizophrenia, novel approaches are essential to better understand its neurobiological basis and improve outcome predictions, as it involves a network of genetic, neural, behavioral, and environmental factors [21].

Among novel approaches, machine learning has been increasingly used in the latest decade for various applications in medicine [22]. Machine learning is a branch of artificial intelligence that deals with teaching computers how to learn from and make predictions or judgments based on data through the use of statistical models and algorithms [23,24]. It focuses on creating systems that, through experience, may naturally perform better on a given task without having to be specifically designed to do so [25]. Data used by machine learning algorithms are referred to as model features [26]. Recent advancements in the field of data science have demonstrated that precision and genomic medicine combined with artificial intelligence have the potential to improve patient health care [27]. Examples of such advancements are the possibility of conducting variant calling, genome annotation and variant classification, and phenotype-to-genotype correspondence by using machine learning algorithms [28]. While existing literature reviews have explored specific applications of machine learning using genomic data for schizophrenia, none, to our knowledge, have comprehensively examined the diverse uses of machine learning at the intersection of these three fields, which could enhance the understanding of schizophrenia, thereby justifying the necessity for a thorough literature review. [29,30]. By identifying the broader applications of machine learning in this context, this overview will help researchers and clinicians pinpoint gaps in current research and pave the way for future applications of machine learning in the study of schizophrenia using genomic data.

This study aims to identify the various applications of machine learning algorithms using genomic data in the field of schizophrenia. By examining these approaches, this research offers an initial exploration into the methods being investigated to address the complexities of schizophrenia, a significant yet challenging mental illness. Therefore, this scoping review aimed to provide a comprehensive overview of these applications, highlighting key areas for future development at the intersection of machine learning, genomic data, and schizophrenia, with the potential to enhance clinical approaches.

Methods

Search Strategies

A comprehensive scoping search was conducted to identify recent studies across several electronic databases, including MEDLINE (PubMed), Web of Science, PsycNet (PsycINFO), and Google Scholar, covering the period from 2013 to 2024. The review was conducted using the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines. The search strategy used both text words and MeSH (Medical Subject Headings) terms, focusing on schizophrenia (eg, "schizophrenia" or "schizophrenic"), genomic data (eg, "genes," "genetic," or "genomic"), and machine learning (eg, "artificial intelligence" or "machine learning"). These topics were selected to align with the study's objectives. Detailed search strategies are provided in Multimedia Appendix 1. The search methodology was developed by the corresponding author, with searches executed by AH and cross-validated by MB. No restrictions were applied regarding setting or geography. The PRISMA checklist is provided in Multimedia Appendix 2.

Study Eligibility

Studies were included based on the following criteria: (1) the population of interest consisted of patients diagnosed with schizophrenia or the study of schizophrenia, (2) the study used a machine learning approach, and (3) the machine learning model incorporated genomic data features to find specific outcomes. Studies were included regardless of whether they used a single algorithm or multiple algorithms. Excluded from consideration were unpublished literature and studies using artificial intelligence algorithms outside the scope of machine learning. Examples of artificial intelligence algorithms outside the scope of machine learning include search algorithms, expert systems that are not data-driven, and heuristic-based systems. Studies that used machine learning solely to reduce data from genomic datasets were excluded. The search was limited to sources in English and French. Gray literature was not included.

Data Extraction

Data extraction was performed using a standardized form in Microsoft Excel and was independently counter-verified for consistency and integrity by two authors (AH and MB). Any disagreements regarding the inclusion or exclusion of a study were mutually resolved by the authors. The systematically extracted information included authors, population (sample), primary uses (or intent) of the machine learning algorithms, types of genomic data, types of machine learning algorithm used, main model performances, and key outcomes identified.

Quality Assessment

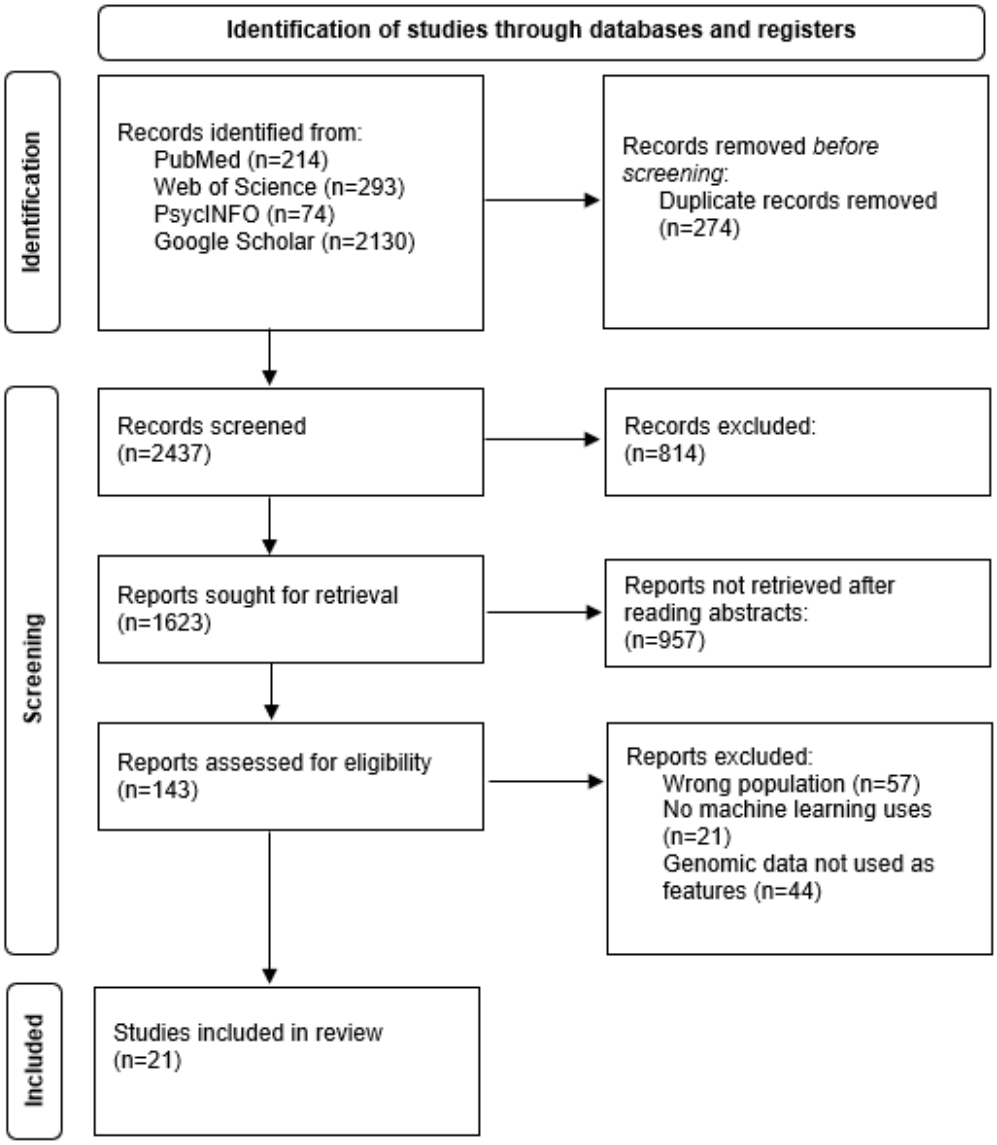
The quality of the identified studies was evaluated using the Newcastle-Ottawa Scale for nonrandomized controlled studies and the Cochrane Risk of Bias Tool for randomized controlled trials [31,32]. The Newcastle-Ottawa Scale is a tool used to assess the quality of cohort and case-control studies. It evaluates studies based on three main domains: selection of study groups, comparability of groups, and ascertainment of exposure or outcome [31]. Each domain includes specific criteria, and studies are awarded stars for meeting these criteria, with a maximum of 9 stars indicating the highest quality [31]. The Cochrane Risk of Bias Tool is a comprehensive framework used to assess the risk of bias in randomized controlled trials [32]. It evaluates 7 specific domains: random sequence generation, allocation concealment, blinding of participants and personnel, blinding of outcome assessment, incomplete outcome data, selective reporting, and other potential sources of bias [32]. Each domain is rated as having a low, high, or unclear risk of bias based on predefined criteria [32]. In this scoping review, studies with 1-4 stars on the Newcastle-Ottawa Scale or a high risk of bias by the Cochrane Risk of Bias Tool will be identified as low in quality, 4-6 stars as moderate, and 7-9 stars (or low risk of bias) as high.

Results

Description of Studies

The scoping review evaluated studies at the intersection of schizophrenia, genomic data, and machine learning. Initially, the literature search identified 2437 eligible articles after removing duplicates. A total of 814 studies were excluded based on a first analysis of the titles and abstract. Following a second round of abstract screening, 143 full-text articles were thoroughly assessed, with 122 subsequently excluded. This left 21 studies for detailed analysis. A flowchart illustrating the inclusion process is provided in Figure 1, and the specific details of the included studies are available in Multimedia Appendix 3. The studies meeting the inclusion criteria included various algorithms for different tasks. The most common application of machine learning was predicting schizophrenia using genomic data (n=10), followed by identifying features to enhance the understanding of schizophrenia (n=6), drug discovery for patients with schizophrenia (n=2), classifying schizophrenia amongst other mental health disorders (n=2), and predicting the quality of life and global functioning of patients with schizophrenia (n=1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart for the inclusion of studies.



Algorithms Used

Several algorithms have been identified in the 21 included studies. The algorithms the most frequently used were support vector machine classifiers (SVM; n=14), random forest (RF) algorithms (n=9), various implementations of neural networks (NN; n=7), and eXtreme Gradient Boosting (XGboost; n=5). Definitions of these popular algorithms are listed below:

- RF: It constitutes an ensemble learning technique. During training, it creates several decision trees and outputs the class, which is the average of the classes of each individual tree [33]. By merging the predictions of several trees, each trained on a different sample of the data, this method increases accuracy and helps avoid overfitting [33].
- SVM: It is an algorithm for supervised machine learning that is applied to regression and classification problems [34]. Finding the ideal hyperplane to divide the data into distinct classes is the fundamental notion behind SVM [34]. Different kernels (a function that quantifies the similarity between a pair of data points) can be used to enhance the performance of the SVM to best fit the data points [35].
- NN: These algorithms are modeled after the composition and operations of the human brain [36]. They are made up of networked layers of nodes, also called neurons, that process and change incoming data to create outputs [36].
- XGboost: It is founded on the gradient boosting principle, which entails building an ensemble of weak learners (usually decision trees) in a stepwise manner [37]. Every new tree seeks to fix the mistakes committed by the ones that came before it [37].

The remaining algorithms can be found in Multimedia Appendix 3.

Predicting Schizophrenia

Prediction of schizophrenia was identified as the main objective of 10 studies, all of which were deemed of high quality as per the Newcastle-Ottawa Scale ratings. The data used in these studies included differentially expressed genes, polygenic risk scores, genotype and human leukocyte antigen alleles, gene expression microarray data, single nucleotide polymorphisms, long non-coding RNAs, DNA methylation in blood, exomes, and G72 protein levels.

Li et al [38] used differentially expressed gene data from the Gene Expression Omnibus database, applying RF and SVM algorithms, and identified 15 key genes correlated with immune cell infiltration, achieving high diagnostic accuracy for schizophrenia with an area under the curve (AUC) of 0.77 in their test set. Another study, by Bracher-Smith et al [39], used data from the UK Biobank, applied machine learning algorithms such as least absolute shrinkage and selection operator, ridge-penalized logistic regression, SVM, RF, XGboost, NN, and stacked models, and found that while machine learning models incorporating polygenic risk scores and demographic factors showed good discrimination (AUC=0.71), they did not significantly outperform logistic regression in predicting schizophrenia. However, they reported that permutation features importance identified polygenic risk score for schizophrenia (PRS-SZ) as the most important predictor of schizophrenia [39].

Using data from the iPSYCH2012 case cohort, another study integrated genetics and registry data with a deep learning approach to stratify 19,636 patients with schizophrenia with or without major depressive disorder into clinically distinct subgroups characterized by unique disorder severities and comorbidity signatures, with predictive models achieving AUCs of 0.55 to 0.97, and therefore emphasized the importance of data-driven stratification for improving psychiatric diagnosis and prognosis [40]. Similarly, Qi et al [41] analyzed gene expression datasets from untreated schizophrenia patients and controls, identified 14 key gene probes, and used artificial NN to achieve diagnostic accuracy of 91.2% in training and 87.9% in testing and highlighted the potential of machine learning in identifying clinically useful biomarkers for schizophrenia. Another study introduced a sparse deep NN approach for identifying interpretable features for schizophrenia case–control classification using gray matter volume and single nucleotide polymorphism data, demonstrating slightly improved performance over traditional methods and highlighting key brain regions related to schizophrenia [42].

Studies with smaller sample sizes also reported several genomic data-enhanced methodologies to predict schizophrenia. Zhu et al [43] demonstrated that a machine learning model using the expression levels of 6 genes (*GNAI1*, *FYN*, *PRKCA*, *YWHAZ*, *PRKCB*, and *LYN*) in peripheral blood effectively distinguish schizophrenia patients from healthy controls, with the SVM model achieving the highest accuracy (AUC=0.993). Another study also reported the importance of long non-coding RNAs as they provided higher accuracy than coding genes in distinguishing schizophrenia from healthy controls [44].

Also focusing on predicting schizophrenia, a machine learning classifier based on DNA methylation in blood, specifically using correlated regions of systemic interindividual epigenetic variation (CoRSIV) regions and sparse partial least squares regression for discrimination analysis (SPLS-DA), effectively distinguishes schizophrenia patients from controls with a highly positive predictive value (PPV) of 80%, outperforming models based on polygenic risk scores (PRS) [45]. Another machine learning implementation used whole exome sequencing data to identify individuals at high risk for schizophrenia, achieving an accuracy of 85.7% with the XGBoost algorithm and providing further insights into the genetic basis of the disorder [46]. Finally, the last identified study used machine learning algorithms to demonstrate that G72 protein levels alone, without incorporating G72 genetic variations, are effective in distinguishing patients with schizophrenia from healthy controls with high specificity (0.9503) and sensitivity (0.8765) [47].

Identifying Features of Schizophrenia

A total of 6 included studies aimed at identifying features of schizophrenia or phenotyping using machine learning and genomic data, all of which were assessed as being of high quality. Feng et al [48] identified 6 candidate genes (*SFN*, *KDM5B*, *MYLK*, *IRF3*, *IRF7*, and *ID1*) with diagnostic significance for schizophrenia using machine learning on gene expression data. Another study by Zhu et al [49] attempted to identify immune-related biomarkers in peripheral blood in patients diagnosed with schizophrenia and reported that the mRNA expression of *CLIC3* was significantly decreased in the schizophrenia samples compared with the healthy controls. By using machine learning methods to analyze RNA sequencing data from the dorsolateral prefrontal cortex and amygdala in a postmortem investigation, Liu et al [50] aimed to identify driving biological signals representing schizophrenia. In doing so, they identified 18 genes added to known schizophrenia-associated pathways and expanded the gene network. These results provide a more comprehensive understanding of schizophrenia pathogenesis [50].

De Rosa et al [51] identified biological signals representing schizophrenia in brain tissues of the dorsolateral prefrontal cortex and hippocampus samples from postmortem brains of nonpsychiatric controls and patients with schizophrenia. Using an RF approach, they found 103 additional gene interactions were expanded to schizophrenia-associated networks, which were shared amongst both the dorsolateral prefrontal cortex and amygdala regions [51]. Another study by Feng and Shen [52] used neural networks using programmed cell-death-related genes as features and found 10 candidate hub genes (*DPF2*, *ATG7*, *GSK3A*, *TFDP2*, *ACVR1*, *CX3CR1*, *AP4M1*, *DEPDC5*, *NR4A2*, and *IKBKB*). Finally, a study on fresh frozen postmortem brain tissue aimed to identify DNA methylation patterns specific to patients with schizophrenia.

A cohort of 73 subjects diagnosed with schizophrenia and 52 control samples was analyzed using an unsupervised machine learning approach. As the results were not convincing, the authors reported that, if there are methylation changes associated with schizophrenia, they are diverse, complex, and have a small effect size [53].

Drug Discovery

A total of 2 studies reported the use of machine learning specifically for drug discovery (or related issues) for patients diagnosed with schizophrenia. Both of them were deemed of high quality. The first study focusing on 2307 patients with schizophrenia from the Chinese Antipsychotics Pharmacogenomics Consortium, 1379 from the Chinese Antipsychotics Pharmacogenetics Consortium, 275 healthy controls used several SVM and RF implementations and identified 6 risk genes for schizophrenia (*LINC01795*, *DDHD2*, *SBNO1*, *KCNG2*, *SEMA7A*, and *RUFY1*), which are involved

in cortical morphology and were identified as having genetic-epigenetic interactions linked to treatment response [54]. The other study, by Zhao and So [55], used the expression database ConnectivityMap that contains transcriptomic changes for *HL60*, *PC3*, and *MCF* over several machine learning implementations and reported that the predictive performance of their 5 approaches in cross-validation did not differ substantially, with SVM slightly outperforming the others while stating that repositioning hits are enriched for psychiatric medications considered in clinical trials [55].

Classifying Schizophrenia Among Other Mental Health Disorders

A total of 2 studies aiming at classifying schizophrenia amongst other mental disorders using machine learning were identified.

The first study by Yang et al [56] aimed at distinguishing schizophrenia from individuals with bipolar disorder, major depressive disorders, and healthy controls. To do so, the authors used differentially expressed genes from 268 individuals (67 patients with schizophrenia, 40 patients with bipolar disorder, 57 patients with major depressive disorders, and 104 healthy controls) over an SVM implementation that achieved an AUC of 0.96 for the schizophrenia group and of 0.71 for the independent set of the classification model. They reported that their model has a strong capacity to classify samples among multiple groups of mental illnesses [56]. Considering the opacity of the implementation, the quality was assessed as moderate for this study.

The other study, by Saardar et al [57], used the dbGaP database (schizophrenia) and the NDAR database (autism spectrum disorder) to compare whole exomes to differentiate between schizophrenia and autism using an XGboost model. They achieved an average validation accuracy of over 5 folds was 88% for both the single nucleotide variants-based model and gene-based model and reported that the ion transmembrane transport, neurotransmitter transport, and microtubule or cytoskeleton processes were of importance for schizophrenia [57]. The quality of this study was determined to be high based on our assessment.

Predicting Quality-of-Life and Global Functioning

Only one of the included studies focused on predicting the quality of life and global functioning of patients diagnosed with schizophrenia. This study was of high quality as per the quality assessment. Using data from 302 patients with schizophrenia in the Taiwanese population, Lin et al [58] compared a bagged ensemble of several machine learning algorithms to different permutations of these algorithms to predict functional outcomes of patients with schizophrenia. Their analysis revealed that the bagging ensemble algorithm with feature selection outperformed other predictive algorithms in forecasting the quality-of-life functional outcome of schizophrenia using the G72 rs2391191 and MET rs2237717 SNPs [58].

Discussion

Principal Results

This scoping review aimed to identify the different ways machine learning algorithms can be applied to genomic data in the study of schizophrenia. A total of 21 studies were fully analyzed, and 5 uses of machine learning algorithms on genomic data were identified: predicting schizophrenia, identifying features of schizophrenia, drug discovery, classifying schizophrenia amongst other mental health disorders, and predicting quality-of-life and global functioning. The studies were overall of high quality.

Comparison With Previous Work

The application of predictive models to forecast mental health disorders, such as schizophrenia, is gaining importance in medical research [59]. These models hold the potential to significantly assist clinicians in patient evaluation, particularly given the heterogeneity inherent to schizophrenia [60]. However, as observed in the identified studies, these models vary greatly in their implementation with diverging accuracy and validation methodologies. It is important to consider the implementation of these models as well as their accuracy and the techniques used to cross-validate the model, especially when using genomic data, as this could hinder their external validity [61]. The results found in the identified studies reinforce the premise that the genetic architecture of schizophrenia has proven to be very complex, heterogeneous, and polygenic and that a vast array of features could be integrated to improve predictive models [62]. Similarly, finding genomic-related risk factors of schizophrenia in such a model could help in distinguishing between this disease and other mental disorders, which may explain why classifying schizophrenia among other mental health disorders was one of the identified uses.

It is unsurprising that machine learning has been used to identify features of schizophrenia, as this has been done in other medical fields. Using candidate genes, it can be possible for clinicians to better understand common diseases and complex traits [63]. In psychiatry, psychiatric genomics is a rapidly advancing field that shows great promise for enhancing risk prediction, prevention, diagnosis, treatment selection, and the understanding of the pathogenesis of patients' symptoms [64]. As an example, some genes and functional genomic data linked to complex features of schizophrenia demonstrated that specific alleles may confer risk to the disorder by directly affecting synaptic function in adulthood [65].

As for drug discovery, literature reviews on the subject support that machine learning techniques can improve decision-making in pharmaceutical data across various applications [66,67]. It is also reported that combining machine learning techniques with genomic data has the potential to speed up the process and reduce failure rates in drug discovery and development [67]. This may explain why two studies focused specifically on schizophrenia in the context of drug discovery were identified. There is an increasing effort to develop pharmaceutical treatments, given the 20%-30% rate of treatment resistance observed in patients with this disorder [4].

Finally, quality-of-life assessment and functioning of patients with schizophrenia is trending in this field, which may explain why this use was identified in one study [68,69]. Another recent study on quality of life and genome-wide analyses of quality of life in psychosis, which used linear regression on 3684 participants (including 1119 psychosis patients), reported that numerous clinical and genetic associations with quality of life can be used in the daily care of these patients and enhance their overall well-being. These findings support the idea that more work should be conducted in this area in the future [70].

In the future, the information gathered by the use of machine learning in this area may provide the basis for more research projects. Through the identification of current knowledge gaps, scientists can narrow their attention and investigate novel genetic and biological markers that may have escaped their notice in the past in the development of machine learning models. This information may pave the way for the development of innovative therapeutic approaches, individualized treatment programs, and a better comprehension of the fundamental pathology of schizophrenia. To effectively handle the intricate problems presented by schizophrenia, machine learning techniques might need to be integrated with genomic data as they develop, and the genes identified in this review might help researchers select key features to enhance their mathematical models. This addition might lead to advancements in both basic science and therapeutic applications.

Limitations of This Study

This scoping review highlighted the various applications of machine learning algorithms using genomic data in the field of schizophrenia. Despite the relevance of this recension, it has a few limitations. The heterogeneity of diagnostic criteria for schizophrenia is a significant concern, as it is not addressed in half of the studies reviewed. Furthermore, the limited number of studies identified indicates the novelty of this field, necessitating future reviews to confirm findings. There is also a lack of external validation in samples differing from the training sample, such as those from different nationalities, raising questions about the generalizability of the results. Notably, no studies have concretely tested these algorithms in clinical settings, particularly for the prediction of schizophrenia, which remains an unmet need in the research. Due to the heterogeneity of the identified studies and the varying metrics used to assess precision and validate the machine learning models, performance comparisons were not conducted. Furthermore, studies on generic models using genomic data to predict overall mental health, rather than specifically focusing on schizophrenia, were excluded, as well as unpublished literature. This may have led to the omission of a small portion of relevant studies.

Conclusions

Considering the heterogeneity of clinical presentations observed in schizophrenia, genomic data combined with machine learning algorithms have been implemented to address several facets of this disorder. From the 21 studies analyzed, 5 main uses were identified: predicting schizophrenia, identifying schizophrenia features, discovering drugs, classifying schizophrenia amongst other mental health disorders, and predicting the quality of life of patients. These uses have potential implications as they could assist clinicians in providing a more personalized approach to their patients diagnosed with schizophrenia, considering the complexity of this diagnosis. There is still a limited amount of literature on the subject, and this study provides a first overview of machine learning applications of genomic data for schizophrenia. Future research is essential to further evaluate the portability of the models identified and their potential clinical applications.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Electronic search strategy for the scoping review conducted.
[DOCX File , 15 KB - bioinform_v5i1e62752_app1.docx]

Multimedia Appendix 2
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[PDF File (Adobe PDF File), 116 KB - bioinform_v5i1e62752_app2.pdf]

Multimedia Appendix 3
Systematic review study selection detailed results.
[DOCX File , 26 KB - bioinform_v5i1e62752_app3.docx]

References

1. Wartelsteiner F, Mizuno Y, Frajo-Apor B, Kemmler G, Pardeller S, Sondermann C, et al. Quality of life in stabilized patients with schizophrenia is mainly associated with resilience and self-esteem. Acta Psychiatr Scand 2016;134(4):360-367. [doi: 10.1111/acps.12628] [Medline: 27497263]
2. Cuthbert BN, Morris SE. Evolving concepts of the schizophrenia spectrum: a research domain criteria perspective. Front Psychiatry 2021;12:641319 [FREE Full text] [doi: 10.3389/fpsyt.2021.641319] [Medline: 33716834]

3. Fazel S, Gulati G, Linsell L, Geddes JR, Grann M. Schizophrenia and violence: systematic review and meta-analysis. PLoS Med 2009;6(8):e1000120 [FREE Full text] [doi: 10.1371/journal.pmed.1000120] [Medline: 19668362]
4. Kotzeva A, Mittal D, Desai S, Judge D, Samanta K. Socioeconomic burden of schizophrenia: a targeted literature review of types of costs and associated drivers across 10 countries. J Med Econ 2023;26(1):70-83 [FREE Full text] [doi: 10.1080/13696998.2022.2157596] [Medline: 36503357]
5. Efthimiou O, Taipale H, Radua J, Schneider-Thoma J, Pinzón-Espinosa J, Ortuño M, et al. Efficacy and effectiveness of antipsychotics in schizophrenia: network meta-analyses combining evidence from randomised controlled trials and real-world data. Lancet Psychiatry 2024;11(2):102-111. [doi: 10.1016/S2215-0366(23)00366-8] [Medline: 38215784]
6. Bighelli I, Çıray O, Salahuddin N, Leucht S. Cognitive behavioural therapy without medication for schizophrenia. Cochrane Database Syst Rev 2024;2(2):CD015332. [doi: 10.1002/14651858.CD015332.pub2] [Medline: 38323679]
7. Remington G, Hahn MK, Agarwal SM, Chintoh A, Agid O. Schizophrenia: antipsychotics and drug development. Behav Brain Res 2021;414:113507. [doi: 10.1016/j.bbr.2021.113507] [Medline: 34352293]
8. Fernández A, Gómez C, Hornero R, López-Ibor JJ. Complexity and schizophrenia. Prog Neuropsychopharmacol Biol Psychiatry 2013;45:267-276. [doi: 10.1016/j.pnpbp.2012.03.015] [Medline: 22507763]
9. Bassett DS, Nelson BG, Mueller BA, Camchong J, Lim KO. Altered resting state complexity in schizophrenia. Neuroimage 2012;59(3):2196-2207. [doi: 10.1016/j.neuroimage.2011.10.002] [Medline: 22008374]
10. Häfner H. From onset and prodromal stage to a life-long course of schizophrenia and its symptom dimensions: how sex, age, and other risk factors influence incidence and course of illness. Psychiatry J 2019;2019:9804836 [FREE Full text] [doi: 10.1155/2019/9804836] [Medline: 31139639]
11. Lee R, Leighton SP, Thomas L, Gkoutos GV, Wood SJ, Fenton SH, et al. Prediction models in first-episode psychosis: systematic review and critical appraisal. Br J Psychiatry 2022;220(Spec Iss 4 Themed Iss Precision Medicine and Personalised Healthcare in Psychiatry):1-13 [FREE Full text] [doi: 10.1192/bjp.2021.219] [Medline: 35067242]
12. Owen MJ. Genomic insights into schizophrenia. R Soc Open Sci 2023;10(2):230125 [FREE Full text] [doi: 10.1098/rsos.230125] [Medline: 36844807]
13. All of Us Research Program Genomics Investigators. Genomic data in the all of us research program. Nature 2024;627(8003):340-346 [FREE Full text] [doi: 10.1038/s41586-023-06957-x] [Medline: 38374255]
14. Middleton A, Milne R, Almarri MA, Anwer S, Atutornu J, Baranova EE, et al. Global public perceptions of genomic data sharing: what shapes the willingness to donate DNA and health data? Am J Hum Genet 2020;107(4):743-752. [doi: 10.1016/j.ajhg.2020.08.023] [Medline: 32946764]
15. Tatusova T. Update on genomic databases and resources at the national center for biotechnology information. Methods Mol Biol 2016;1415:3-30. [doi: 10.1007/978-1-4939-3572-7_1] [Medline: 27115625]
16. Eisenberg L. Are genes destiny? Have adenine, cytosine, guanine and thymine replaced lachesis, clotho and atropos as the weavers of our fate? World Psychiatry 2005;4(1):3-8 [FREE Full text] [Medline: 16633494]
17. Daniels H, Jones KH, Heys S, Ford DV. Exploring the use of genomic and routinely collected data: narrative literature review and interview study. J Med Internet Res 2021;23(9):e15739 [FREE Full text] [doi: 10.2196/15739] [Medline: 34559060]
18. Liu J, Zhou Y, Liu S, Song X, Yang X, Fan Y, et al. The coexistence of copy number variations (CNVs) and single nucleotide polymorphisms (SNPs) at a locus can result in distorted calculations of the significance in associating SNPs to disease. Hum Genet 2018;137(6-7):553-567. [doi: 10.1007/s00439-018-1910-3]
19. Legge SE, Santoro ML, Periyasamy S, Okewole A, Arsalan A, Kowalec K. Genetic architecture of schizophrenia: a review of major advancements. Psychol. Med 2021;51(13):2168-2177. [doi: 10.1017/s0033291720005334]
20. Merikangas AK, Shelly M, Knighton A, Kotler N, Tanenbaum N, Almasy L. What genes are differentially expressed in individuals with schizophrenia? A systematic review. Mol Psychiatry 2022;27(3):1373-1383 [FREE Full text] [doi: 10.1038/s41380-021-01420-7] [Medline: 35091668]
21. Haller CS, Padmanabhan JL, Lizano P, Torous J, Keshavan M. Recent advances in understanding schizophrenia. F1000Prime Rep 2014;6:57 [FREE Full text] [doi: 10.12703/P6-57] [Medline: 25184047]
22. Bohr A, Memarzadeh K. The rise of artificial intelligence in healthcare applications. Artificial Intelligence in Healthcare 2020:25-60. [doi: 10.1016/B978-0-12-818438-7.00002-2]
23. Lepakshi V. Machine learning and deep learning based aI tools for development of diagnostic tools. Computational Approaches for Novel Therapeutic and Diagnostic Designing to Mitigate SARS-CoV-2 Infection 2022:399-420. [doi: 10.1016/b978-0-323-91172-6.00011-x]
24. Lovis C. Unlocking the power of artificial intelligence and big data in medicine. J Med Internet Res 2019;21(11):e16607 [FREE Full text] [doi: 10.2196/16607] [Medline: 31702565]
25. Sarker IH. AI-based modeling: techniques, applications and research issues towards automation, intelligent and smart systems. SN Comput Sci 2022;3(2):158. [doi: 10.1007/s42979-022-01043-x] [Medline: 35194580]
26. Pudjihartono N, Fadason T, Kempa-Liehr AW, O'Sullivan JM. A review of feature selection methods for machine learning-based disease risk prediction. Front Bioinform 2022;2:927312 [FREE Full text] [doi: 10.3389/fbinf.2022.927312] [Medline: 36304293]

27. Quazi S. Artificial intelligence and machine learning in precision and genomic medicine. Med Oncol 2022;39(8):120 [FREE Full text] [doi: 10.1007/s12032-022-01711-1] [Medline: 35704152]
28. Dias R, Torkamani A. Artificial intelligence in clinical and genomic diagnostics. Genome Med 2019;11(1):70 [FREE Full text] [doi: 10.1186/s13073-019-0689-8] [Medline: 31744524]
29. Bracher-Smith M, Crawford K, Escott-Price V. Machine learning for genetic prediction of psychiatric disorders: a systematic review. Mol Psychiatry 2021;26(1):70-79 [FREE Full text] [doi: 10.1038/s41380-020-0825-2] [Medline: 32591634]
30. Del Fabro L, Bondi E, Serio F, Maggioni E, D'Agostino A, Brambilla P. Machine learning methods to predict outcomes of pharmacological treatment in psychosis. Transl Psychiatry 2023;13(1):75 [FREE Full text] [doi: 10.1038/s41398-023-02371-z] [Medline: 36864017]
31. Stang A. Critical evaluation of the newcastle-ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses. Eur J Epidemiol 2010;25(9):603-605. [doi: 10.1007/s10654-010-9491-z] [Medline: 20652370]
32. Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, et al. The cochrane collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011;343:d5928-d5928. [doi: 10.1136/bmj.d5928] [Medline: 22008217]
33. Rigatti SJ. Random forest. J Insur Med 2017;47(1):31-39. [doi: 10.17849/insm-47-01-31-39.1] [Medline: 28836909]
34. Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. Applications of support vector machine (SVM) learning in cancer genomics. Cancer Genomics Proteomics 2018;15(1):41-51 [FREE Full text] [doi: 10.21873/cgp.20063] [Medline: 29275361]
35. Noble WS. What is a support vector machine? Nat Biotechnol 2006;24(12):1565-1567. [doi: 10.1038/nbt1206-1565] [Medline: 17160063]
36. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to machine learning, neural networks, and deep learning. Transl Vis Sci Technol 2020;9(2):14 [FREE Full text] [doi: 10.1167/tvst.9.2.14] [Medline: 32704420]
37. Moore A, Bell M. XGBoost, a novel explainable aI technique, in the prediction of myocardial infarction: a UK biobank cohort study. Clin Med Insights Cardiol 2022;16:11795468221133611 [FREE Full text] [doi: 10.1177/11795468221133611] [Medline: 36386405]
38. Li Z, Li X, Jin M, Liu Y, He Y, Jia N, et al. Identification of potential biomarkers and their correlation with immune infiltration cells in schizophrenia using combinative bioinformatics strategy. Psychiatry Res 2022;314:114658. [doi: 10.1016/j.psychres.2022.114658] [Medline: 35660966]
39. Bracher-Smith M, Rees E, Menzies G, Walters JT, O'Donovan MC, Owen MJ, et al. Machine learning for prediction of schizophrenia using genetic and demographic factors in the UK biobank. Schizophr Res 2022;246:156-164 [FREE Full text] [doi: 10.1016/j.schres.2022.06.006] [Medline: 35779327]
40. Allesøe RL, Nudel R, Thompson WK, Wang Y, Nordentoft M, Børglum AD, et al. Deep learning-based integration of genetics with registry data for stratification of schizophrenia and depression. Sci Adv 2022;8(26):eabi7293 [FREE Full text] [doi: 10.1126/sciadv.abi7293] [Medline: 35767618]
41. Qi B, Boscenco S, Ramamurthy J, Trakadis YJ. Transcriptomics and machine learning to advance schizophrenia genetics: a case-control study using post-mortem brain data. Comput Methods Programs Biomed 2022;214:106590. [doi: 10.1016/j.cmpb.2021.106590] [Medline: 34954633]
42. Chen J, Li X, Calhoun VD, Turner JA, van Erp TGM, Wang L, et al. Sparse deep neural networks on imaging genetics for schizophrenia case-control classification. Hum Brain Mapp 2021;42(8):2556-2568 [FREE Full text] [doi: 10.1002/hbm.25387] [Medline: 33724588]
43. Zhu L, Wu X, Xu B, Zhao Z, Yang J, Long J, et al. The machine learning algorithm for the diagnosis of schizophrenia on the basis of gene expression in peripheral blood. Neurosci Lett 2021;745:135596. [doi: 10.1016/j.neulet.2020.135596] [Medline: 33359735]
44. Liu Y, Qu H, Chang X, Tian L, Qu J, Glessner J, et al. Machine learning reduced gene/non-coding RNA features that classify schizophrenia patients accurately and highlight insightful gene clusters. Int J Mol Sci 2021;22(7):3364 [FREE Full text] [doi: 10.3390/ijms22073364] [Medline: 33805976]
45. Gunasekara CJ, Hannon E, MacKay H, Coarfa C, McQuillin A, Clair DS, et al. A machine learning case-control classifier for schizophrenia based on DNA methylation in blood. Transl Psychiatry 2021;11(1):412 [FREE Full text] [doi: 10.1038/s41398-021-01496-3] [Medline: 34341337]
46. Trakadis YJ, Sardaar S, Chen A, Fulginiti V, Krishnan A. Machine learning in schizophrenia genomics, a case-control study using 5,090 exomes. Am J Med Genet B Neuropsychiatr Genet 2019;180(2):103-112. [doi: 10.1002/ajmg.b.32638] [Medline: 29704323]
47. Lin E, Lin CH, Lai YL, Huang CH, Huang YJ, Lane HY. Combination of genetic variation and G72 protein level to detect schizophrenia: machine learning approaches. Front Psychiatry 2018;9:566 [FREE Full text] [doi: 10.3389/fpsyt.2018.00566] [Medline: 30459659]
48. Feng Y, Shen J, He J, Lu M. Schizophrenia and cell senescence candidate genes screening, machine learning, diagnostic models, and drug prediction. Front Psychiatry 2023;14:1105987 [FREE Full text] [doi: 10.3389/fpsyt.2023.1105987] [Medline: 37113536]

49. Zhu X, Wang CL, Yu JF, Weng J, Han B, Liu Y, et al. Identification of immune-related biomarkers in peripheral blood of schizophrenia using bioinformatic methods and machine learning algorithms. Front Cell Neurosci 2023;17:1256184 [FREE Full text] [doi: 10.3389/fncel.2023.1256184] [Medline: 37841288]
50. Liu Y, Qu HQ, Chang X, Tian L, Glessner J, Sleiman PAM, et al. Expansion of schizophrenia gene network knowledge using machine learning selected signals from dorsolateral prefrontal cortex and amygdala RNA-seq data. Front Psychiatry 2022;13:797329 [FREE Full text] [doi: 10.3389/fpsyt.2022.797329] [Medline: 35386517]
51. De Rosa A, Fontana A, Nuzzo T, Garofalo M, Di Maio A, Punzo D, et al. Machine learning algorithm unveils glutamatergic alterations in the post-mortem schizophrenia brain. Schizophrenia (Heidelb) 2022;8(1):8. [doi: 10.1038/s41537-022-00231-1] [Medline: 35217646]
52. Feng Y, Shen J. Machine learning-based predictive models and drug prediction for schizophrenia in multiple programmed cell death patterns. Front Mol Neurosci 2023;16:1123708. [doi: 10.3389/fnmol.2023.1123708] [Medline: 36993785]
53. Torabi Moghadam B, Etemadikhah M, Rajkowska G, Stockmeier C, Grabherr M, Komorowski J, et al. Analyzing DNA methylation patterns in subjects diagnosed with schizophrenia using machine learning methods. J Psychiatr Res 2019;114:41-47 [FREE Full text] [doi: 10.1016/j.jpsychires.2019.04.001] [Medline: 31022588]
54. Guo LK, Su Y, Zhang YY, Yu H, Lu Z, Li W, et al. Prediction of treatment response to antipsychotic drugs for precision medicine approach to schizophrenia: randomized trials and multiomics analysis. Mil Med Res 2023;10(1):24. [doi: 10.1186/s40779-023-00459-7] [Medline: 37269009]
55. Zhao K, So HC. Drug repositioning for schizophrenia and depression/anxiety disorders: a machine learning approach leveraging expression data. IEEE J. Biomed. Health Inform 2019;23(3):1304-1315. [doi: 10.1109/jbhi.2018.2856535]
56. Yang Q, Xing Q, Yang Q, Gong Y. Classification for psychiatric disorders including schizophrenia, bipolar disorder, and major depressive disorder using machine learning. Comput Struct Biotechnol J 2022;20:5054-5064 [FREE Full text] [doi: 10.1016/j.csbj.2022.09.014] [Medline: 36187923]
57. Sardaar S, Qi B, Dionne-Laporte A, Rouleau GA, Rabbany R, Trakadis YJ. Machine learning analysis of exome trios to contrast the genomic architecture of autism and schizophrenia. BMC Psychiatry 2020;20(1):92. [doi: 10.1186/s12888-020-02503-5] [Medline: 32111185]
58. Lin E, Lin CH, Lane HY. Prediction of functional outcomes of schizophrenia with genetic biomarkers using a bagging ensemble machine learning method with feature selection. Sci Rep 2021;11(1):10179 [FREE Full text] [doi: 10.1038/s41598-021-89540-6] [Medline: 33986383]
59. Montazeri M, Montazeri M, Bahaadinbeigy K, Montazeri M, Afraz A. Application of machine learning methods in predicting schizophrenia and bipolar disorders: a systematic review. Health Sci Rep 2022;6(1):e962 [FREE Full text] [doi: 10.1002/hsr2.962] [Medline: 36589632]
60. Gashkarimov VR, Sultanova RI, Efremov IS, Asadullin AR. Machine learning techniques in diagnostics and prediction of the clinical features of schizophrenia: a narrative review. Consort Psychiatr 2023;4(3):43-53. [doi: 10.17816/CP11030] [Medline: 38249535]
61. McGaugh SE, Lorenz AJ, Flagel LE. The utility of genomic prediction models in evolutionary genetics. Proc Biol Sci 2021;288(1956):20210693 [FREE Full text] [doi: 10.1098/rspb.2021.0693] [Medline: 34344180]
62. Henriksen MG, Nordgaard J, Jansson LB. Genetics of schizophrenia: overview of methods, findings and limitations. Front Hum Neurosci 2017;11:322 [FREE Full text] [doi: 10.3389/fnhum.2017.00322] [Medline: 28690503]
63. Hirschhorn JN. Genetic approaches to studying common diseases and complex traits. Pediatr Res 2005;57(5 Part 2):74R-77R. [doi: 10.1203/01.pdr.0000159574.98964.87]
64. Ward ET, Kostick KM, Lázaro-Muñoz G. Integrating genomics into psychiatric practice: ethical and legal challenges for clinicians. Harv Rev Psychiatry 2019;27(1):53-64. [doi: 10.1097/hrp.0000000000000203]
65. Hall J, Bray NJ. Schizophrenia genomics: convergence on synaptic development, adult synaptic plasticity, or both? Biol Psychiatry 2022;91(8):709-717. [doi: 10.1016/j.biopsych.2021.10.018] [Medline: 34974922]
66. Dara S, Dhamercherla S, Jadav SS, Babu CM, Ahsan MJ. Machine learning in drug discovery: a review. Artif Intell Rev 2022;55(3):1947-1999 [FREE Full text] [doi: 10.1007/s10462-021-10058-4] [Medline: 34393317]
67. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov 2019;18(6):463-477 [FREE Full text] [doi: 10.1038/s41573-019-0024-5] [Medline: 30976107]
68. de Souza LA, Coutinho ES. The quality of life of people with schizophrenia living in community in rio de Janeiro, Brazil. Soc Psychiatry Psychiatr Epidemiol 2006;41(5):347-356. [doi: 10.1007/s00127-006-0042-6] [Medline: 16520882]
69. Beaudoin M, Hudon A, Giguère CE, Potvin S, Dumais A. Prediction of quality of life in schizophrenia using machine learning models on data from clinical antipsychotic trials of intervention effectiveness (CATIE) schizophrenia trial. Schizophrenia (Heidelb) 2022;8(1):29. [doi: 10.1038/s41537-022-00236-w] [Medline: 35314708]
70. Pazoki R, Lin BD, van Eijk KR, Schijven D, de Zwarte S, Guloksuz S, et al. Phenome-wide and genome-wide analyses of quality of life in schizophrenia. BJPsych Open 2020;7(1):e13. [doi: 10.1192/bjo.2020.140] [Medline: 33295273]

Abbreviations

AUC: area under the curve
CNV: copy number variation
CoRSIV: correlated regions of systemic interindividual epigenetic variation
GWAS: genome-wide association studies
MeSH: Medical Subject Headings
NN: neural networks
PPV: positive predictive value
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
PRS: polygenic risk scores
PRS-SZ: polygenic risk score for schizophrenia
RF: random forest
SNP: single nucleotide polymorphism
SPLS-DA: sparse partial least squares regression for discrimination analysis
SVM: support vector machine
XGboost: eXtreme Gradient Boosting

Edited by A Uzun; submitted 30.05.24; peer-reviewed by KK Raja, R Kilshaw, C König; comments to author 02.10.24; revised version received 06.10.24; accepted 16.10.24; published 15.11.24.

Please cite as:
Hudon A, Beaudoin M, Phraxayavong K, Potvin S, Dumais A
Exploring the Intersection of Schizophrenia, Machine Learning, and Genomics: Scoping Review
JMIR Bioinform Biotech 2024;5:e62752
URL: https://bioinform.jmir.org/2024/1/e62752
doi:10.2196/62752
PMID:39546776

Original Paper

Eco-Evolutionary Drivers of Vibrio parahaemolyticus Sequence Type 3 Expansion: Retrospective Machine Learning Approach

Amy Marie Campbell[1,2], PhD; Chris Hauton[1], PhD; Ronny van Aerle[2], PhD; Jaime Martinez-Urtaza[2,3], PhD

[1]School of Ocean and Earth Science, University of Southampton, Southampton, United Kingdom

[2]Centre for Environment, Fisheries and Aquaculture Science (CEFAS), Weymouth, United Kingdom

[3]Department of Genetics and Microbiology, Autonomous University of Barcelona, Barcelona, Spain

Corresponding Author:
Jaime Martinez-Urtaza, PhD
Department of Genetics and Microbiology
Autonomous University of Barcelona
Facultat de Biociènces, oficina C3/109
Campus de la UAB, Bellaterra
Barcelona, 08193
Spain
Phone: 34 93 581 2729
Email: jaime.martinez.urtaza@uab.cat

Abstract

Background: Environmentally sensitive pathogens exhibit ecological and evolutionary responses to climate change that result in the emergence and global expansion of well-adapted variants. It is imperative to understand the mechanisms that facilitate pathogen emergence and expansion, as well as the drivers behind the mechanisms, to understand and prepare for future pandemic expansions.

Objective: The unique, rapid, global expansion of a clonal complex of *Vibrio parahaemolyticus* (a marine bacterium causing gastroenteritis infections) named *Vibrio parahaemolyticus* sequence type 3 (VpST3) provides an opportunity to explore the eco-evolutionary drivers of pathogen expansion.

Methods: The global expansion of VpST3 was reconstructed using VpST3 genomes, which were then classified into metrics characterizing the stages of this expansion process, indicative of the stages of emergence and establishment. We used machine learning, specifically a random forest classifier, to test a range of ecological and evolutionary drivers for their potential in predicting VpST3 expansion dynamics.

Results: We identified a range of evolutionary features, including mutations in the core genome and accessory gene presence, associated with expansion dynamics. A range of random forest classifier approaches were tested to predict expansion classification metrics for each genome. The highest predictive accuracies (ranging from 0.722 to 0.967) were achieved for models using a combined eco-evolutionary approach. While population structure and the difference between introduced and established isolates could be predicted to a high accuracy, our model reported multiple false positives when predicting the success of an introduced isolate, suggesting potential limiting factors not represented in our eco-evolutionary features. Regional models produced for 2 countries reporting the most VpST3 genomes had varying success, reflecting the impacts of class imbalance.

Conclusions: These novel insights into evolutionary features and ecological conditions related to the stages of VpST3 expansion showcase the potential of machine learning models using genomic data and will contribute to the future understanding of the eco-evolutionary pathways of climate-sensitive pathogens.

(JMIR Bioinform Biotech 2024;5:e62747) doi:10.2196/62747

KEYWORDS

pathogen expansion; climate change; machine learning; ecology; evolution; vibrio parahaemolyticus; sequencing; sequence type 3; VpST3; genomics

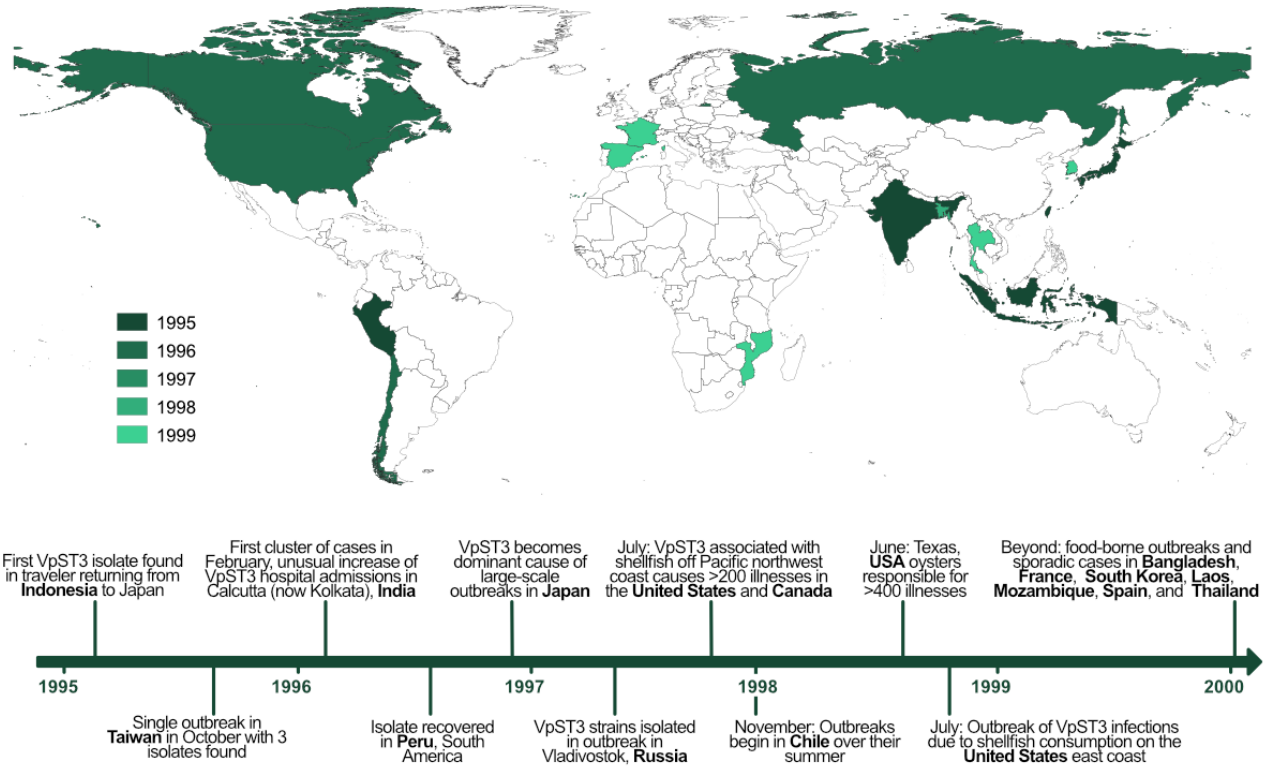
Introduction

Background

Climate change is likely to impact environmentally sensitive pathogens in terms of shifts in seasonality, expansion of suitable habitats, and the emergence and global dispersal of well-adapted variants. This has already been observed for *Vibrio parahaemolyticus* [1], a marine bacterium inhabiting coastal waters that causes acute gastroenteritis when transmitted to humans by ingestion of contaminated seafood, contributing to a large percentage of foodborne infections worldwide. Recent decades have seen this highly adaptable bacterium spread globally and increasingly cause outbreaks [1].

Before the 1990s, *Vibrio* infections were considered an exotic outcome of travel to Asia, where *Vibrio* bacteria were historically considered endemic. Up to this point, only particular strains of *Vibrio cholerae* had been designated as epidemic variants, characterized by global expansion and pandemic potential. However, transcontinental spread has now been reported for 2 *V parahaemolyticus* clonal types: sequence type 3 from Southeast Asia and, more recently, sequence type 36 from the Pacific Northwest [2,3]. The first of these instances, involving the clonal type *Vibrio parahaemolyticus* sequence type 3 (VpST3), was identified in 1996, when the unique variant, which had not been previously reported, was found to be responsible for up to 80% of the cases in a notable increase of *V parahaemolyticus* infections in Calcutta (now Kolkata), India, in 1996 [4]. This outbreak was unusual, with all recovered isolates clustered into a single homogeneous group, unlike previous outbreaks [3]. Similar isolates were then observed from outbreaks in distinct locations around the world, including Peru, Japan, Russia, Chile, and the United States [5-8] (Figure 1 [4-15]), where the variant was emerging concurrently. This included regions with conditions previously considered adverse for the presence of such pathogens. The epidemic radiations that followed in these diverse regions were the first observed for *V parahaemolyticus* and resulted in the variant supplanting local populations and rapidly becoming the most dominant *V parahaemolyticus* variant globally. As a consequence of this expansion, *V parahaemolyticus* became the second human pathogenic *Vibrio* species with an epidemic nature and, along with *V cholerae*, the only *Vibrio* species with strains capable of worldwide expansions and causing infections at a global level.

Figure 1. Timeline and map of *Vibrio parahaemolyticus* sequence type 3 (VpST3) initial expansion based on reported isolates and outbreaks in the literature (shapefile provided by Database of Global Administrative Areas).



This process of expansion results from an epidemic bacterial population structure, as described in the study by Smith et al [16]. Upon a background of numerous rapidly recombining genotypes, a limited number of very frequent genotypes are superimposed, known as clonal complexes, that have originated from highly adaptive ancestral genotypes [16]. The mechanisms behind the rise of these clonal complexes are largely unknown; yet, it is imperative to identify the conditions that allow a pathogen to emerge in such diverse locations and become dominant, as well as the drivers behind these processes, to understand and prepare for future pandemic expansions. When considering environmentally sensitive pathogens such as *Vibrio*, possible drivers can be categorized as either ecological or evolutionary. Evolutionary drivers include the processes of adaptation, mutations that increase fitness, or the uptake or horizontal transfer of beneficial accessory genes. Both ends of

the spectrum of genetic diversity—generalists and specialists—are associated with pandemic expansions. Ecological drivers can range from the local environment, which affects pathogen survival and growth, to environmental corridors and transport mechanisms. Importantly, these ecological and evolutionary drivers are not exclusive and, instead, interact significantly, with this interplay known as "eco-evolutionary." A key example of this would be adaptative selection occurring after arrival to a distinct marine environment. While more studies are considering the coeffect of ecological and evolutionary factors on larger species (such as vertebrates and invertebrates), little attention has been paid to environmentally sensitive pathogens. Focusing on these pathogens would provide novel insights into how particular pathogenic strains emerge [17].

V parahaemolyticus is a uniquely placed species, with a history of pandemic expansion that facilitates the study of such eco-evolutionary drivers. First, *Vibrio* are phylogenetically diverse with highly variable genomic backgrounds shaped by recombination and horizontal gene transfer [3], from which specialized variants can emerge. Second, *V parahaemolyticus* exhibits well-characterized environmental thresholds and tolerances, rapidly responding to changes in its marine environment, such as water temperature [18-23] and salinity [18,21,24-26]. Notably, anomalously high temperatures were observed between 1996 and 1998 [27,28] around the emergence of VpST3—pertinent amid the preference of *Vibrio* for warmer waters. From a combined eco-evolutionary perspective, *Vibrio* have high genome plasticity, which facilitates rapid adaptation in response to environmental changes [29], resulting in a large diversity of causative strains and resulting infection dynamics [30,31]. It would be simplistic to assume that all these diverse *V parahaemolyticus* variants respond to environmental change homogeneously, opening up the eco-evolutionary response landscape for exploration.

Study Objectives

We reconstructed this global expansion using publicly available genome sequences of VpST3 from clinical and environmental sources, isolated from around the world over the period of expansion of this clone, to identify population structure and demographic shifts indicative of the different stages of expansion, including emergence and establishment. We investigated the possible drivers of the expansion and our ability to predict the dynamics of VpST3 by testing a range of evolutionary and ecological drivers in a combined approach using machine learning models to elucidate the complex mechanisms that, when combined, may facilitate such a rapid, global expansion. Machine learning has been credited for its ability to harness the predictive power of evolution, using pattern recognition to uncover complex associations between biological processes [32]; therefore, it is well-placed for the novel exploration of interacting eco-evolutionary mechanisms in combination. Understanding the evolutionary features and ecological conditions related to the stages of pathogen expansion is a crucial step in understanding future eco-evolutionary pathways of climate-sensitive pathogens.

Methods

The protocol for preprocessing evolutionary and ecological data into a data frame for machine learning analysis is summarized in a graphical representation in Figure S1 in Multimedia Appendix 1.

Ethical Considerations

Ethical approval was not sought for the present study because it consisted of neither human nor animal experimentation and all genomic datasets used had been previously made publicly available with reference to their ethical approval in the papers associated with these submissions.

Genomic Data

Raw sequencing datasets from a collection of 311 VpST3 isolates, representing a range of geographic areas, were acquired from public databases for genomic analyses (Table S1 in Multimedia Appendix 1). The 311 isolates covered a temporal range from 1996 to 2021, with 162 (52.1%) from Asia, 78 (25.1%) from North America, and 71 (22.8%) from South America. A series of genetic markers were used to confirm that the isolates were VpST3 using multilocus sequence typing in MLST (version 2.11) [33]. Our analysis was restricted to isolates that were submitted with accompanying isolation date and location details because such metadata were required for the downstream linkage with environmental variables. The raw sequences were processed using default parameters within Bactopia (version 2.0.2) [34], including quality filtering, assembly, and annotation. Core single nucleotide polymorphisms (SNPs) were identified across all sequences using parsnp (version 1.5.6) [35] to create a core genome alignment, mapped to the *V parahaemolyticus* reference genome RIMD2210633. Gubbins (version 3.1.6) [36] was used to remove recombining regions to provide a final nonrecombining core genome alignment.

Phylogenetic Analysis

TempEst (version 1.5.3) [37] was used to confirm a temporal signal and conformation to a molecular clock, followed by BEAST2 (version 2.7.6) [38] analysis to reconstruct the global phylogenetic dynamics of *V parahaemolyticus*, using BEAUti [39] and a structured coalescent within a MultiTypeTree template [40]. After sensitivity analyses on a range of models, the selected model used a relaxed log normal clock model and a general time reversible (GTR) substitution model, with a normal distribution substitution rate prior. The tip dates and discrete location attributes were used to situate the genomic evolution in space and time. The Markov chain Monte Carlo was run for 250 million states until all outputs converged (effective sample size >200), confirmed by Tracer (version 1.7.1) [41]. The final maximum clade credibility tree was generated using TreeAnnotator within BEAST2.

Encoding of Expansion Dynamics

The Bayesian phylogenetic analysis and subsequent tree structure informed the designation of a variety of classifications representing VpST3 dynamics. These classifications included populations within the collection, temporal divergence, the

success of introductions, and the stages of expansion. We assigned each of these classifications to each of the 311 isolates, using set criteria applied to the phylogenetic tree (Textbox 1), and converted them into binary or categorical features to provide target variables for machine learning analysis.

Textbox 1. Criteria for encoding the numerical and categorical variables of expansion dynamics. The terms in parentheses refer to the column names of the expansion dynamics in the data frame input.

Populations (wave)

Populations within the *Vibrio parahaemolyticus* sequence type 3 (VpST3) collection were identified using TreeStructure (version 0.1.0) [42], which identifies genealogical patterns to infer population structure from time-scaled phylogenies by performing 100,000 tree simulations with a significance threshold set at $P<.001$

Temporal evolution (earlylate)

Very few VpST3 isolates were recovered in 2003, after the initial global population expansion; therefore, we specified this year as a split between the early colonizers found before this date and the later isolates recovered after the expansion

Success (success)

Failed introductions were monophyletic branches that did not split into further nodes in the phylogenetic tree, while successful introductions were those that saw downstream nodes in the same reported location

Stages of expansion (stages)

We split expansion into five defined stages: (1) initial introduction (the first node or nodes in a clade), (2) established population (the nodes in the clade after this introduction), (3) secondary introduction (the first node or nodes in a clade in a new location from the original introduction), (4) secondary establishment (the nodes in the clade after this introduction in the new location), and (5) bottleneck (the last node of a clade or a location within the clade)

Stages of expansion: binary (stages_binary)

A simplified version of the previous stages of expansion classification, reducing it to a binary classification of introduced isolates (the first instances in a clade or location) and established isolates (those that followed these introductory nodes within this clade)

Extraction of Evolutionary Driver Data

Genomic analysis was used to extract features representing possible evolutionary drivers for each isolate. Quantifying the gene content variation in the accessory pangenome in natural populations is important to understand the plasticity and adaptability of populations to environmental perturbations [43]. To obtain a metric of total genes present in each isolate, we used Roary (version 3.13.0) [44] to construct the pangenome and annotate each gene present in each isolate. We used Scoary (version 1.6.16) [45] to identify shell genes (present in 15%-95% of the population) whose presence was statistically associated ($P<.01$) with the previously assigned labels representing introduction, establishment, or success. We retained a selection of these that were common accessory genes (with a presence ranging from 5% to 95% across the isolates in the collection), annotated their function, and generated features representing the binary presence or absence features. We used single-likelihood ancestor counting within HyPhy (version 2.5.48) [46] to estimate the ratio of nonsynonymous to synonymous substitutions (dN/dS) and identify sites under significant diversifying or purifying selection ($P<.05$) in the genes of interest.

SNP mutations of relevance to the expansion process were selected using pcadapt (version 4.3.3) [47] for outlier detection based on population structure. The outliers were inferred based on principal component analysis, using the parameter K=2 and a desired false discovery rate threshold of 0.1 (q-threshold) to identify discriminatory SNP mutations associated with local adaptation. These SNPs were annotated to predict functional effects on genes using SnpEff (version 5.1) [48] and the *V parahaemolyticus* RIMD2210633 genome annotations as a reference. SNPs predicted to have nonsynonymous missense variants were retained for downstream analyses. We recorded the base found at this position for every isolate to assess whether these mutations would help the model define the evolutionary classification. To convert these into numerical values fit for machine learning applications, we reclassified the letters representing bases into numbers (A=1, C=2, T=3, G=4, and N-polymorphic=5).

Extraction of Ecological Driver Data

Time series data for sea surface temperature (SST) and salinity—2 of the most well-reported environmental drivers of *V parahaemolyticus* in the marine environment [49] —were acquired from the European Centre for Medium-Range Weather Forecasts Reanalysis version 5 [50] and the Met Office Hadley Centre's EN4.2.2 quality-controlled ocean dataset [51,52], respectively, covering the period from 1995 to 2021. We zonally extracted the climate time series data for the country of isolate recovery, using Database of Global Administrative Areas country zones provided as shapefiles, extending into coastal waters by 2 decimal degrees to extract the local conditions of the marine environment. Although the climate data were available at a monthly resolution, the majority of the genomic isolates only contained an annual resolution. Instead of averaging across the whole year, we created metrics for maximum, minimum, and mean values for each season across the year, alongside generated lagged variables from the previous year. Alongside environmental drivers, the seafood industry, including fisheries [53], seafood consumption and trade [54], and fish market contamination [55], has been previously hypothesized as a possible mechanism for the emergence and spread of *Vibrio* bacteria. We therefore extracted shellfish import

data for each country from the FishStatJ database of the Food and Agricultural Organization of the United Nations as annual totals measured in 100 kg of net product weight [56] to explore the potential of this driver.

Machine Learning Approach

We combined the ecological and evolutionary driver metrics and the classification for each of the criteria into a single data frame for each isolate, with a total of 311 data points.

For our machine learning analysis, we chose a random forest classifier model, an ensemble learning method that uses bootstrapping across decision tree classifiers, due to its high interpretability and implemented the models using the Python module *scikit-learn* (version 1.3.0) [57]. For each expansion dynamic, we created three separate models: 1 model used only the evolutionary drivers as features, a second only ecological drivers as features, and a final model used both ecological and evolutionary drivers in a combined approach. We trained the random forest classifier (using 100 estimators, setting the maximum number of features to consider for best split to the square root of the total number of features, and using bootstrap samples to build trees) on a randomly selected subset of 70.1% (218/311) of the data, retaining the remaining 29.9% (93/311) as an unseen test dataset. In total, there were 109 features used to predict each evolutionary dynamic (Table S2 in Multimedia Appendix 1), of which 60 (55%) represented evolutionary drivers, and 49 (45%) represented ecological drivers. The classification output classes were either binary or categorical based on the expansion dynamic being predicted.

To test the accuracy of these predictions and provide insight into our ability to predict the expansion dynamics of VpST3, we reported 4 accuracy metrics, both per class and across all predictions, when the models were applied to the unseen test data. These metrics included precision (positive prediction rate, affected by false positives), recall (sensitivity rate, affected by false negatives), the F_1-score (a harmonic mean of precision and sensitivity, often used for comparative machine learning performance assessments), and overall accuracy (taking into account all components). We calculated the feature importance for all ecological and evolutionary drivers involved in each model using the Gini importance attribute within the random forest implementation in *scikit-learn* (version 1.3.0) (57), which is computed by the mean and SD of the accumulating impurity decrease within each tree due to the addition of each specific feature.

To assess the collinearity effects from cross-correlations between the ecological and evolutionary driver metrics contained in our model, we calculated the Spearman rank correlation coefficient between the driver variables. During model development of the individual ecological and evolutionary models, we selected features that did not exhibit significant (P<.05) collinearity. However, collinearity between evolutionary and ecological features in the combined model was explored, rather than omitted, to gain greater insight into potential eco-evolutionary associations. These significant relationships were visualized in a heat map using the *seaborn* (version 0.12.2) Python library [58].

Region-Specific Analysis

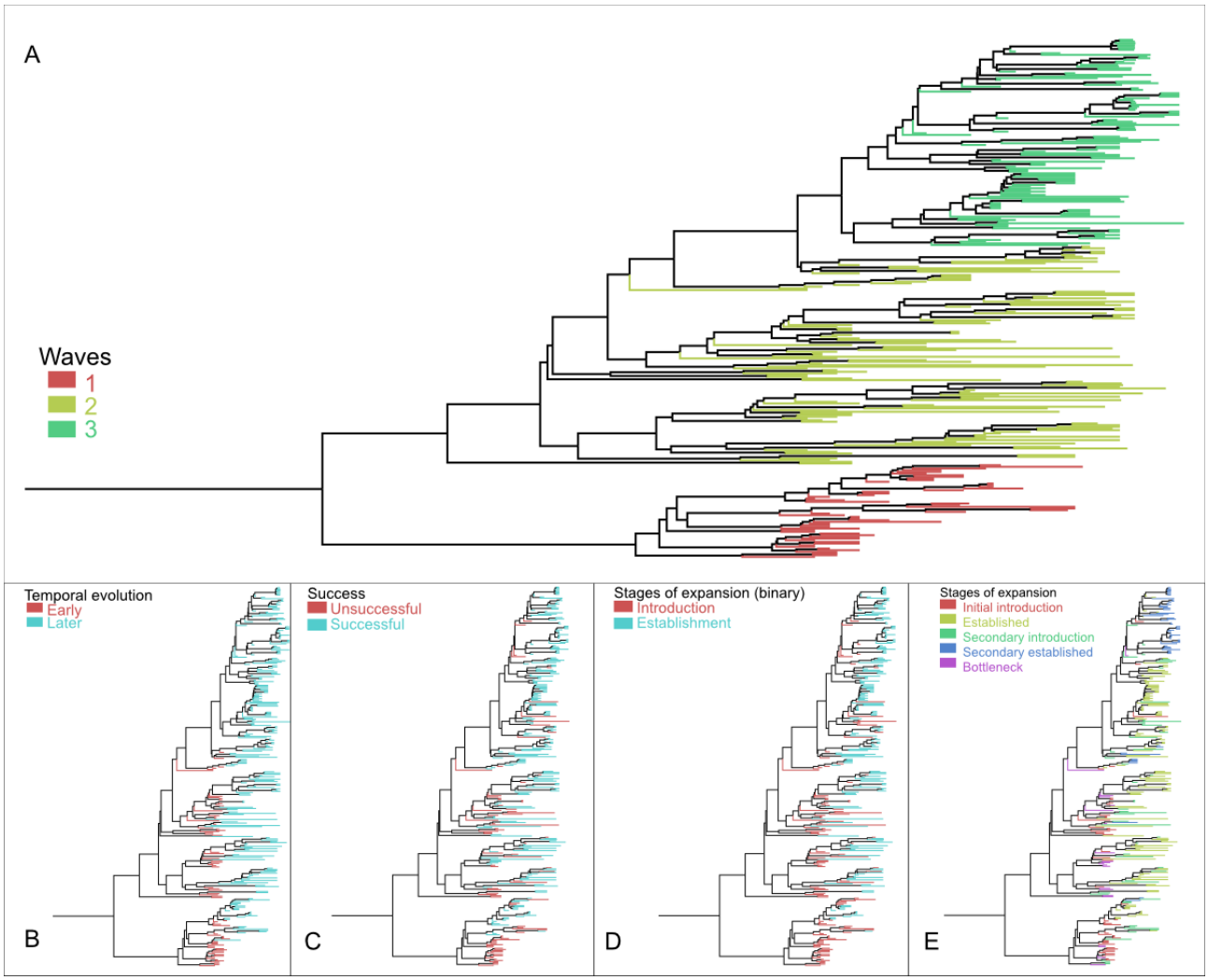
To explore the potential to understand the successful expansion of VpST3 in particular regions, we developed 2 region-specific models representing an endemic area and an area where VpST3 emerged: China and Peru, respectively. These regions were chosen because they reported the most VpST3 isolates within their respective continents and consist of distinct geographic characteristics to establish whether a regional focus on an area with specific local conditions to drive eco-evolutionary dynamics improves our ability to predict successful expansions. These models were trained and tested on regional subsets of the original data frame, with the same model parameters and features.

Results

Phylogeny Characterization

The phylogeny revealed an evolutionary population structure within the VpST3 genomes, with multiple introductions into geographically distinct locations, including secondary migrations and introductions. Our phylogenetic analysis found 3 clear population “waves” within VpST3, comprising 56 (18%), 131 (42.1%), and 124 (39.9%) of the 311 isolates (Figure 2A). In terms of temporal evolution, of the 311 isolates, 73 (23.5%) were classified as early colonizers (before 2003), and 238 (76.5%) were isolated after the initial expansion after 2003 (Figure 2B). Regarding expansion success, of the 311 isolates, 86 (27.7%) were classified as unsuccessful and 225 (72.3%) as successful (Figure 2C). With regard to the stages of expansion, of the 311 isolates, under a binary classification, 121 (38.9%) were classified as introduced, with 190 (61.1%) being classified as established (Figure 2D). When this was scaled up to the 5 stages of classification, of the 311 isolates, 49 (15.8%) were classified as initial introductions, 131 (42.1%) as established, 52 (16.7%) as secondary introductions, 58 (18.6%) as secondarily established, and finally 21 (6.8%) as representing bottlenecked populations (Figure 2E).

Figure 2. Maximum clade credibility tree with branches color coded by expansion dynamic metrics: (A) population waves, (B) temporal evolution, (C) expansion success, and (D and E) stages of expansion.



Evolutionary Features Extracted

We detected 194 potential adaptive SNP outliers within the collection of genomic isolates, of which 44 (22.7%) were predicted to be missense variants, altering an amino acid within a protein, with predicted moderate effects on particular genes (Table S3 in Multimedia Appendix 1). These SNPs were chosen as evolutionary features for the machine learning analysis. Overall, the total number of genes in each isolate ranged from 4292 to 4735, with no clear temporal signal (R^2=0.08). We identified 400 accessory genes present in 15% to 95% of the entire VpST3 collection and reduced these to 15 (3.8%) genes of interest as evolutionary features for the machine learning analysis (Table 1). This selection was based on genes that were associated with particular expansion metrics; the presence of all 15 selected genes was significantly associated (P<.01) with the binary classification delineating introduced and established isolates, and 5 (33%) were further associated with the successful classification metric. Annotation of these genes of interest found that most (n=5, 33%) were functionally associated with survival in the environment and tolerance to environmental conditions (Table 1). In addition, some of them (n=8, 53%) were involved in bacterial transport mechanisms, such as putrescine pathways, that promote biofilm formation. On 4 occasions, 2 versions of a gene with a similar function were identified within this group—for *pilT*, *ttcA*, CARB β-lactamase, and DeoR family transcriptional regulators. Of these 15 accessory genes, no evidence for positive diversifying selection was found (as determined by HyPhy single-likelihood ancestor counting [46]); however, 10 (67%) genes had evidence of negative, purifying selection (P value threshold <.10), ranging from 1 to 17 sites under purifying selection.

Table 1. Significant associations identified between accessory gene presence and key expansion dynamics.

Annotation	Function	Significance of presence association with classification labels, *P* value		
		Introduction	Establishment	Success
Lactoylglutathione lyase	Enzyme used for methylglyoxal detoxification, contributes to bacterial survival in the environment [59]	<.001	<.001	<.001
HTH[a]-type transcriptional regulator (*puuR*)	Recombinant protein, involved in putrescine pathways [60]	<.001	<.001	<.001
RNA polymerase sigma factor (*RpoS*)	Proteins that regulate transcription in bacteria, activated in response to different environmental conditions	<.001	<.001	—[b]
Type IV pilus twitching motility protein (*pilT*)	Involved in the transport (motility) of the bacteria itself, biofilms, and virulence [61]	<.001	<.001	—
Sodium:proton antiporter	Antiporters (in this case moving sodium ions in or out of a cell) play an important role in tolerance to salt stress [62]	<.001	<.001	<.001
N-carbamoylputrescine amidase (*aguB*)	Involved in biofilm production by converting N-carbamoylputrescine to putrescine [63]	<.001	<.001	<.001
Agmatine deiminase (*aguC*)	Involved in a putrescine pathway [64]	<.001	<.001	<.001
DeoR family transcriptional regulator	Primarily drives the sensing of environmental stimuli and life cycle responses [65]	<.001	<.001	—
Type IV pilus twitching motility protein (*pilT*)	Involved in the transport (motility) of the bacteria itself, biofilms, and virulence [61]	<.001	<.001	—
Carbenicillin-hydrolyzing class A beta-lactamase CARB-23	Expresses β-lactamase for resistance to antibiotic penicillins [66]	<.001	<.001	—
Ribonuclease III (*rnc*)	Modulates pathogenicity: motility, invasiveness, biofilm formation ability, and virulence [67]	<.001	<.001	—
tRNA 2-thiocytidine(32) synthetase (*ttcA*)	Involved in bacterial growth, resistance to biocides, biofilm formation, and swimming motility [68]	<.001	<.001	—
Carbenicillin-hydrolyzing class A beta-lactamase CARB-23	Expresses β-lactamase for resistance to antibiotic penicillins [66]	<.001	<.001	—
DeoR family transcriptional regulator	Primarily drives the sensing of environmental stimuli and life cycle responses [65]	<.001	<.001	—
tRNA 2-thiocytidine(32) synthetase (*ttcA*)	Involved in bacterial growth, resistance to biocides, biofilm formation, and swimming motility [68]	<.001	<.001	—

[a]HTH: helix-turn-helix.

[b]Not applicable.

Predictive Power

Overall accuracies for the different expansion metrics ranged from 0.722 to 0.967 for models using a combined eco-evolutionary approach (Figure 3). In our analysis, a combined eco-evolutionary approach almost always improved the accuracy of predicting expansion dynamics compared to using evolutionary or ecological drivers in isolation (Table 2). This was notably apparent for the predictions of the population structure within the phylogeny, in terms of the identification of 3 clear groups, in which evolutionary and ecological features individually produced accuracies of 0.733 and 0.744, respectively, but the combined approach increased the accuracy to 0.922. The only exception occurred when characterizing the success of emergence, where the ecological-only approach achieved the same accuracy as the combined approach. We could distinguish which isolates would be "successfully introduced" to an accuracy of 82% using both ecological and evolutionary data, but 13% of these were false positives, suggesting that our analysis could have overlooked a limiting factor that prevents an isolate from successfully establishing in an area.

Figure 3. Confusion matrices visualizing the predictions of random forest classifier models for each expansion dynamic when applied to unseen test datasets: (A) population waves, (B) temporal evolution, (C) expansion success, (D) binary stages of expansion, and (E) categorical stages of expansion. (B, C, and D) For binary expansion dynamics, the matrix represents (clockwise from top left) true negatives, false positives, true positives, and false negatives. (A and E) For categorical expansion dynamics, the matrix shows correct class membership and misclassified class memberships for each category.

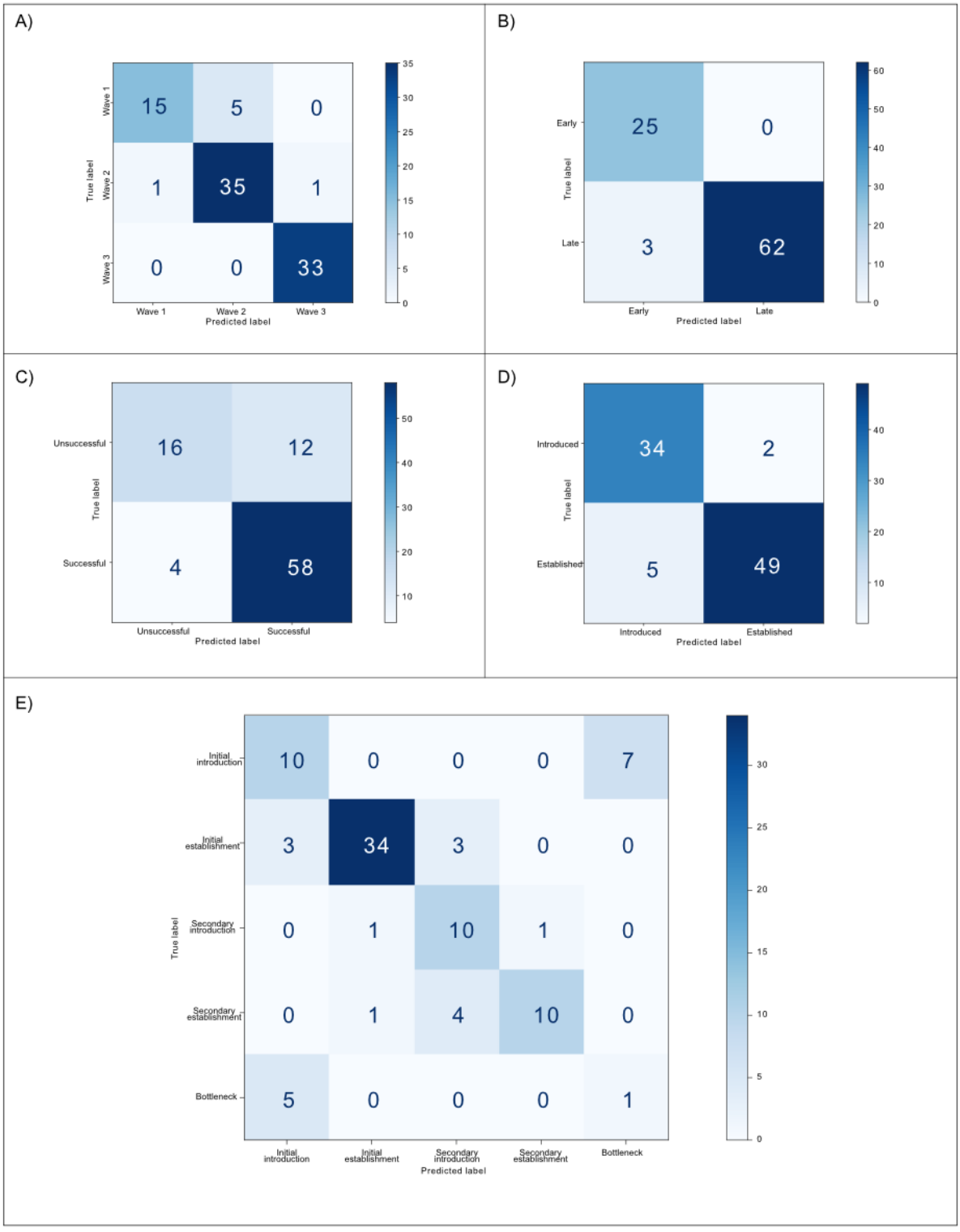


Table 2. Accuracy metrics of random forest classifier models predicting unseen test data of each expansion dynamic.

Expansion dynamics	Combined eco-evolutionary approach				Evolutionary features only				Ecological features only			
	Precision	Recall	F_1-score	Overall accuracy	Precision	Recall	F_1-score	Overall accuracy	Precision	Recall	F_1-score	Overall accuracy
Populations				0.922				0.733				0.744
Wave 1	0.938	0.750	0.833		0.529	0.450	0.486		0.900	0.450	0.600	
Wave 2	0.875	0.946	0.909		0.676	0.676	0.676		0.642	0.919	0.756	
Wave 3	0.971	1.000	0.985		0.889	0.970	0.928		0.889	0.727	0.800	
Unweighted average	0.928	0.899	0.909		0.698	0.698	0.697		0.810	0.699	0.719	
Weighted average	0.924	0.922	0.920		0.721	0.733	0.726	0.733	0.790	0.744	0.737	
Temporal evolution				0.967				0.767				0.956
Early (before 2002)	0.893	1.000	0.943		0.583	0.560	0.571		0.862	1.000	0.926	
Late (after 2003)	1.000	0.954	0.976		0.833	0.846	0.840		1.000	0.938	0.968	
Unweighted average	0.946	0.977	0.960		0.708	0.703	0.706		0.931	0.969	0.947	
Weighted average	0.970	0.967	0.967		0.764	0.767	0.765		0.962	0.956	0.956	
Stages of expansion				0.722				0.511				0.733
Initial introduction	0.556	0.588	0.571		0.412	0.412	0.412		0.526	0.588	0.556	
Established population	0.944	0.850	0.895		0.605	0.575	0.590		0.944	0.850	0.895	
Secondary introduction	0.588	0.833	0.690		0.231	0.250	0.240		0.750	0.750	0.750	
Secondary established population	0.909	0.667	0.769		0.688	0.733	0.710		0.800	0.800	0.800	
Population bottleneck	0.125	0.167	0.143		0.333	0.333	0.333		0.125	0.167	0.143	
Unweighted average	0.624	0.621	0.614		0.454	0.461	0.457		0.629	0.631	0.629	
Weighted average	0.763	0.722	0.735		0.514	0.511	0.512		0.761	0.733	0.745	
Stages of expansion (binary)				0.922				0.667				0.911
Introduction	0.872	0.944	0.907		0.583	0.583	0.583		0.868	0.917	0.892	
Establishment	0.961	0.907	0.933		0.722	0.722	0.722		0.942	0.907	0.925	
Unweighted average	0.916	0.926	0.920		0.6533	0.653	0.653		0.905	0.912	0.908	
Weighted average	0.925	0.922	0.923		0.667	0.667	0.667		0.913	0.911	0.911	
Success				0.822				0.711				0.822
Unsuccessful	0.800	0.571	0.667		0.545	0.429	0.480		0.800	0.571	0.667	
Successful	0.829	0.935	0.879		0.765	0.839	0.800		0.829	0.935	0.879	

Expansion dynamics	Combined eco-evolutionary approach				Evolutionary features only				Ecological features only			
	Precision	Recall	F_1-score	Overall accuracy	Precision	Recall	F_1-score	Overall accuracy	Precision	Recall	F_1-score	Overall accuracy
Unweighted average	0.814	0.753	0.773		0.655	0.634	0.640		0.814	0.753	0.773	
Weighted average	0.820	0.822	0.813		0.696	0.711	0.700		0.820	0.822	0.813	

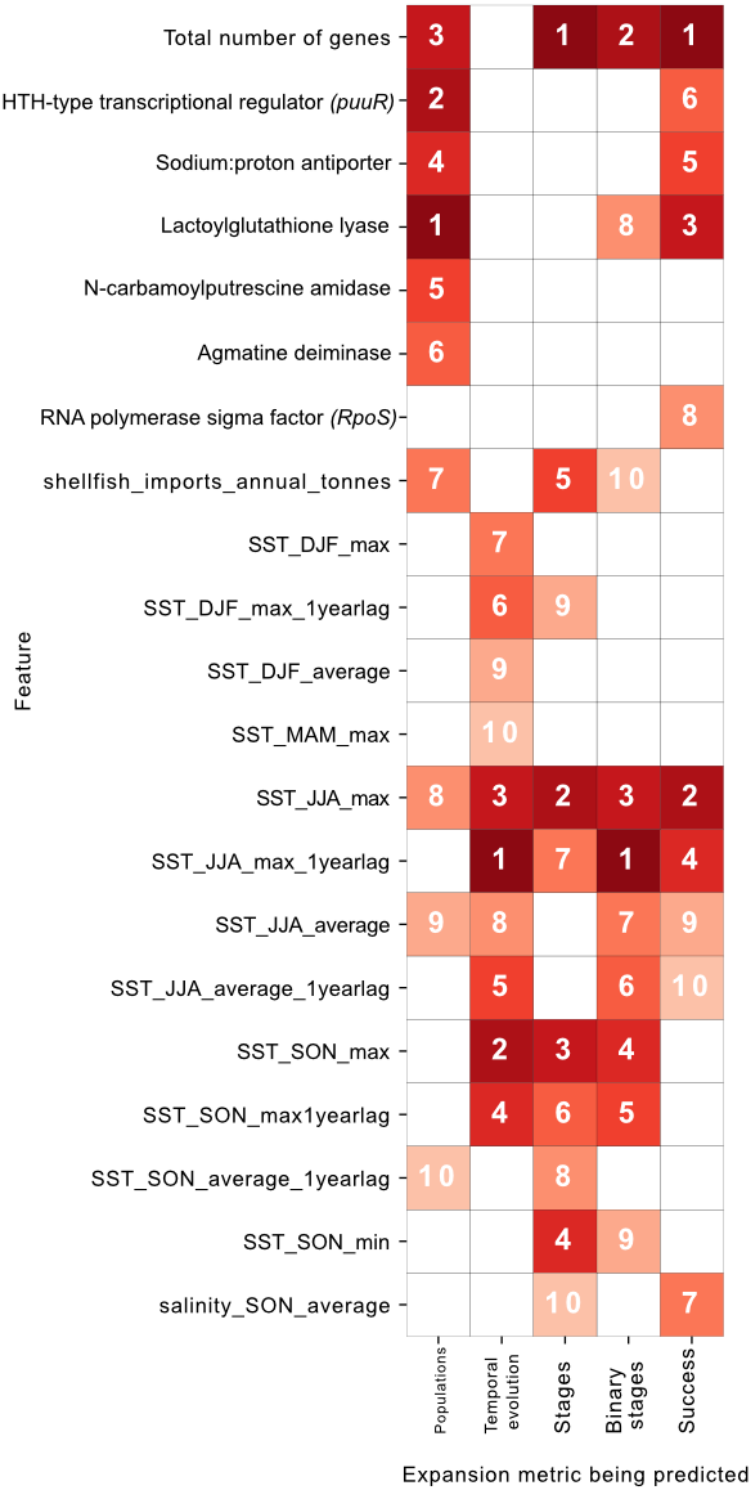
Classes that were particularly difficult to predict, with the lowest accuracies reported, were genetic bottlenecks (which were almost always misclassified as initial introductions) and the eco-evolutionary drivers that result in an isolate's failure to establish successfully. It was harder to predict initial introductions compared to predicting established populations using the categorical "stages of expansion" metric, but when this was reduced to a binary problem, accuracy increased by 0.2, suggesting that separating the stages into initial and secondary introductions (from an established population to a new area) hindered the prediction process.

Exploring the spatiotemporal presence of the errors identified when testing our eco evolutionary models on 90 unseen data points (Table S4 in Multimedia Appendix 1) revealed that the raw highest frequency of errors was found in Asia compared to other continents; however, the relative error rate considering the number of Asian isolates (n=50, 56% of the 90 data points) was the lowest across continents. Notable successes include a strong ability to predict population structure in Asia (in which all 50 samples were accurately predicted) and a low error rate (5.8%) when predicting successful expansions into the United States. Success was more difficult to predict in geographic locations with little representation in the test dataset; or example, there was 1 isolate each from Canada, Japan, Mexico, and Singapore in the test dataset, and only the success of the Canadian isolate was successfully predicted. Temporally, a greater number of errors occurred earlier in the time series, during the initial expansion period of VpST3.

Eco-Evolutionary Feature Importance

In general, ecological metrics performed stronger than evolutionary metrics individually (Figure S2 in Multimedia Appendix 1). Some of the notably important eco-evolutionary drivers included 3 accessory genes, which were almost always present in introduced isolates (and subsequently eroded) and which provided salt stress tolerance, survival advantages, and biofilm formation for motility, as well as summer maximum sea temperatures from both the year of isolate discovery and the year prior. Of the 109 total features used for training and prediction, the 10 (9.2%) strongest predictive features for each metric, based on feature importance, were collated into a data frame (Figure 4). A small range of these ecological and evolutionary metrics featured within the 10 most important features across all 5 expansion dynamics. For 1 expansion dynamic only—temporal evolution—the strongest predictive features were all environmental features, suggesting that the influence of environmental temporal trends outweighed that of the evolutionary drivers. The total number of genes was an important feature for 4 (80%) of the 5 predicted expansion dynamics and located in the top 3 most important features for each of these, suggesting that genetic diversity was a key distinguishing factor between the classes. In addition, the maximum temperatures during June, July, and August were strong predictor variables, appearing in the top 10 features of all models. Lagged sea temperature effects also offered significant information, notably the SSTs during June, July, and August from the previous year. Although salinity variables did not often appear in the top 10 features, the average salinity during September, October, and November was a useful predictor for classifying the stages of expansion and assessing the success chances of an isolate. Shellfish imports featured as important predictors in classifying population waves and the stages of expansion. Accessory gene presences were stronger predictors in classifying population waves and the chance of success than the stages of expansion themselves.

Figure 4. Feature importance for each of the expansion dynamic predictions within each random forest classifier model (where 1 indicates the most important and 10 the 10th most important). DJF: December-January-February; HTH: helix-turn-helix; JJA: June-July-August; MAM: March-April-May; SON: September-October-November; SST: sea surface temperature.



In terms of notable relationships identified, the success of an isolate was generally associated with higher average and maximum SSTs, particularly during June, July, and August. The presence of certain accessory genes, including *puuR*, *aguB*, and *aguC*, was more important in the classification of "introduced" isolates than in the classification of "established" isolates. The isolates that were predicted to be "introduced" (as opposed to "established") almost always had these genes present compared to greater variation in the isolates that were predicted to be "established." This was even more evident in predicting the population to which an isolate belonged, where multiple accessory genes were absent in the third and most recent population wave. Shellfish imports emerged as an important driver in the distinction of the 3 separate populations, with a

higher prediction range seen for isolates belonging to the third population wave.

Cross-Correlation Between Ecological and Evolutionary Drivers

We explored the relationships between the ecological and evolutionary features included in the model and found multiple significant correlations (Figure S3 in Multimedia Appendix 1). Notably, the selected accessory genes exhibited strong correlations with the SST metrics (both positive and negative) as well as with shellfish imports; some genes also correlated with the salinity metrics. In addition, some adaptive SNPs exhibited correlations; for example, the SNP at position 597 had slight negative associations with maximum SSTs and slight positive associations with minimum salinities. The total number of genes had slight positive associations with most of the SST metrics.

Region-Specific Eco-Evolutionary Models

When generating region-specific models to identify which isolates would be specifically successful in China or Peru, as representative countries, we found the model predictions for China to be largely more accurate, with accuracies ranging from 0.778 to 0.917, compared to model predictions for Peru, with accuracies ranging from 0.529 to 0.706 (Table 3). However, while the model was able to successfully classify successful isolates in China, it had difficulty in classifying the unsuccessful isolates, with poor specificity. The Peru model had more balanced predictions between these 2 classes. In both cases, the ecological features–only model was the best approach, providing the best accuracy. Total gene diversity was the top feature for the combined eco-evolutionary approach (and the evolutionary features–only model). For Peru, the remainder of the top 10 important features were ecological features; however, for China it was an even split between ecological and evolutionary drivers, including the type IV pilus twitching motility protein and the SNP at position 603 (Table S3 in Multimedia Appendix 1), which had not appeared previously among the important features. In the ecological features–only model, the top features were December to February minimum sea temperatures and June to August average temperatures a year prior for China and Peru, respectively.

Table 3. Accuracy metrics of region-specific random forest classifier models predicting unseen test data of each expansion dynamic.

Expansion dynamics	Combined eco-evolutionary approach				Evolutionary features only				Ecological features only			
	Precision	Recall	F_1-score	Accuracy	Precision	Recall	F_1-score	Accuracy	Precision	Recall	F_1-score	Accuracy
Success in China				0.889				0.778				0.917
Unsuccessful	0.500	0.250	0.330		0.167	0.250	0.200		1.000	0.250	0.400	
Successful	0.912	0.969	0.939		0.900	0.844	0.871		0.914	1.000	0.955	
Unweighted average	0.706	0.609	0.636		0.533	0.547	0.535		0.957	0.625	0.678	
Weighted average	0.866	0.889	0.872		0.819	0.778	0.796		0.924	0.917	0.894	
Success in Peru				0.529				0.647				0.706
Unsuccessful	0.500	0.375	0.429		0.750	0.375	0.500		0.714	0.625	0.667	
Successful	0.545	0.667	0.600		0.615	0.889	0.727		0.700	0.778	0.737	
Unweighted average	0.523	0.521	0.514		0.683	0.632	0.614		0.707	0.701	0.702	
Weighted average	0.524	0.529	0.519		0.679	0.647	0.620		0.707	0.706	0.704	

Discussion

Principal Findings

Our analysis suggests that VpST3, as a clonal complex, exhibited a high degree of efficacy in propagation during its expansion, evidenced by the numerous introductions in geographically distinct places at similar times. We found evolutionary features that provided mechanisms for this process, including accessory genes linked to functions that facilitate motility and biofilm formation for attachment-based transport mechanisms. The total number of genes within an isolate was an important predictor in the machine learning models for most expansion dynamics. Although we found no trend in gene numbers over time, the model associated higher gene numbers with isolates classified as within established populations, evidenced by a higher prediction range for established isolates. This suggests that isolates that became established could have acquired genes specific to survival in the local conditions, with this plasticity allowing it to colonize new geographic regions. The declining presence of certain accessory genes (*puuR*, *aguB*, and *aguC*) under purifying selection signals suggests that the genes involved in initial introduction may become less useful for population establishment, resulting in reduced selection pressure for these genes. This is corroborated by the prediction ranges of our model for “introduced” isolates, in which these

genes were both important features and characterized as almost always present in introduced isolates.

Assessment of the Eco-Evolutionary Approach

Our analysis has confirmed the hypothesis that considering ecological and evolutionary features in a combined approach to explore the drivers of pathogen expansion yields higher accuracy than dealing with these drivers individually. This is a novel use of the framework described in the study by Campbell et al [17] for characterizing *V parahaemolyticus* expansion dynamics.

From the ecological perspective, SST was a strong predictor variable, as expected from well-established interactions between *V parahaemolyticus* and SST [69]; however, maximum temperatures during June, July, and August emerged as the strongest driver, alongside lagged effects from the previous year. More than two-thirds of our genomic isolates (240/311, 77.2%) were isolated in the northern hemisphere, where these months would be the warmest; this period has previously been described as the "*Vibrio* season" [69]—the characteristics of this season each year seem to drive expansion. In addition, the importance of SSTs in September, October, and November as well as in March, April, and May is pertinent to recent studies that have found expansions in seasonal suitability into cooler months, approximately a 1-month increase every 30 years [70]. Although the period from June to August is the coldest in the southern hemisphere for the South American isolates, it could still drive expansion dynamics when the maximum sea temperatures exceed the minimum for *V parahaemolyticus* survival in the environment, allowing the bacteria to persist in their environmental reservoirs until optimum conditions resume, a phenomenon known as overwintering [71]. Sea temperatures can drive both survival and community composition changes [22,72], with mostly positive associations between SST and the successful established isolates in our analyses. In laboratory studies, increases in seawater temperature have been found to upregulate the expression of virulence factors involved in adhesion processes, such as biofilm formation [73], which could facilitate transport mechanisms via attachment to marine organisms that aid expansion and settlement in new areas.

Generally, the models using evolutionary features only had a lower predictive potential; however, the inclusion of evolutionary features improved the ecological models when combined. The evolutionary features themselves potentially did not offer enough predictive information independently, but when linked to the specific local environmental conditions in which the evolutionary processes provide survival benefits, the evolutionary features were able to provide useful information within the model on pathogen expansion. The evolutionary features might lack meaning outside of ecological contexts or indeed play a different biological role in different ecological contexts. This is supported by the cross-correlations identified between several evolutionary features and the associated environmental conditions (Figure S3 in Multimedia Appendix 1), indicating that these interacting factors themselves, in the form of dynamic evolutionary responses to environmental conditions, can provide predictors of pathogen expansion. This justifies the inclusion of both ecological and evolutionary features in the same predictive model to account for the interactions between them. We observed a specific eco-evolutionary mechanism in our analysis, where SSTs were significantly associated with the presence of multiple accessory genes (Figure S3 in Multimedia Appendix 1), which could indicate an introduced selection pressure in the environment, with changes in SST representing a myriad of implications for the microbial community. However, it is important to note that these cross-correlations provide limited information and could also be purely reflecting the strength of the temporal trends of accessory gene presence, as the result of 2 concurrent or diverging trends, with sea temperature gradually increasing over the time period and accessory gene presence either increasing or decreasing steadily.

Shellfish imports were an important driver for the classification of the third population wave, which could allude to a population opportunistically taking advantage of shellfish movements as a transport mechanism. This would explain why this population has purged multiple accessory genes offering transport mechanisms, such as biofilm pathways. While the role of live aquatic animal transport in contributing to *V parahaemolyticus* expansion is currently unclear, studies have found that this method of transport introduces new populations, facilitates the exchange of genetic material, and promotes adaptation [74]. Further analysis will need to explore whether this subpopulation has undergone innovation to improve host-pathogen attachment mechanisms, particularly involving shellfish.

Few of the SNP mutations identified during outlier detection featured heavily in model decisions, despite our methodology aiming to identify mutations affecting proteins that could promote expansion dynamics. While we encoded the SNPs as categorical features in our machine learning analysis, alternative encoding techniques, such as one-hot encoding, have been explored, and it was found that including information on not only the mutation but also the position of mutation can improve accuracy [75]. Further analysis or different approaches should be explored to improve the identification of mutations critical to expansion processes.

While the models were designed generically to predict a range of expansion metrics, they could be further refined for specific purposes. There were several instances of a large discrepancy between recall and precision, particularly for smaller, underrepresented classes such as bottlenecks, which is a common issue in machine learning when dealing with imbalanced datasets. The models here were not developed individually to obtain the greatest accuracy, as the aim was to facilitate the comparison of accuracy metrics when combining ecological and evolutionary features. However, these imbalances can be remedied on a per-model basis in the future using techniques such as class weights to assign higher weights to minority classes during training or through oversampling (of the minority classes) and undersampling (of the majority classes), as demonstrated by DeLuca et al [76]. The difficulties in separating initial introductions and bottlenecks can be simplified into understanding why a particular introduction is successful or unsuccessful. We did find a potential limiting factor when predicting this success as a separate expansion

metric, resulting in a high proportion of false positives where unsuccessful isolates were misclassified as successful.

We propose that a potential limiting factor here could be plankton presence, which has been found to offer nutrients for growth and host protection [77], which was not included in the analysis. This is relevant given the biofilm-related accessory genes identified, which facilitate attachment to plankton, in which these eco-evolutionary factors could combine to provide further information on isolate success. Similarly, plankton abundance was found to significantly increase the presence of 2 major virulence factors of *V parahaemolyticus*, *tdh* and *trh* [78], underlining another eco-evolutionary mechanism driving *V parahaemolyticus* dynamics. There are difficulties in quantifying marine plankton presence for such a global collection spanning decades. Earth Observation data offers a suitable source for ecological driver data in the future, providing consistent time series data at a sufficient resolution; however, satellite observations of plankton (using chlorophyll-a concentration as a proxy) are only available from late 1997; the key preceding year that represents the pivotal early introductions of the expansion of VpST3 is missing.

The spatiotemporal trends of error counts discussed (Table S4 in Multimedia Appendix 1) offer insights into model limitations and areas for future improvement, such as improving our predictive capabilities during the initial emergence of a pathogen strain and in geographic regions reporting few isolates (as is common during initial expansion).

Regional Predictive Performance for a Globally Expanding Pathogen

The difference in accuracy between the China and Peru regional models is likely due to the consequences of class imbalances. The Chinese isolates had a much higher proportion of successful isolates (108/120, 90%) than Peru (27/55, 49%), which meant that, although we were able to predict successful isolates with high precision and recall, it was very difficult to predict the minority class of unsuccessful isolates (F_1-score=0.33). Such class imbalances result in overfitting of the majority class, enabling the model to achieve a high accuracy of 90% even if it simply predicted all isolates to be successful. This can be seen in the China model using only ecological features, in which the majority class (successful isolates) was predicted perfectly due to 97% (35/36) of the data points being predicted as successful. Further evidence for overfitting is provided by a large difference between the area under the receiver operating characteristic curve values of the training and test data, which were 0.860 and 0.949, respectively. To overcome such overfitting during the future development of regional models, per-class and alternative accuracy metrics need to be considered and imbalances addressed through methods previously outlined. In the Peru model, the number of successful and unsuccessful isolates were much more balanced, resulting in lower but more balanced per-class accuracy metrics. Currently, this would suggest that we can predict the success of a pathogenic variant isolate more accurately in an endemic region than in an emerging one but at the expense of possible overfitting, providing areas for improvement. In both cases, we found that ecological drivers alone were the best approach, suggesting that the evolutionary features were introducing noise into the model. This suggests that focusing on common features in the whole group that might facilitate expansion on a global scale might not be as valuable as more region-specific evolutionary drivers, such as those representative of adaptation to local conditions of a particular region, which would need to be extracted for a more successful regional approach.

However, it is important to note that while models can be improved specifically for particular geographic regions, for example, based on the ranges of local environmental conditions, this comes at the expense of declining applicability. Such applicability could be seen as a priority for a globally expanding pathogen such as VpST3, requiring a model that is able to function in a range of distinct geographic regions. Future work could mediate this trade-off through the introduction of regional encoders as features [24] or through engineering environmental features to be more comparable, such as through normalized anomalies rather than raw values.

Future Predictive Potential

While this analysis focused on the 3 continents reporting the most VpST3 isolates (Asia, North America, and South America), in the future, the focus will need to shift to countries that lie on the periphery of the environmental tolerance ranges of *V parahaemolyticus*, representing the potential locations of future expansion. These include Europe, which, in recent years, has observed the emergence of *Vibrio* lineages and increases in vibriosis incidence as an emerging public health issue [79]. Increased genomic surveillance is required in these countries to test the ability of this framework to identify expansion potential into these new regions.

In addition, the eco-evolutionary analysis was limited by the annual resolution of the genomic isolate metadata and shellfish movement data. The majority of the isolates in our collection were submitted to public databases with limited metadata, specifying only a country and a year; however, higher-resolution metadata, such as a district and a day, week, or month, as suggested by Campbell et al [17], would greatly improve the specificity of the related ecological data that we could then append to this isolate, which is available at a very high resolution. This is particularly necessary to account for the rapid evolutionary timescales on which bacteria such as *Vibrio* function [29]. Future models would benefit from higher spatiotemporal–resolution datasets for machine learning training that facilitate the characterization of more specific eco-evolutionary drivers and increase predictive accuracy.

Conclusions

This pilot study provides a precedent for combining ecological and evolutionary driver data using machine learning to predict pathogen expansion metrics. This both aids our understanding of historic expansion and, through further refinement and development, could be operationalized into a trained database through which a new recovered isolate could be submitted and predictions made as to its introduction or establishment potential to track pathogen expansion in near real time. The current limitations preventing such operationalization include sufficient genomic surveillance, data accessibility, and interdisciplinary

analysis requirements. Accuracy would need to be refined to the appropriate confidence values based on user requirements of model sensitivity. Further exploration of applicability to a range of climate-sensitive pathogens will require sufficient genomic surveillance, which is currently limited by poor spatiotemporal resolution. Combining state-of-the art analyses of both ecological and evolutionary pathogen drivers will provide new insights into future eco-evolutionary pathways of climate-sensitive pathogens.

Acknowledgments

AMC was supported by the Natural Environment Research Council (NE/S007210/1) and Centre for Environment, Fisheries and Aquaculture Science internal Seedcorn funding. JM-U was funded by grants from the Ministerio de Ciencia e Innovación (Spain; PID2021-127107NB-I00) and the Generalitat de Catalunya (Spain; 2021 SGR 00526). This work is funded within the framework of the Horizon Europe program through the project IDAlert.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables and figures.
[DOCX File , 536 KB - bioinform_v5i1e62747_app1.docx]

References

1. Martinez-Urtaza J, Bowers JC, Trinanes J, DePaola A. Climate anomalies and the increasing risk of Vibrio parahaemolyticus and Vibrio vulnificus illnesses. Food Res Int 2010 Aug;43(7):1780-1790. [doi: 10.1016/j.foodres.2010.04.001]
2. Abanto M, Gavilan RG, Baker-Austin C, Gonzalez-Escalona N, Martinez-Urtaza J. Global expansion of Pacific Northwest Vibrio parahaemolyticus sequence type 36. Emerg Infect Dis 2020 Feb;26(2):323-326 [FREE Full text] [doi: 10.3201/eid2602.190362] [Medline: 31961301]
3. Baker-Austin C, Oliver JD, Alam M, Ali A, Waldor MK, Qadri F, et al. Vibrio spp. infections. Nat Rev Dis Primers 2018 Jul 12;4(1):8-19. [doi: 10.1038/s41572-018-0005-8] [Medline: 30002421]
4. Nair GB, Ramamurthy T, Bhattacharya SK, Dutta B, Takeda Y, Sack DA. Global dissemination of Vibrio parahaemolyticus serotype O3:K6 and its serovariants. Clin Microbiol Rev 2007 Jan;20(1):39-48 [FREE Full text] [doi: 10.1128/CMR.00025-06] [Medline: 17223622]
5. Arakawa E, Murase T, Shimada T, Okitsu T, Yamai S, Watanabe H. Emergence and prevalence of a novel Vibrio parahaemolyticus O3:K6 clone in Japan. Jpn J Infect Dis 1999 Dec;52(6):246-247. [Medline: 10738364]
6. Gil AI, Miranda H, Lanata CF, Prada A, Hall ER, Barreno CM, et al. O3:K6 serotype of Vibrio parahaemolyticus identical to the global pandemic clone associated with diarrhea in Peru. Int J Infect Dis 2007 Jul;11(4):324-328 [FREE Full text] [doi: 10.1016/j.ijid.2006.08.003] [Medline: 17321179]
7. González-Escalona N, Cachicas V, Acevedo C, Rioseco ML, Vergara JA, Cabello F, et al. Vibrio parahaemolyticus diarrhea, Chile, 1998 and 2004. Emerg Infect Dis 2005 Jan;11(1):129-131 [FREE Full text] [doi: 10.3201/eid1101.040762] [Medline: 15705337]
8. Centers for Disease Control and Prevention (CDC). Outbreak of Vibrio parahaemolyticus infection associated with eating raw oysters and clams harvested from Long Island Sound--Connecticut, New Jersey, and New York, 1998. MMWR Morb Mortal Wkly Rep 1999 Jan 29;48(3):48-51 [FREE Full text] [Medline: 9935142]
9. Okuda J, Ishibashi M, Hayakawa E, Nishino T, Takeda Y, Mukhopadhyay AK, et al. Emergence of a unique O3:K6 clone of Vibrio parahaemolyticus in Calcutta, India, and isolation of strains from the same clonal group from Southeast Asian travelers arriving in Japan. J Clin Microbiol 1997 Dec;35(12):3150-3155 [FREE Full text] [doi: 10.1128/jcm.35.12.3150-3155.1997] [Medline: 9399511]
10. Centers for Disease Control and Prevention (CDC). Outbreak of Vibrio parahaemolyticus infections associated with eating raw oysters--Pacific Northwest, 1997. MMWR Morb Mortal Wkly Rep 1998 Jun 12;47(22):457-462 [FREE Full text] [Medline: 9639367]
11. Bag PK, Nandi S, Bhadra RK, Ramamurthy T, Bhattacharya SK, Nishibuchi M, et al. Clonal diversity among recently emerged strains of Vibrio parahaemolyticus O3:K6 associated with pandemic spread. J Clin Microbiol 1999 Jul;37(7):2354-2357 [FREE Full text] [doi: 10.1128/JCM.37.7.2354-2357.1999] [Medline: 10364615]
12. -. Vibrio parahaemolyticus, Japan, 1996-1998. Wkly Epidemiol Rec 1999 Oct 29;74(43):361-363. [Medline: 10887775]
13. Chiou CS, Hsu SY, Chiu SI, Wang TK, Chao CS. Vibrio parahaemolyticus serovar O3:K6 as cause of unusually high incidence of food-borne disease outbreaks in Taiwan from 1996 to 1999. J Clin Microbiol 2000 Dec;38(12):4621-4625 [FREE Full text] [doi: 10.1128/JCM.38.12.4621-4625.2000] [Medline: 11101606]

14. Smolikova L, Lomov IM, Khomenko TV, Murnachev GP, Kudriakova TA, Fetsaĭlova OP, et al. [Studies on halophilic vibrios causing a food poisoning outbreak in the city of Vladivostok]. Zh Mikrobiol Epidemiol Immunobiol 2001(6):3-7. [Medline: 11881491]
15. Ansaruzzaman M, Lucas M, Deen JL, Bhuiyan NA, Wang XY, Safa A, et al. Pandemic serovars (O3:K6 and O4:K68) of Vibrio parahaemolyticus associated with diarrhea in Mozambique: spread of the pandemic into the African continent. J Clin Microbiol 2005 Jun;43(6):2559-2562 [FREE Full text] [doi: 10.1128/JCM.43.6.2559-2562.2005] [Medline: 15956363]
16. Smith JM, Feil EJ, Smith NH. Population structure and evolutionary dynamics of pathogenic bacteria. Bioessays 2000 Dec;22(12):1115-1122. [doi: 10.1002/1521-1878(200012)22:12<1115::AID-BIES9>3.0.CO;2-R] [Medline: 11084627]
17. Campbell AM, Hauton C, Baker-Austin C, van Aerle R, Martinez-Urtaza J. An integrated eco-evolutionary framework to predict population-level responses of climate-sensitive pathogens. Curr Opin Biotechnol 2023 Apr;80:102898 [FREE Full text] [doi: 10.1016/j.copbio.2023.102898] [Medline: 36739640]
18. Deeb R, Tufford D, Scott GI, Moore JG, Dow K. Impact of climate change on Vibrio vulnificus abundance and exposure risk. Estuaries Coast 2018 Dec 7;41(8):2289-2303 [FREE Full text] [doi: 10.1007/s12237-018-0424-5] [Medline: 31263385]
19. Galanis E, Otterstatter M, Taylor M. Measuring the impact of sea surface temperature on the human incidence of Vibrio sp. infection in British Columbia, Canada, 1992-2017. Environ Health 2020 May 27;19(1):58 [FREE Full text] [doi: 10.1186/s12940-020-00605-x] [Medline: 32460848]
20. Hartwick MA, Urquhart EA, Whistler CA, Cooper VS, Naumova EN, Jones SH. Forecasting seasonal Vibrio parahaemolyticus concentrations in New England shellfish. Int J Environ Res Public Health 2019 Nov 07;16(22):4341 [FREE Full text] [doi: 10.3390/ijerph16224341] [Medline: 31703312]
21. Kopprio GA, Neogi SB, Rashid H, Alonso C, Yamasaki S, Koch BP, et al. Vibrio and bacterial communities across a pollution gradient in the Bay of Bengal: unraveling their biogeochemical drivers. Front Microbiol 2020 Apr 15;11:594 [FREE Full text] [doi: 10.3389/fmicb.2020.00594] [Medline: 32351470]
22. Siboni N, Balaraju V, Carney R, Labbate M, Seymour JR. Spatiotemporal dynamics of Vibrio spp. within the Sydney Harbour Estuary. Front Microbiol 2016 Apr 12;7:460 [FREE Full text] [doi: 10.3389/fmicb.2016.00460] [Medline: 27148171]
23. Xu M, Cao C, Wang D, Kan B. Identifying environmental risk factors of cholera in a coastal area with geospatial technologies. Int J Environ Res Public Health 2014 Dec 29;12(1):354-370 [FREE Full text] [doi: 10.3390/ijerph120100354] [Medline: 25551518]
24. Campbell AM, Racault MF, Goult S, Laurenson A. Cholera risk: a machine learning approach applied to essential climate variables. Int J Environ Res Public Health 2020 Dec 15;17(24):9378 [FREE Full text] [doi: 10.3390/ijerph17249378] [Medline: 33333823]
25. Froelich B, Gonzalez R, Blackwood D, Lauer K, Noble R. Decadal monitoring reveals an increase in Vibrio spp. concentrations in the Neuse River Estuary, North Carolina, USA. PLoS One 2019 Apr 23;14(4):e0215254 [FREE Full text] [doi: 10.1371/journal.pone.0215254] [Medline: 31013284]
26. Martinez-Urtaza J, Lozano-Leon A, Varela-Pet J, Trinanes J, Pazos Y, Garcia-Martin O. Environmental determinants of the occurrence and distribution of Vibrio parahaemolyticus in the rias of Galicia, Spain. Appl Environ Microbiol 2008 Jan;74(1):265-274 [FREE Full text] [doi: 10.1128/AEM.01307-07] [Medline: 17981951]
27. Daniels NA, MacKinnon L, Bishop R, Altekruse S, Ray B, Hammond RM, et al. Vibrio parahaemolyticus infections in the United States, 1973-1998. J Infect Dis 2000 May;181(5):1661-1666. [doi: 10.1086/315459] [Medline: 10823766]
28. McLaughlin JB, DePaola A, Bopp CA, Martinek KA, Napolilli NP, Allison CG, et al. Outbreak of Vibrio parahaemolyticus gastroenteritis associated with Alaskan oysters. N Engl J Med 2005 Oct 06;353(14):1463-1470. [doi: 10.1056/NEJMoa051594] [Medline: 16207848]
29. Oberbeckmann S, Fuchs BM, Meiners M, Wichels A, Wiltshire KH, Gerdts G. Seasonal dynamics and modeling of a Vibrio community in coastal waters of the North Sea. Microb Ecol 2012 Apr 28;63(3):543-551. [doi: 10.1007/s00248-011-9990-9] [Medline: 22202887]
30. Banerjee SK, Rutley R, Bussey J. Diversity and dynamics of the Canadian coastal Vibrio community: an emerging trend detected in the temperate regions. J Bacteriol 2018 Aug 01;200(15):e00787-e00717 [FREE Full text] [doi: 10.1128/JB.00787-17] [Medline: 29735763]
31. Montánchez I, Ogayar E, Plágaro AH, Esteve-Codina A, Gómez-Garrido J, Orruño M, et al. Analysis of Vibrio harveyi adaptation in sea water microcosms at elevated temperature provides insights into the putative mechanisms of its persistence and spread in the time of global warming. Sci Rep 2019 Jan 22;9(1):289 [FREE Full text] [doi: 10.1038/s41598-018-36483-0] [Medline: 30670759]
32. Thompson CL, Alberti M, Barve S, Battistuzzi FU, Drake JL, Goncalves GC, et al. Back to the future: reintegrating biology to understand how past eco-evolutionary change can predict future outcomes. Integr Comp Biol 2022 Feb 05;61(6):2218-2232. [doi: 10.1093/icb/icab068] [Medline: 33964141]
33. Jolley KA, Maiden MC. BIGSdb: scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics 2010 Dec 10;11(1):595 [FREE Full text] [doi: 10.1186/1471-2105-11-595] [Medline: 21143983]
34. Petit RA, Read TD. Bactopia: a flexible pipeline for complete analysis of bacterial genomes. mSystems 2020 Aug 25;5(4). [doi: 10.1128/msystems.00190-20]

35. Treangen TJ, Ondov BD, Koren S, Phillippy AM. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol 2014 Nov 19;15(11):524 [FREE Full text] [doi: 10.1186/s13059-014-0524-x] [Medline: 25410596]
36. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res 2015 Feb 18;43(3):e15 [FREE Full text] [doi: 10.1093/nar/gku1196] [Medline: 25414349]
37. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol 2016 Jan 10;2(1):vew007 [FREE Full text] [doi: 10.1093/ve/vew007] [Medline: 27774300]
38. Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, et al. BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. PLoS Comput Biol 2019 Apr 8;15(4):e1006650 [FREE Full text] [doi: 10.1371/journal.pcbi.1006650] [Medline: 30958812]
39. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 2012 Aug;29(8):1969-1973 [FREE Full text] [doi: 10.1093/molbev/mss075] [Medline: 22367748]
40. Vaughan TG, Kühnert D, Popinga A, Welch D, Drummond AJ. Efficient Bayesian inference under the structured coalescent. Bioinformatics 2014 Aug 15;30(16):2272-2279 [FREE Full text] [doi: 10.1093/bioinformatics/btu201] [Medline: 24753484]
41. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior summarization in Bayesian phylogenetics using tracer 1.7. Syst Biol 2018 Sep 01;67(5):901-904 [FREE Full text] [doi: 10.1093/sysbio/syy032] [Medline: 29718447]
42. Volz EM, Carsten W, Grad YH, Frost SD, Dennis AM, Didelot X. Identification of hidden population structure in time-scaled phylogenies. Syst Biol 2020 Sep 01;69(5):884-896 [FREE Full text] [doi: 10.1093/sysbio/syaa009] [Medline: 32049340]
43. Conrad RE, Viver T, Gago JF, Hatt JK, Venter SN, Rossello-Mora R, et al. Toward quantifying the adaptive role of bacterial pangenomes during environmental perturbations. ISME J 2022 May;16(5):1222-1234 [FREE Full text] [doi: 10.1038/s41396-021-01149-9] [Medline: 34887548]
44. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 2015 Nov 15;31(22):3691-3693 [FREE Full text] [doi: 10.1093/bioinformatics/btv421] [Medline: 26198102]
45. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biol 2016 Nov 25;17(1):238 [FREE Full text] [doi: 10.1186/s13059-016-1108-8] [Medline: 27887642]
46. Kosakovsky Pond SL, Frost SD. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 2005 May;22(5):1208-1222. [doi: 10.1093/molbev/msi105] [Medline: 15703242]
47. Luu K, Bazin E, Blum MG. pcadapt: an R package to perform genome scans for selection based on principal component analysis. Mol Ecol Resour 2017 Jan;17(1):67-77. [doi: 10.1111/1755-0998.12592] [Medline: 27601374]
48. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 2012 Oct 27;6(2):80-92 [FREE Full text] [doi: 10.4161/fly.19695] [Medline: 22728672]
49. Namadi P, Deng Z. Optimum environmental conditions controlling prevalence of Vibrio parahaemolyticus in marine environment. Mar Environ Res 2023 Jan;183:105828. [doi: 10.1016/j.marenvres.2022.105828] [Medline: 36423461]
50. Soci C, Hersbach H, Simmons A, Poli P, Bell B, Berrisford P, et al. The ERA5 global reanalysis from 1940 to 2022. Q J R Meteorol Soc 2024 Jul 31. [doi: 10.1002/qj.4803]
51. Good SA, Martin MJ, Rayner NA. EN4: quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates. JGR Oceans 2013 Dec 11;118(12):6704-6716. [doi: 10.1002/2013jc009067]
52. Gouretski V, Cheng L. Correction for systematic errors in the global dataset of temperature profiles from mechanical bathythermographs. J Atmos Ocean Technol 2020;37(5):841-855. [doi: 10.1175/JTECH-D-19-0205.1]
53. Sarkar BL, Nair GB, Banerjee AK, Pal SC. Seasonal distribution of Vibrio parahaemolyticus in freshwater environs and in association with freshwater fishes in Calcutta. Appl Environ Microbiol 1985 Jan;49(1):132-136 [FREE Full text] [doi: 10.1128/aem.49.1.132-136.1985] [Medline: 3977307]
54. Morris JGJ, Wilson R, Davis BR, Wachsmuth IK, Riddle CF, Wathen HG, et al. Non-O group 1 Vibrio cholerae gastroenteritis in the United States: clinical, epidemiologic, and laboratory characteristics of sporadic cases. Ann Intern Med 1981 May;94(5):656-658. [doi: 10.7326/0003-4819-94-5-656] [Medline: 7235397]
55. Pal SC, Sircar BK, Nair GB, Deb BC. Epidemiology of bacterial diarrhoeal diseases in India with special reference to Vibrio parahaemolyticus infections. In: Takeda Y, Miwatani T, editors. Bacterial Diarrheal Diseases. Dordrecht, The Netherlands: Springer; 1985:65-73.
56. FishStatJ, F. A. O. A tool for fishery statistics analysis. FAO Fisheries and Aquaculture Department, FIPS–Statistics and Information. 2017. URL: https://scholar.google.com/scholar_lookup?&title=FishStatJ%2C%20a%20tool%20for%20fishery%20statistics%20analysis&publication_year=2024 [accessed 2024-11-12]
57. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12(85):2825-2830 [FREE Full text]

58. Waskom ML. seaborn: statistical data visualization. J Open Source Softw 2021 Apr;6(60):3021 [FREE Full text] [doi: 10.21105/joss.03021]
59. Chakraborty S, Gogoi M, Chakravortty D. Lactoylglutathione lyase, a critical enzyme in methylglyoxal detoxification, contributes to survival of Salmonella in the nutrient rich environment. Virulence 2015 Dec 17;6(1):50-65 [FREE Full text] [doi: 10.4161/21505594.2014.983791] [Medline: 25517857]
60. Nemoto N, Kurihara S, Kitahara Y, Asada K, Kato K, Suzuki H. Mechanism for regulation of the putrescine utilization pathway by the transcription factor PuuR in Escherichia coli K-12. J Bacteriol 2012 Jul;194(13):3437-3447 [FREE Full text] [doi: 10.1128/JB.00097-12] [Medline: 22522900]
61. Merz AJ, So M, Sheetz MP. Pilus retraction powers bacterial twitching motility. Nature 2000 Sep 07;407(6800):98-102. [doi: 10.1038/35024105] [Medline: 10993081]
62. Akram U, Song Y, Liang C, Abid MA, Askari M, Myat AA, et al. Genome-wide characterization and expression analysis of NHX gene family under salinity stress in Gossypium barbadense and its comparison with Gossypium hirsutum. Genes (Basel) 2020 Jul 16;11(7):803 [FREE Full text] [doi: 10.3390/genes11070803] [Medline: 32708576]
63. Liu Z, Hossain SS, Morales Moreira Z, Haney CH. Putrescine and its metabolic precursor arginine promote biofilm and c-di-GMP synthesis in pseudomonas aeruginosa. J Bacteriol 2022 Jan 18;204(1):e0029721 [FREE Full text] [doi: 10.1128/JB.00297-21] [Medline: 34723645]
64. del Rio B, Linares DM, Ladero V, Redruello B, Fernández M, Martin MC, et al. Putrescine production via the agmatine deiminase pathway increases the growth of Lactococcus lactis and causes the alkalinization of the culture medium. Appl Microbiol Biotechnol 2015 Jan 24;99(2):897-905. [doi: 10.1007/s00253-014-6130-8] [Medline: 25341400]
65. Pis Diez CM, Juncos MJ, Villarruel Dujovne M, Capdevila DA. Bacterial transcriptional regulators: a road map for functional, structural, and biophysical characterization. Int J Mol Sci 2022 Feb 16;23(4):2179 [FREE Full text] [doi: 10.3390/ijms23042179] [Medline: 35216300]
66. Li P, Liu C, Li B, Ma Q. Structural analysis of the CARB β-lactamase from Vibrio parahaemolyticus facilitates application of the β-lactam/β-lactamase inhibitor therapy. Biochimie 2020 Apr;171-172:213-222. [doi: 10.1016/j.biochi.2020.03.011] [Medline: 32179166]
67. Lee J, Shin E, Park J, Lee M, Lee K. Regulator of ribonuclease activity modulates the pathogenicity of Vibrio vulnificus. J Microbiol 2021 Dec 09;59(12):1133-1141. [doi: 10.1007/s12275-021-1518-5] [Medline: 34751908]
68. Zhou G, Wang YS, Peng H, Huang XM, Xie XB, Shi QS. Role of Ttca of Citrobacter Werkmanii in bacterial growth, biocides resistance, biofilm formation and swimming motility. Int J Mol Sci 2018 Sep 06;19(9):2644 [FREE Full text] [doi: 10.3390/ijms19092644] [Medline: 30200616]
69. Froelich BA, Daines DA. In hot water: effects of climate change on Vibrio-human interactions. Environ Microbiol 2020 Oct;22(10):4101-4111. [doi: 10.1111/1462-2920.14967] [Medline: 32114705]
70. Trinanes J, Martinez-Urtaza J. Future scenarios of risk of Vibrio infections in a warming planet: a global mapping study. Lancet Planet Health 2021 Jul;5(7):e426-e435 [FREE Full text] [doi: 10.1016/S2542-5196(21)00169-8] [Medline: 34245713]
71. Martinez-Urtaza J, van Aerle R, Abanto M, Haendiges J, Myers RA, Trinanes J, et al. Genomic variation and evolution of Vibrio parahaemolyticus ST36 over the course of a transcontinental epidemic expansion. mBio 2017 Nov 14;8(6):e01425-17 [FREE Full text] [doi: 10.1128/mBio.01425-17] [Medline: 29138301]
72. Logar-Henderson C, Ling R, Tuite AR, Fisman DN. Effects of large-scale oceanic phenomena on non-cholera vibriosis incidence in the United States: implications for climate change. Epidemiol Infect 2019 Jan;147:e243 [FREE Full text] [doi: 10.1017/S0950268819001316] [Medline: 31364581]
73. Billaud M, Seneca F, Tambutté E, Czerucka D. An increase of seawater temperature upregulates the expression of Vibrio parahaemolyticus virulence factors implicated in adhesion and biofilm formation. Front Microbiol 2022 Mar 8;13:840628 [FREE Full text] [doi: 10.3389/fmicb.2022.840628] [Medline: 35350627]
74. Fu S, Wang Q, Zhang Y, Yang Q, Hao J, Liu Y, et al. Dynamics and microevolution of Vibrio parahaemolyticus populations in shellfish farms. mSystems 2021 Jan 12;6(1):e01161-20 [FREE Full text] [doi: 10.1128/mSystems.01161-20] [Medline: 33436516]
75. Singh VK, Maurya NS, Mani A, Yadav RS. Machine learning method using position-specific mutation based classification outperforms one hot coding for disease severity prediction in haemophilia 'A'. Genomics 2020 Nov;112(6):5122-5128 [FREE Full text] [doi: 10.1016/j.ygeno.2020.09.020] [Medline: 32927010]
76. DeLuca NM, Zaitchik BF, Guikema SD, Jacobs JM, Davis BJ, Curriero FC. Evaluation of remotely sensed prediction and forecast models for Vibrio parahaemolyticus in the Chesapeake Bay. Remote Sens Environ 2020 Dec 1;250:112016. [doi: 10.1016/j.rse.2020.112016]
77. Hartwick MA, Berenson A, Whistler CA, Naumova EN, Jones SH. The seasonal microbial ecology of plankton and plankton-associated Vibrio parahaemolyticus in the Northeast United States. Appl Environ Microbiol 2021 Jul 13;87(15):e0297320 [FREE Full text] [doi: 10.1128/AEM.02973-20] [Medline: 33990304]
78. Turner JW, Malayil L, Guadagnoli D, Cole D, Lipp EK. Detection of Vibrio parahaemolyticus, Vibrio vulnificus and Vibrio cholerae with respect to seasonal fluctuations in temperature and plankton abundance. Environ Microbiol 2014 Apr;16(4):1019-1028. [doi: 10.1111/1462-2920.12246] [Medline: 24024909]

79. Amato E, Riess M, Thomas-Lopez D, Linkevicius M, Pitkänen T, Wołkowicz T, et al. Epidemiological and microbiological investigation of a large increase in vibriosis, northern Europe, 2018. Euro Surveill 2022 Jul;27(28):2101088 [FREE Full text] [doi: 10.2807/1560-7917.ES.2022.27.28.2101088] [Medline: 35837965]

Abbreviations

dN/dS: ratio of nonsynonymous to synonymous substitutions
GTR: general time reversible
SNP: single nucleotide polymorphism
SST: sea surface temperature
VpST3: Vibrio parahaemolyticus sequence type 3

Edited by S Hacking; submitted 30.05.24; peer-reviewed by M Jani, R Voicu; comments to author 18.08.24; revised version received 12.09.24; accepted 02.10.24; published 28.11.24.

Please cite as:
Campbell AM, Hauton C, van Aerle R, Martinez-Urtaza J
Eco-Evolutionary Drivers of Vibrio parahaemolyticus Sequence Type 3 Expansion: Retrospective Machine Learning Approach
JMIR Bioinform Biotech 2024;5:e62747
URL: https://bioinform.jmir.org/2024/1/e62747
doi:10.2196/62747
PMID:

Viewpoint

Ethical Considerations in Human-Centered AI: Advancing Oncology Chatbots Through Large Language Models

James C L Chow[1,2], PhD; Kay Li[3], PhD

[1]Department of Radiation Oncology, University of Toronto, Toronto, ON, Canada

[2]Princess Margaret Cancer Centre, University Health Network, Toronto, ON, Canada

[3]Department of English, University of Toronto, Toronto, ON, Canada

Corresponding Author:
James C L Chow, PhD
Princess Margaret Cancer Centre
University Health Network
7/F, Rm 7-606
700 University Ave
Toronto, ON, M5G 1X6
Canada
Phone: 1 4169464501
Fax: 1 4169466566
Email: james.chow@uhn.ca

Abstract

The integration of chatbots in oncology underscores the pressing need for human-centered artificial intelligence (AI) that addresses patient and family concerns with empathy and precision. Human-centered AI emphasizes ethical principles, empathy, and user-centric approaches, ensuring technology aligns with human values and needs. This review critically examines the ethical implications of using large language models (LLMs) like GPT-3 and GPT-4 (OpenAI) in oncology chatbots. It examines how these models replicate human-like language patterns, impacting the design of ethical AI systems. The paper identifies key strategies for ethically developing oncology chatbots, focusing on potential biases arising from extensive datasets and neural networks. Specific datasets, such as those sourced from predominantly Western medical literature and patient interactions, may introduce biases by overrepresenting certain demographic groups. Moreover, the training methodologies of LLMs, including fine-tuning processes, can exacerbate these biases, leading to outputs that may disproportionately favor affluent or Western populations while neglecting marginalized communities. By providing examples of biased outputs in oncology chatbots, the review highlights the ethical challenges LLMs present and the need for mitigation strategies. The study emphasizes integrating human-centric values into AI to mitigate these biases, ultimately advocating for the development of oncology chatbots that are aligned with ethical principles and capable of serving diverse patient populations equitably.

(JMIR Bioinform Biotech 2024;5:e64406) doi:10.2196/64406

KEYWORDS

artificial intelligence; humanistic AI; ethical AI; human-centered AI; machine learning; large language models; natural language processing; oncology chatbot; transformer-based model; ChatGPT; health care

Introduction

Overview

The development of oncology chatbots underscores the critical need for systems grounded in human-centered artificial intelligence (AI) principles that prioritize empathy, accuracy, and personalized patient support. In the context of oncology, where patients and their families often face significant emotional and informational challenges, these chatbots are essential tools for addressing their unique concerns [1-6]. However, as the adoption of large language models (LLMs) such as GPT-3 and GPT-4 becomes increasingly common in health care, the ethical considerations surrounding their use have grown in importance. It is vital that oncology chatbots adhere to ethical standards that ensure fairness, transparency, accountability, and respect for user privacy and autonomy. These systems should be designed to serve diverse user groups, particularly those from underrepresented communities, by avoiding biases and ensuring equitable treatment [7,8]. Human-centered AI in oncology focuses on creating systems that prioritize the needs and experiences of patients and health care providers, thereby

enhancing care, empathy, and support. Ethical AI extends beyond mere technical functionality; it involves embedding principles that safeguard the well-being, dignity, and rights of patients. This includes building trust through transparency, securing patient data, and delivering accurate and bias-free information [9-12].

This review explores the integration of generative AI and LLMs into oncology chatbots, aiming to create tools that embody these human-centered AI principles. The customization and personalization of chatbots are essential to meet the specific needs of each user, transforming traditional chatbots from basic information providers into empathetic, patient-focused tools that significantly enhance the care experience [1,13,14]. The primary goal of this review is to examine the challenges and ethical concerns associated with deploying AI in sensitive health care settings, particularly oncology. As these technologies become more widespread, it is crucial to ensure that they align with human-centered ethical principles. This study is motivated by the need to address potential biases in AI systems, which could inadvertently harm the very patients they are designed to support.

The paper contributes to the field by identifying and analyzing key ethical challenges associated with oncology chatbots, with a specific focus on biases in the datasets used to train these models. Unlike previous studies that provide broad discussions on AI ethics, this review specifically addresses the unique ethical dilemmas faced in oncology, where the stakes are exceptionally high. The study also offers practical strategies for developers and health care providers to enhance the ethical development of AI, proposing a framework for human-centered AI in oncology. The findings of this study reveal that oncology chatbots often endure biases rooted in their training data, leading to unfair or ineffective outcomes. To address these issues, the paper provides strategic recommendations, such as using more diverse and representative datasets, implementing continuous monitoring, and refining training methodologies. These measures aim to ensure that AI-driven tools in oncology are not only effective but also ethically sound. In comparison to existing literature, this study offers a focused analysis of the ethical implications specific to oncology chatbots, an area that has been relatively underexplored. By providing a detailed examination of the sources of bias and presenting practical solutions, this paper advances the conversation on ethical AI in health care, particularly within the critical field of oncology.

Enhancing Oncology Chatbots With Ethical and Human-Centered AI

In oncology, ethical principles like beneficence, nonmaleficence, autonomy, and justice are crucial to ensure patient well-being. Oncology chatbots, designed to support patients and families, must adhere to these guidelines. For example, a chatbot for patients with breast cancer can provide personalized treatment information and emotional support, ensuring that the information is accurate, culturally sensitive, and delivered with empathy [13]. Such chatbots can significantly ease the burden on patients by offering timely and relevant information. However, these chatbots also face ethical challenges, particularly in maintaining privacy and data security. For instance, a pediatric oncology chatbot must securely handle sensitive data, requiring robust encryption and transparent data usage policies [14]. Additionally, regular updates and monitoring are essential to prevent biases or inaccuracies that could harm patients. Transparency is another critical concern. Oncology chatbots must clearly disclose their AI nature to users. For instance, in end-of-life care, failing to inform users that they are interacting with an AI could lead to mistrust and harm the health care organization's reputation. A proactive approach with clear self-disclosure at the start of interactions is essential to maintain trust [15]. In health care domains like nephrology, similar ethical considerations apply, with a focus on patient consent, privacy, and bias mitigation. In educational settings, oncology chatbots can also be valuable, but they must follow ethical frameworks to ensure accurate information delivery and fair AI operation. By adhering to these principles, oncology chatbots can effectively bridge learning gaps while maintaining trust and integrity [6,16-18].

Designing Ethical and Trustworthy Oncology Chatbots With Human-Centered AI

Ethical chatbots, therefore, need to adhere to certain principles. They should prioritize transparency, providing users with clear indications when they are interacting with AI rather than a human. Respecting user privacy, obtaining informed consent, mitigating biases, ensuring data security, and promoting responsible AI use in education are central to developing ethical chatbots. By integrating ethical frameworks and considering societal impact, chatbots can contribute positively while upholding ethical standards in their interactions with users.

There are 2 concerns to build an ethical oncology chatbot in human-centered AI—first, it has to build trust in the users. Second, how can it build trust? Building a human-centered approach to AI-driven chatbots involves a strategic integration of several key elements. First, the design should prioritize the user's needs and expectations. Rather than merely dispensing information mechanically, the chatbot should discern and address human needs relevantly. This personalized approach fosters a more trustworthy relationship between the user and the AI. Trust emerges as a pivotal concern in the development of AI chatbots.

Second, essential strategies can be used to cultivate trust in the oncology chatbots. The first involves personalization tailored to each user, enhancing the sense of individual relevance and reliability [19]. The second entails infusing the oncology chatbot with a human-like persona, creating a relatable and approachable interaction for users [20]. Implementing these qualities in the design of AI chatbots requires a thoughtful technical strategy. While the focus here is less on technical aspects and more on user experience and interaction, achieving personalization involves machine learning algorithms capable of understanding and adapting to individual user preferences [21]. Meanwhile, instilling a human-like persona necessitates sophisticated natural language processing (NLP) techniques and dialogue design that emulate human conversational patterns [22]. In essence, the development of human-centered AI chatbots revolves around creating an experience that seamlessly integrates technical prowess with an empathetic understanding of human needs

through web-based inputs with the users. By bridging the gap between technological sophistication and human-like interaction, these chatbots can truly serve as effective companions in addressing users' queries and needs.

Ethical Challenges in Implementing Transformer-Based AI Models for Health Care Enhancement

One way is in the design of transformers. In 2019, transformers were used to create LLMs such as Bidirectional Encoder Representations from Transformers (BERT) and GPT-2 [23]. The integration of AI technologies, specifically LLMs, holds immense potential for improving efficiency and decision support in health care settings. However, ethical considerations become paramount when deploying such models, especially in critical domains like health care.

GPT-4, the underlying model of ChatGPT, has demonstrated significant potential in conversational AI applications [24]. This advancement has sparked discussions about the ethical implications of deploying such powerful models in health care. One primary concern is the potential inaccuracies in generated content. LLMs can produce convincing yet incorrect information, posing a risk of errors in medical records. Compounding this issue is the opacity of training data, making it challenging to assess accuracy effectively [25]. To address this concern, it is crucial for LLMs like GPT-4 to train on precise and validated medical datasets [26].

The growing integration of AI chatbots, exemplified by tools such as ChatGPT and Google Bard, in health care, introduces critical security implications [27,28]. While these AI-driven systems hold significant promise for improving patient care and public health, their reliance on massive datasets, including sensitive patient information, raises concerns about data security. During the pandemic, health care chatbots have become extensively used, addressing tasks like appointment scheduling and providing health information [29]. However, this increased usage magnifies security risks and privacy challenges that remain understudied. AI chatbots, like ChatGPT, also pose unique challenges in ensuring patient privacy and compliance with regulations such as the Health Insurance Portability and Accountability Act (HIPAA) [30]. Recent viewpoints in medical journals highlight the need for providers to navigate HIPAA compliance while safeguarding patient data [31]. Additionally, the safety of medical AI chatbots in patient interactions becomes a paramount consideration, necessitating measures to protect patient data, maintain information accuracy, and ensure user understanding [32]. Ethical considerations, including privacy and data security concerns, further complicate the widespread adoption of conversational AI in health care, emphasizing the need for comprehensive guidelines and robust encryption methods to build trust and safeguard sensitive health information in this era of AI-driven health care communication. Another critical ethical consideration is model bias [33]. LLMs may inadvertently perpetuate biases present in their training data, leading to medically inaccurate and discriminatory responses. Biases can stem from various sources such as sampling, programming, and compliance, necessitating careful consideration to avoid perpetuating harmful stereotypes. Striking a balance between model accuracy and unbiased responses is essential for responsible deployment in health care settings.

Privacy, a fundamental principle in health care, adds another layer of ethical complexity when using public LLMs. The potential risks associated with data sharing must be mitigated through strict agreements and HIPAA-compliant training protocols. Ensuring patient privacy is paramount in the implementation of AI technologies in health care [34].

Despite the potential benefits of using AI technologies, particularly transformer-based models, in health care, careful consideration of ethical principles is crucial. Addressing concerns related to accuracy, bias, and privacy will facilitate responsible and patient-centered implementation, benefiting both health care professionals and patients.

The insights from the Megatron transformer underscore the ethical considerations in deploying transformer models like ChatGPT [35]. Trained on vast datasets, Megatron suggests AI's incapacity to independently ensure ethical behavior, emphasizing its tool-like nature dependent on human usage. Addressing the potential biases in transformer models, especially in health care, demands a focus on fairness metrics, proactive bias detection, and diverse training data. Continuous user feedback becomes crucial for iterative refinement, and bias-awareness training for stakeholders fosters a culture of ethical responsibility. Integrating these strategies into the deployment of transformer models is imperative, ensuring more equitable and inclusive AI-generated content across diverse applications.

Ethical Considerations in Deploying LLMs in Health Care and Education

Using LLMs raises ethical considerations, including the potential for biased outputs, breaches of privacy, and the risk of misuse. These may have serious implications in medical settings. Addressing these concerns requires the adoption of transparent development practices, the responsible handling of data, and the integration of fairness mechanisms.

The integration of LLMs, such as ChatGPT, in medical practice and research raises crucial ethical issues concerning bias, trust, authorship, equitability, and privacy [32]. Although this technology has the potential to revolutionize medicine and medical research, being mindful of its potential consequences is essential. An outright ban on the use of this technology would be shortsighted. Instead, establishing guidelines that aim to responsibly and effectively use LLMs is crucial.

LLMs, like BioGPT [36] and LaMDA (Google Brain) [37], are currently under exploration for various applications in the medical field, showcasing versatility in tasks such as text generation, summarization, and aiding in clinical documentation and academic writing. The integration of LLMs, including oncology chatbots powered by ChatGPT, holds promise for streamlining essential health care tasks, including template creation, summarizing academic content, and enhancing the clarity of clinical notes. This potential introduces significant time-saving and efficiency gains in medical settings.

However, the incorporation of LLMs, particularly oncology chatbots, into health care applications also presents ethical challenges that demand careful consideration to ensure responsible use. Recent research underscores concerns related to the attribution of credit and rights for content generated by LLMs. Users may encounter difficulties in fully claiming credit for positive outcomes while potentially facing responsibility for unintended consequences, such as the generation of misinformation. This highlights the pressing need for updated perspectives on responsibility and the establishment of clear guidelines addressing issues like authorship, disclosure, educational applications, and intellectual property in the context of oncology chatbots and LLMs in general. Navigating the ethical implications of integrating oncology chatbots and LLMs into the medical field requires a comprehensive approach to foster responsible and transparent use of these powerful language models in health care settings.

In the field of education, LLMs show potential in automating tasks such as question generation, feedback provision, and essay grading. However, concerns about practicality and ethics, including technological readiness, transparency, and privacy considerations, must be addressed. A systematic scoping review identifies these challenges and recommends updating innovations with state-of-the-art models, open-sourcing models or systems, and adopting a human-centered approach in development. Therefore, the ethical considerations surrounding the use of LLMs in various fields in medicine and education, necessitate a careful and responsible approach. Establishing clear guidelines such as ensuring transparency and incorporating human oversight are essential steps in harnessing the benefits of LLMs while mitigating potential risks.

As a research group focused on human-centered AI and the ethical integration of AI principles into medical and oncology chatbots [1-6], particularly leveraging LLMs [32], our analysis delves into the historical evolution and the transformative potential of LLMs. We aim to spotlight the continuum of advancements in computational theory that has shaped our technological landscape, emphasizing the pivotal role of integrating humanistic and ethical considerations into AI for health care.

LLMs and NLP Unveil New Potential for Human-Centered AI in Oncology Chatbots

Neural Networks and Machine Learning

Neural networks, fundamental to modern AI, emulate the structure and functioning of the human brain, forming the basis for various applications [38]. In the medical context, the integration of LLMs like ChatGPT brings forth unprecedented possibilities. LLMs are part of the NLP domain and are built on architectures such as GPT and BERT [23,39]. Unlike rule-based models, LLMs learn unsupervised from extensive text data during pretraining, gaining a profound understanding of syntax, grammar, and context. Fine-tuning follows, adapting their knowledge for tasks like text generation and sentiment analysis.

Within the broader landscape of human-centered AI, the principles of neural networks and machine learning persist. The capacity of neural networks to capture complex patterns in data, combined with machine learning algorithms, remains instrumental. In the realm of human-centered AI, LLMs and NLP play a crucial role. NLP focuses on enabling machines to comprehend and generate human language, aligning with the principles of human-centered AI [40]. LLMs, as a significant advancement in NLP, excel in understanding and generating human-like language, enhancing natural interactions between AI systems and users. In the context of oncology chatbots, the integration of LLMs is pivotal. These advanced models empower chatbots to comprehend medical queries, respond empathetically, and adapt to diverse communication styles, ultimately improving the user experience in health care interactions. The use of LLMs in oncology chatbots not only fosters effective communication but also reinforces the human-centered aspect by creating more empathetic and context-aware interactions within the medical domain.

LLMs in Oncology Chatbot

LLMs have achieved remarkable breakthroughs, innovating the field of NLP with their capacity to generate human-like text and excel in a multitude of NLP tasks. A compelling example is their application in the development of oncology chatbots [23,32,41]. These chatbots have the ability to communicate with users in a natural and coherent manner, offering invaluable assistance to both health care professionals and patients. LLMs have enabled oncology chatbots to generate human-like responses, providing users with a web-based and intuitive experience. These chatbots can understand complex medical queries, extract relevant information from patients' descriptions of their symptoms, and generate responses that are not only accurate but also easily comprehensible to laypersons. This human-like text-generation capability significantly enhances the user experience, fostering trust and improving communication between patients and health care providers [42].

Furthermore, LLMs empower oncology chatbots to perform diverse NLP tasks within the health care domain. They can extract critical information from medical records, assisting in patient diagnosis and treatment recommendations. These chatbots can also provide medication information, offer guidance on healthy lifestyles, and even support mental health through empathetic conversations [43,44]. Their versatility makes them invaluable tools in health care, augmenting the capabilities of medical professionals and providing accessible, round-the-clock health care information and support. Figure 1 shows the various applications of an oncology chatbot powered by LLMs.

Figure 1. Various applications of an LLM-powered oncology chatbot. LLM: large language model.

Applications and Implications of LLMs for Human-Centered AI in Oncology Chatbots

Practical Applications of LLMs in Human-Centered AI

LLMs have demonstrated extensive practical applications across diverse domains, showcasing their versatility and transformative potential within the framework of human-centered AI. In the realm of language translation, these models have markedly enhanced the precision and fluency of machine translation systems [45]. They adeptly translate text among multiple languages, fostering seamless cross-cultural communication and bolstering global business operations. Within text generation, LLMs exhibit proficiency in crafting human-like text for multifarious purposes, aiding content creation by drafting papers, generating marketing copy, or assisting authors in producing creative content [46]. Moreover, LLMs find use in chatbots and web-based assistants, delivering natural and contextually sensitive responses in customer support, health care, and various other industries [47,48]. An illustration in Figure 2 depicts an example of the RT Bot used in radiotherapy education, epitomizing the integration of LLMs within the sphere of human-centered AI applications.

In software development, LLMs have demonstrated their prowess in code generation and code completion tasks. They can assist programmers by generating code snippets, fixing bugs, and enhancing productivity [49]. Moreover, in data analytics, LLMs are used for natural language querying of databases, simplifying data exploration and analysis for nontechnical users [50]. Moreover, LLMs are invaluable in the health care sector, where they aid in medical record analysis, diagnosis support, and drug discovery [51]. They can sift through vast amounts of medical literature to extract relevant information and assist health care professionals in making informed decisions. LLMs are also used in sentiment analysis and social media monitoring, helping businesses gauge public opinion, and adapt their strategies accordingly [52].

Figure 2. The RT Bot providing education in radiotherapy.

Humanistic and Ethical AI

Humanistic AI refers to the approach in AI development that prioritizes human values, well-being, and understanding in the design and implementation of AI systems [53]. It emphasizes creating AI technologies that align with human principles, fostering empathy, compassion, and a deeper understanding of human needs and emotions. On the other hand, ethical AI involves adhering to moral principles and guidelines in the development and deployment of AI systems [54]. It encompasses considerations of fairness, transparency, accountability, privacy, and the societal impact of AI applications. Ethical AI aims to ensure that AI technologies benefit individuals and communities while minimizing potential harm or biases. Incorporating humanistic and ethical AI principles into oncology chatbots is crucial. Humanistic AI prioritizes empathy and understanding of human needs, while ethical AI ensures fairness, transparency, and accountability. This dual focus not only aligns with societal expectations but also safeguards against biases and harm, ensuring AI benefits individuals and communities in the medical domain [55-57].

The development of LLMs and oncology chatbots is deeply intertwined with the concepts of humanistic and ethical AI. LLMs, such as GPT-3 and GPT-4, are designed to generate human-like text and have been applied to various domains, including health care [58]. Oncology chatbots powered by LLMs aim to provide assistance, information, and even preliminary diagnosis to users [59]. Humanistic AI in oncology chatbots based on LLMs involves creating interfaces and interactions that are more empathetic, understandable, and accommodating to human emotions and concerns [60]. It seeks to imbue these AI systems with a human touch, making them more relatable and comforting for users seeking medical information or support. Ethical considerations in the development of LLM-based oncology chatbots are crucial. These AI systems must maintain patient privacy, ensure the accuracy and reliability of information provided, mitigate biases in data and responses, and offer transparent explanations for their suggestions or advice [61]. In addition, ethical AI in this context involves clearly delineating the capabilities and limitations of oncology chatbots to users, ensuring informed decision-making and responsible use of the technology.

Humanistic and ethical AI principles guide the responsible development and deployment of LLM-based oncology chatbots, promoting trust, reliability, and user satisfaction while addressing societal concerns and ethical implications associated with these AI-driven health care solutions [62].

Societal and Ethical Implications of LLMs in Deploying Oncology Chatbots

The deployment of LLMs in the health care sector, particularly in the form of oncology chatbots, presents both significant benefits and ethical challenges. On one hand, oncology chatbots powered by LLMs can enhance access to health care information

and provide quick assistance to users with medical queries. They offer a convenient means for individuals to seek information about symptoms, treatments, or health care recommendations. However, ethical concerns emerge when considering issues of privacy, security, and misinformation [63]. Oncology chatbots may inadvertently expose sensitive patient information if not properly secured, raising concerns about data breaches and privacy violations. Moreover, LLMs can potentially propagate medical misinformation, leading to incorrect self-diagnoses or treatment decisions [64]. The responsible development and deployment of oncology chatbots must involve robust data protection measures, continuous monitoring for accuracy, and adherence to medical ethics guidelines to ensure that these technological advancements contribute positively to health care while mitigating potential risks. Balancing the benefits of LLM-powered oncology chatbots with these ethical considerations is essential for their responsible and effective use in the health care domain [32]. Above all, the training datasets can be biased, and fall short of the underrepresented communities such as women, aboriginal people, persons with disabilities, and members of visible minorities [65,66]. The oncology chatbots still have to be trained to answer the needs of these communities.

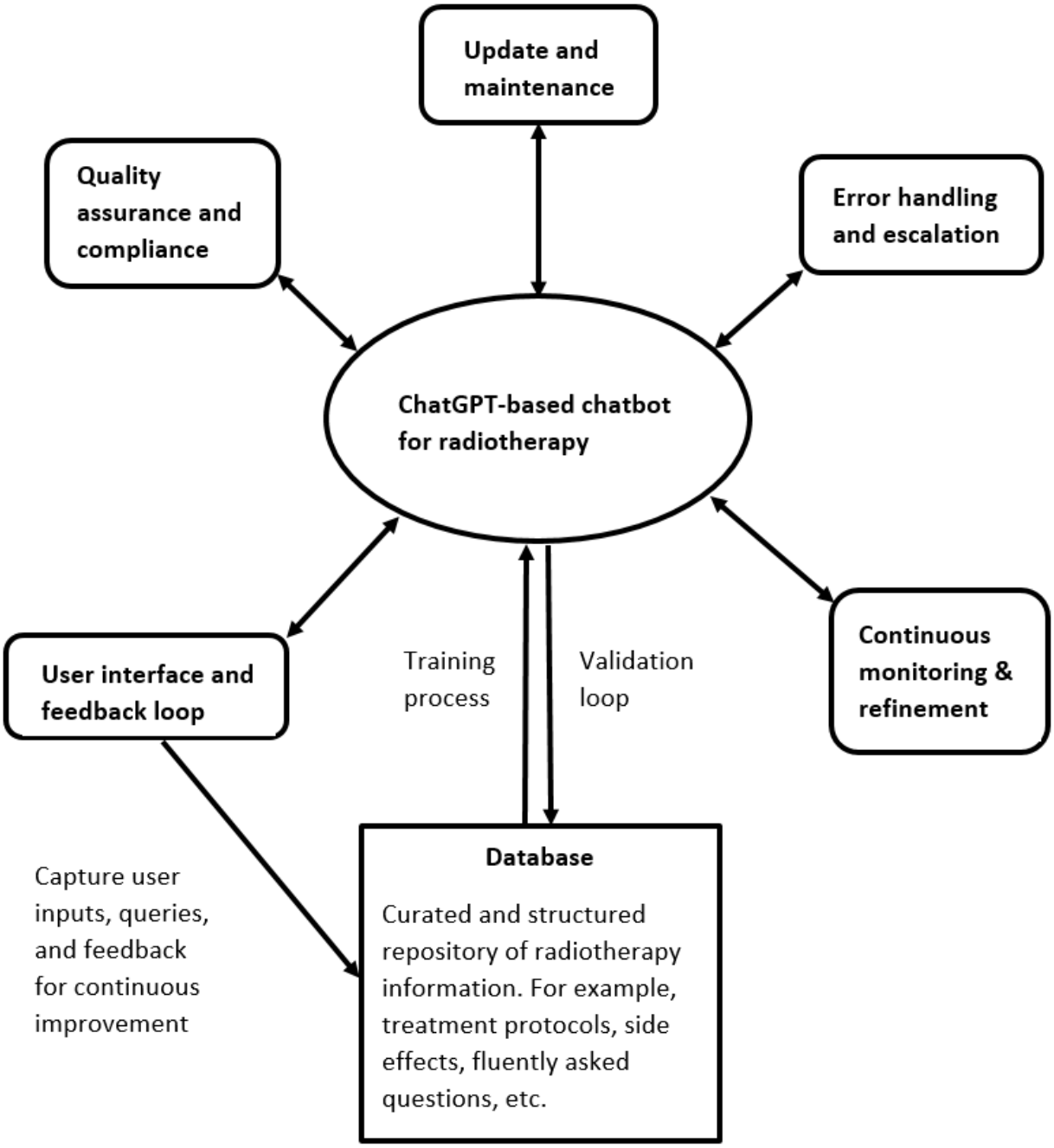
Challenges and Limitations

Incorporating Humanistic and Ethical Principles Into LLM-Driven Oncology Chatbots

The application of LLMs in oncology chatbots not only presents a promising avenue for enhancing health care accessibility and support but also introduces critical ethical considerations within the realm of human-centered AI [20]. Despite the potential benefits, the integration of these systems raises significant ethical concerns and safety considerations. One prevalent issue pertains to the potential perpetuation of biases and discrimination within these AI systems. LLMs, learning from extensive datasets that may inherently contain societal biases, risk generating skewed recommendations or responses that could adversely affect specific demographics, thus perpetuating health care disparities [67,68]. Moreover, the deployment of AI-driven chatbots might inadvertently impede individuals' autonomy, recourse, and rights by overshadowing or dismissing their unique health care needs or preferences [69]. Transparency also remains a significant challenge, as these models often generate outcomes that are nontransparent, difficult to explain, or seemingly unjustifiable, making it challenging for users to comprehend or challenge the decisions made by the AI [32]. Furthermore, there are concerns regarding user privacy breaches, as personal health information shared with these chatbots may not always be adequately secured [70]. Additionally, the reliance on AI-driven interactions might risk isolation and the deterioration of the patient-doctor relationship, potentially undermining the crucial social connections essential for holistic health care [71]. Ensuring the reliability and safety of outcomes produced by these chatbots remains a concern, as inaccuracies or poor-quality responses could have detrimental consequences on patient health and well-being [72,73]. Mitigating these ethical challenges and ensuring the safety of LLM-based oncology chatbots necessitate robust frameworks, stringent regulations, and ongoing scrutiny to address potential harms and uphold ethical standards within the domain of human-centered AI in health care. Figure 3 shows a proposed framework of a radiotherapy chatbot based on ChatGPT. The chatbot is anchored by a robust core powered by ChatGPT, interfacing seamlessly with a meticulously curated database of verified medical information [74]. The model undergoes domain-specific training to enhance its comprehension of radiotherapy intricacies, while a continuous feedback loop ensures that validated data inform its responses and are cross-verified for accuracy. To enhance ethical AI practices, the framework should incorporate bias mitigation strategies by diversifying data sources, ensuring transparency about the chatbot's capabilities and limitations, implementing robust user privacy measures, establishing continuous ethical reviews, providing user education on verifying information, and creating accessible feedback mechanisms for reporting inaccuracies. This iterative approach fosters a dynamic, reliable, and ethically responsible ecosystem for delivering accurate and up-to-date information within the scope of radiotherapy [75].

Therefore, integrating humanistic and ethical principles into LLM-based oncology chatbots stands as a significant challenge in contemporary AI development [76]. Achieving this integration requires a comprehensive approach. First, prioritizing patient confidentiality and data security remains pivotal. Implementing robust encryption measures and stringent access controls can effectively mitigate risks associated with sensitive medical information [77]. Second, infusing empathy and sensitivity into the chatbot's responses poses a significant hurdle. It necessitates the development of algorithms capable of understanding and empathetically responding to patients' emotional states, demanding extensive research into sentiment analysis and contextually appropriate language generation [78]. Moreover, carefully considering the ethical implications of decision-making in medical scenarios is crucial. Collaborative efforts among AI developers, ethicists, and medical professionals are vital to embed ethical guidelines into the chatbot's algorithms, ensuring alignment with medical ethics and patient welfare [14,79,80]. Striking a balance between technical functionality and ethical considerations is key to fostering trust and acceptance of LLM-based oncology chatbots in the health care ecosystem. Continuous vigilance, ongoing refinement, and transparent communication about the chatbot's capabilities and limitations are essential steps in responsibly integrating humanistic and ethical principles into this advancing technology [63,81].

Figure 3. Schematic diagram showing the framework of medical chatbot based on large language model–based ChatGPT, focused on radiotherapy, ensuring accuracy, compliance, and continuous refinement.



Approaches to Mitigate Bias in LLM-Driven Oncology Chatbots

To avoid potential bias in LLM-based oncology chatbots, it is crucial to adopt a comprehensive approach. First, ensure that the training data are diverse and representative of the entire population the chatbot aims to assist. This involves incorporating information from various demographic groups, ethnicities, genders, and socioeconomic backgrounds to prevent the model from learning and perpetuating biases present in specific subsets of data [82,83]. Moreover, ethical data collection practices should be a priority, with developers implementing strict guidelines to eliminate unintentional biases. Transparently communicate ethical standards to users and stakeholders to foster trust and accountability in the development process [84]. Incorporating bias detection and correction algorithms during both the training and deployment phases is essential [85,86]. These mechanisms should be designed to identify and rectify biased outputs in real time, with regular updates to adapt to evolving data and user interactions. In addition, transparency is key in addressing bias; therefore, the chatbot should be designed to provide clear explanations for its decisions. This not only enhances user trust but also enables health care professionals to understand the reasoning behind the chatbot's recommendations. Continuous monitoring and evaluation are also important to the chatbot's success [87,88]. Regularly assess its performance over time, ensuring that potential biases are identified and corrected promptly. User feedback integration

further enhances the system, allowing diverse user groups to report biases and contribute to ongoing improvements [1,2]. Furthermore, collaboration with health care professionals is paramount. Involving experts in the development and validation processes helps refine the chatbot's responses, ensuring accuracy and minimizing biases that may arise from a lack of medical context [32]. Finally, regulatory compliance with health care and data protection standards is vital. Adhering to established regulations ensures that the chatbot operates within ethical and legal boundaries, building trust among users and health care providers alike [89]. Table 1 summarizes the strategies for mitigating bias in LLM-based oncology chatbots.

Table 1. Strategies for mitigating bias in large language model–based oncology chatbots.

Strategies	Description
Diverse and representative training data	Use data that reflect the diversity of the target population
Ethical data collection practices	Implement strict ethical guidelines for data collection
Bias detection and correction algorithms	Integrate algorithms to identify and correct biased outputs
Explainability and transparency	Design the chatbot to provide clear explanations for decisions
User feedback integration	Continuously monitor and evaluate the chatbot's performance
Collaboration with health care professionals	Encourage user feedback to identify and address biases
Privacy-preserving models	Involve health care experts in development and validation
Regulatory compliance	Adhere to health care and data protection regulations

Navigating the AI Frontier: Challenges and Ethical Considerations

The rise of LLMs, exemplified by GPT-4, has sparked both excitement and apprehension. Geoffrey Hinton, a prominent figure in deep learning, acknowledges their potential to surpass human intelligence [90]. However, this rapid progress raises ethical and safety concerns. Despite having significantly fewer connections than the human brain, LLMs exhibit remarkable learning capabilities. Their ability to generalize from limited examples challenges conventional wisdom. Hinton argues that their occasional errors and hallucinations are features, akin to human imperfections. His fears extend beyond mere intelligence; he emphasizes the risk of AI misuse by malicious actors. Whether in elections or warfare, AI's capacity to create subgoals and manipulate environments demands urgent attention. Responsible development and regulation are imperative. Hinton envisions a hybrid intelligence—a fusion of learning and communication—where machines outperform humans in both domains. This transformative era requires collective action and societal discussions akin to historical agreements on chemical weapons. As AI development outpaces regulation, Hinton questions whether our existing social structures can handle the implications. Responsible AI deployment necessitates interdisciplinary collaboration and thoughtful governance. While some may dismiss Hinton's concerns, the stakes are high. As we navigate the path toward AI advancement, we must grapple with the potential consequences and strive for ethical, human-centered progress.

Concerns of Datasets in LLMs or NLP for Ethical and Human-Centered AI

The ethical considerations surrounding the datasets used in training LLMs and NLP systems are critical for advancing human-centered AI, particularly in the context of oncology chatbots. The datasets used for training these models often reflect societal biases, which can lead to ethical dilemmas when the outputs of these chatbots are applied in real-world health care settings [57]. For instance, commonly used datasets like the Common Crawl, Wikipedia, and clinical databases may overrepresent affluent, Western demographics while underrepresenting minority groups, non-Western cultures, and marginalized communities. This bias can result in oncology chatbots that are less effective in serving diverse patient populations, potentially exacerbating health disparities. Moreover, the ethical implications of dataset bias become evident when examining specific LLMs like GPT-3 and GPT-4. These models are often fine-tuned on domain-specific datasets, which can inadvertently amplify existing biases. For example, if an oncology chatbot is trained predominantly on datasets from high-income health care systems, it may lack the cultural competency required to address the needs of patients from low-income or diverse backgrounds [91]. Such a scenario not only limits the chatbot's effectiveness but also raises concerns regarding equity in health care delivery. Table 2 outlines how to address the concerns related to datasets in LLM or NLP for ethical and human-centered AI, specifically in the context of oncology chatbots. This table provides examples of datasets, identifies potential biases, and suggests strategies for mitigating these biases. By systematically addressing these issues, we seek to illuminate the vital importance of ethical dataset selection and its influence on developing effective, human-centered oncology chatbots. Our findings underscore the need for continuous evaluation and modification of datasets to reduce bias, ensuring that LLMs accurately represent and serve the diverse populations they aim to support in the field of oncology.

Table 2. Overview of dataset, potential biases, and strategies for mitigation in large language model or natural language processing medical chatbot.

Dataset	Description	Potential biases	Strategies for mitigation
Common Crawl	A large dataset collected from web pages across the internet	Overrepresentation of Western cultures, socioeconomic status	Ensure diverse sourcing and include localized health care data from various regions
Wikipedia	Open-source encyclopedia with content generated by volunteers	Gender and racial biases due to contributor demographics	Use guidelines for inclusive contributions and diversify contributor base
MIMIC-III	Critical care database with deidentified health data	Predominantly includes data from urban hospitals; underrepresents rural populations	Incorporate data from a variety of health care settings, including rural and underserved areas
Health-related Twitter data	Tweets related to health topics used for sentiment analysis	Possible bias in language and topics relevant to affluent groups	Filter and include tweets from diverse socioeconomic backgrounds and global populations
Clinical trials data	Data from clinical trials used to evaluate treatments	Limited representation of minority groups in trial participants	Prioritize inclusion of diverse populations in future trials and datasets
PubMed studies	Biomedical literature and research papers	Predominantly Western-centric studies may neglect non-Western medical practices	Integrate literature from diverse geographical regions and cultural contexts
Patient health records	Deidentified patient data for training models	Disparities in data collection practices may overlook marginalized groups	Standardize data collection practices to ensure comprehensive representation

Future Directions

While LLMs have undoubtedly showcased remarkable capabilities, it is crucial to recognize their inherent limitations, especially when applied in the context of oncology chatbots. One of the most pronounced constraints is the absence of genuine understanding [92]. LLMs excel at producing coherent and contextually relevant text, yet they lack true comprehension or reasoning abilities. In the realm of oncology chatbots, this limitation can manifest in responses based solely on patterns from their training data, without a deep grasp of medical principles [93]. Furthermore, there is the risk of unintentionally generating misleading or inaccurate content, a particularly critical concern in health care, where erroneous information can carry significant consequences [94]. Therefore, the deployment of oncology chatbots should be approached as a supplementary aid alongside human medical professionals [95]. It is imperative to navigate their limitations thoughtfully while maintaining vigilant oversight to ensure the precision and reliability of the information they furnish.

Ongoing research in the realm of LLMs is dedicated to confronting their limitations and enhancing their reliability, interpretability, and ethical standing. One particularly promising avenue focuses on the development of more resilient training datasets that seek to mitigate bias and encompass a broader spectrum of perspectives and languages [96]. Researchers are actively exploring methods to render LLMs more interpretable, facilitating users in comprehending and trusting their decision-making processes. Additionally, there is a mounting emphasis on ethical considerations, including the establishment of guidelines and regulations governing LLM deployment, content generation, and the protection of data privacy [97]. Upholding transparency, accountability, and fairness in LLMs is fundamental to their responsible use. Future directions may encompass the creation of hybrid models that combine the strengths of LLMs with other AI techniques, ultimately enhancing their reliability while diminishing the likelihood of generating misleading information [98]. As LLMs assume an increasingly central role across diverse domains, ongoing research and ethical considerations are pivotal forces shaping their development and deployment for the betterment of society.

Future directions for LLMs focus on several key areas of advancement. These include enhancing the models' ability to understand context, nuances, and user intent in natural language, which will lead to more effective human-computer interactions. There is also a growing emphasis on integrating text-based models with vision and audio capabilities, enabling richer and more comprehensive communication. Addressing and reducing biases in LLMs is critical to ensuring fairness and inclusivity in generated content, while customization and fine-tuning of models are becoming increasingly important for specific applications. Efforts are also being made to develop more energy-efficient LLM architectures and training methods, which would reduce their environmental impact and make them accessible on low-power devices. Real-time conversational AI is another area of focus, with the goal of enabling more seamless, natural, and context-aware interactions. Privacy-preserving models are being researched to protect user data, and human-AI collaboration is being advanced to enhance productivity and creativity. Ethical guidelines and regulations are being established to ensure the responsible and safe use of LLMs [99]. In education, LLMs are being used to create personalized and adaptive learning experiences. In the medical field, these models are expanding their role in research, diagnostics, and patient care, with a strong emphasis on adhering to medical ethics and ensuring compliance with standards such as patient confidentiality and informed consent. Finally, the creative capabilities of LLMs are being explored, pushing the boundaries in generating content across various artistic domains.

The integration of LLMs into the domain of oncology chatbots raises intriguing opportunities and concerns, underscoring the significance of human-centered AI within health care applications. While LLMs offer a powerful tool for enhancing human-computer interactions, particularly in health care settings, their application necessitates careful consideration and balance

[100,101]. Historically, expert systems have played a pivotal role in decision support and knowledge representation within these applications. The incorporation of LLMs introduces a novel dimension to this landscape by capitalizing on their remarkable capacity to comprehend and generate human language. However, it is crucial to recognize that akin to expert systems, LLMs possess inherent limitations. While excelling at processing extensive data and generating coherent responses, their actual grasp of intricate medical principles may be constrained. Therefore, the primary challenge lies in harnessing the capabilities of LLMs while ensuring that their responses align with medical accuracy, ethical considerations, and the ethos of human-centered AI in health care.

Conclusions

The emergence of LLMs signifies a transformative leap in computational paradigms, highlighting the central role of human-centered AI in this progression. Models such as GPT-3 and GPT-4 have not only revolutionized machine learning but have also profoundly influenced oncology chatbots through their advanced language processing capabilities. However, as technological advancements persist, the ethical dimensions—particularly concerning biases and misinformation—require meticulous attention. Integrating humanistic and ethical principles into the development of LLMs, especially within oncology chatbots, is crucial for responsible AI integration. Envisioning a future where machines possess unparalleled language abilities alongside adept management of ethical complexities demands a proactive ethical framework.

This comprehensive review explores the evolution, applications, and future trajectories of LLMs in health care and beyond. It is essential to acknowledge the inherent limitations and dynamic nature of technology, suggesting that the landscape of LLMs is rapidly evolving. Future directions outlined herein may witness significant changes or novel developments shortly. Therefore, ongoing research efforts should continuously update and expand this review, encompassing newer LLM iterations, exploring specific health care applications, and conducting empirical studies to validate practical implications and real-world efficacy.

Furthermore, deeper exploration into the ethical implications and societal impacts of widespread LLM implementation remains a critical avenue for future inquiry. Continued research endeavors in these areas will not only enhance our comprehension and use of LLMs but also address emerging challenges and opportunities, aligning with the foundational principles of human-centered AI.

Acknowledgments

This work is supported by the Planning and Dissemination Grants—Institute Community Support, Canadian Institutes of Health Research, Canada (CIHR PCS—168296 and CIHR PCS—191021).

Conflicts of Interest

None declared.

References

1. Xu L, Sanders L, Li K, Chow JCL. Chatbot for health care and oncology applications using artificial intelligence and machine learning: systematic review. JMIR Cancer 2021;7(4):e27850 [FREE Full text] [doi: 10.2196/27850] [Medline: 34847056]
2. Rebelo N, Sanders L, Li K, Chow JCL. Learning the treatment process in radiotherapy using an artificial intelligence-assisted chatbot: development study. JMIR Form Res 2022;6(12):e39443 [FREE Full text] [doi: 10.2196/39443] [Medline: 36327383]
3. Chow JCL, Sanders L, Li K. Design of an educational chatbot using artificial intelligence in radiotherapy. AI 2023;4(1):319-332. [doi: 10.3390/ai4010015]
4. Kovacek D, Chow JCL. An AI-assisted chatbot for radiation safety education in radiotherapy. IOP SciNotes 2021;2(3):034002. [doi: 10.1088/2633-1357/ac1f88]
5. Chow JC. Artificial intelligence in radiotherapy and patient care. In: InArtificial Intelligence in Medicine. Cham: Springer International Publishing; 2021:1-13.
6. Chow JCL, Wong V, Sanders L, Li K. Developing an AI-assisted educational chatbot for radiotherapy using the IBM Watson assistant platform. Healthcare (Basel) 2023;11(17):2417 [FREE Full text] [doi: 10.3390/healthcare11172417] [Medline: 37685452]
7. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023;183(6):589-596 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
8. Sidlauskiene J, Joye Y, Auruskeviciene V. AI-based chatbots in conversational commerce and their effects on product and price perceptions. Electron Mark 2023;33(1):24 [FREE Full text] [doi: 10.1007/s12525-023-00633-8] [Medline: 37252674]
9. Bond RR, Mulvenna MD, Wan H, Finlay DD, Wong A, Koene A, et al. Human Centered Artificial Intelligence: Weaving UX into Algorithmic Decision Making. Romania: International Conference on Human-Computer Interaction; 2019:2-9.
10. Shneiderman B. Bridging the gap between ethics and practice. ACM Trans Interact Intell Syst 2020;10(4):1-31. [doi: 10.1145/3419764]

11. Schwartz Reisman Institute for Technology and Society. University of Toronto. 2024. URL: https://srinstitute.utoronto.ca/ [accessed 2024-10-08]
12. Wilson L, Marasoiu M. The development and use of chatbots in public health: scoping review. JMIR Hum Factors 2022;9(4):e35882 [FREE Full text] [doi: 10.2196/35882] [Medline: 36197708]
13. Tayebi Arasteh S, Han T, Lotfinia M, Kuhl C, Kather JN, Truhn D, et al. Large language models streamline automated machine learning for clinical studies. Nat Commun 2024;15(1):1603 [FREE Full text] [doi: 10.1038/s41467-024-45879-8] [Medline: 38383555]
14. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. JAMA 2023;330(9):866-869. [doi: 10.1001/jama.2023.14217] [Medline: 37548965]
15. Suppadungsuk S, Thongprayoon C, Miao J, Krisanapan P, Qureshi F, Kashani K, et al. Exploring the potential of chatbots in critical care nephrology. Medicines (Basel) 2023;10(10):58 [FREE Full text] [doi: 10.3390/medicines10100058] [Medline: 37887265]
16. Kaur A, Singh S, Chandan JS, Robbins T, Patel V. Qualitative exploration of digital chatbot use in medical education: a pilot study. Digit Health 2021;7:20552076211038151 [FREE Full text] [doi: 10.1177/20552076211038151] [Medline: 34513002]
17. KC GP, Ranjan S, Ankit T, Kumar V. A personalized medical assistant chatbot: Medibot. Int J Sci Technol Eng 2019;5(7).
18. Nißen M, Rüegger D, Stieger M, Flückiger C, Allemand M, V Wangenheim F, et al. The effects of health care chatbot personas with different social roles on the client-chatbot bond and usage intentions: development of a design codebook and web-based study. J Med Internet Res 2022;24(4):e32630 [FREE Full text] [doi: 10.2196/32630] [Medline: 35475761]
19. Patel N, Trivedi S. Leveraging predictive modeling, machine learning personalization, NLP customer support, and AI chatbots to increase customer loyalty. Empir Quests Manag Essences 2020;3(3):1-24.
20. King MR. The future of AI in medicine: a perspective from a chatbot. Ann Biomed Eng 2023;51(2):291-295. [doi: 10.1007/s10439-022-03121-w] [Medline: 36572824]
21. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
22. Ali H. Generative pre‐trained transformer 4 in healthcare: challenges, opportunities, and recommendations. Med Adv 2023;1(2):163-166. [doi: 10.1002/med4.21]
23. Sheth S, Baker HP, Prescher H, Strelzow JA. Ethical considerations of artificial intelligence in health care: examining the role of generative pretrained transformer-4. J Am Acad Orthop Surg 2024;32(5):205-210. [doi: 10.5435/JAAOS-D-23-00787] [Medline: 38175996]
24. Truhn D, Loeffler CM, Müller-Franzes G, Nebelung S, Hewitt KJ, Brandner S, et al. Extracting structured information from unstructured histopathology reports using generative pre-trained transformer 4 (GPT-4). J Pathol 2024;262(3):310-319. [doi: 10.1002/path.6232] [Medline: 38098169]
25. Singh SK, Kumar S, Mehra PS. : IEEE; 2023 Presented at: 2023 International Conference on IoT, Communication and Automation Technology (ICICAT); 2023 June 04; Gorakhpur, India p. 1-6. [doi: 10.1109/icicat57735.2023.10263706]
26. Li J. Security implications of AI chatbots in health care. J Med Internet Res 2023;25:e47551 [FREE Full text] [doi: 10.2196/47551] [Medline: 38015597]
27. Arshad HB, Butt SA, Khan SU, Javed Z, Nasir K. ChatGPT and artificial intelligence in hospital level research: potential, precautions, and prospects. Methodist Debakey Cardiovasc J 2023;19(5):77-84 [FREE Full text] [doi: 10.14797/mdcvj.1290] [Medline: 38028967]
28. Waters MR, Aneja S, Hong JC. Unlocking the power of ChatGPT, artificial intelligence, and large language models: practical suggestions for radiation oncologists. Pract Radiat Oncol 2023;13(6):e484-e490. [doi: 10.1016/j.prro.2023.06.011] [Medline: 37598727]
29. Woo WL, Gao B, Al-Nima RRO, Ling W. Development of conversational artificial intelligence for pandemic healthcare query support. Int J Auto AI Mach Learn 2020;1(1):54-79. [doi: 10.61797/ijaaiml.v1i1.35]
30. Chow JCL, Sanders L, Li K. Impact of ChatGPT on medical chatbots as a disruptive technology. Front Artif Intell 2023;6:1166014 [FREE Full text] [doi: 10.3389/frai.2023.1166014] [Medline: 37091303]
31. Stahl BC, Eke D. The ethics of ChatGPT – exploring the ethical issues of an emerging technology. Int J Inf Manage 2024;74:102700. [doi: 10.1016/j.ijinfomgt.2023.102700]
32. Haupt CE, Marks M. AI-generated medical advice-GPT and beyond. JAMA 2023;329(16):1349-1350. [doi: 10.1001/jama.2023.5321] [Medline: 36972070]
33. Yao X. A review of evolutionary artificial neural networks. Int J Intell Syst 1993;8(4):539-567. [doi: 10.1002/int.4550080406]
34. Alawida M, Mejri S, Mehmood A, Chikhaoui B, Isaac Abiodun O. A comprehensive study of ChatGPT: advancements, limitations, and ethical considerations in natural language processing and cybersecurity. Information 2023;14(8):462. [doi: 10.3390/info14080462]
35. Min B, Ross H, Sulem E, Veyseh APB, Nguyen TH, Sainz O, et al. Recent advances in natural language processing via large pre-trained language models: a survey. ACM Comput Surv 2023;56(2):1-40. [doi: 10.1145/3605943]
36. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform 2022;23(6):bbac409. [doi: 10.1093/bib/bbac409] [Medline: 36156661]

37. Thoppilan R, De FD, Hall J, Shazeer N, Kulshreshtha A, Cheng HT, et al. Lamda: language models for dialog applications. arXiv Preprint posted online Jan 20, 2022. [doi: 10.48550/arXiv.2201.08239]
38. Alshemali B, Kalita J. Improving the reliability of deep neural networks in NLP: a review. Knowl Based Syst 2020;191:105210. [doi: 10.1016/j.knosys.2019.105210]
39. Gilbert S, Harvey H, Melvin T, Vollebregt E, Wicks P. Large language model AI chatbots require approval as medical devices. Nat Med 2023;29(10):2396-2398. [doi: 10.1038/s41591-023-02412-6] [Medline: 37391665]
40. Siddique S, Chow JCL. Machine learning in healthcare communication. Encyclopedia 2021;1(1):220-239. [doi: 10.3390/encyclopedia1010021]
41. Chen S, Wu M, Zhu KQ, Lan K, Zhang Z, Cui L. LLM-empowered chatbots for psychiatrist and patient simulation: application and evaluation. arXiv Preprint posted online May 23, 2023. [doi: 10.48550/arXiv.2305.13614]
42. Lim S, Schmälzle R. Artificial intelligence for health message generation: an empirical study using a large language model (LLM) and prompt engineering. Front Commun 2023;8:1129082. [doi: 10.3389/fcomm.2023.1129082]
43. Zhu W, Liu H, Dong Q, Xu J, Kong L, Chen J, et al. Multilingual machine translation with large language models: empirical results and analysis. arXiv Preprint posted online Apr 10, 2023. [doi: 10.18653/v1/2024.findings-naacl.176]
44. van Heerden AC, Pozuelo JR, Kohrt BA. Global mental health services and the impact of artificial intelligence-powered large language models. JAMA Psychiatry 2023;80(7):662-664. [doi: 10.1001/jamapsychiatry.2023.1253] [Medline: 37195694]
45. Jeon J, Lee S. Large language models in education: a focus on the complementary relationship between human teachers and ChatGPT. Educ Inf Technol 2023;28(12):15873-15892. [doi: 10.1007/s10639-023-11834-1]
46. Meyer JG, Urbanowicz RJ, Martin PCN, O'Connor K, Li R, Peng P, et al. ChatGPT and large language models in academia: opportunities and challenges. BioData Min 2023;16(1):20 [FREE Full text] [doi: 10.1186/s13040-023-00339-9] [Medline: 37443040]
47. Ouyang S, Zhang JM, Harman M, Wang M. LLM is like a box of chocolates: the non-determinism of ChatGPT in code generation. arXiv Preprint posted online Aug 5, 2023. [doi: 10.48550/arXiv.2308.02828]
48. Peng L, Zhang Y, Shang J. Generating efficient training data via LLM-based attribute manipulation. arXiv Preprint posted online Jul 14, 2023. [doi: 10.48550/arXiv.2307.07099]
49. Tasioulas J. Artificial intelligence, humanistic ethics. Daedalus 2022;151(2):232-243. [doi: 10.1162/daed_a_01912]
50. De Angelis L, Baglivo F, Arzilli G, Privitera GP, Ferragina P, Tozzi AE, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front Public Health 2023;11:1166120 [FREE Full text] [doi: 10.3389/fpubh.2023.1166120] [Medline: 37181697]
51. Karabacak M, Margetis K. Embracing large language models for medical applications: opportunities and challenges. Cureus 2023;15(5):e39305 [FREE Full text] [doi: 10.7759/cureus.39305] [Medline: 37378099]
52. Siau K, Wang W. Artificial intelligence (AI) ethics: ethics of AI and ethical AI. J Database Manag 2020;31(2):74-87. [doi: 10.4018/jdm.2020040105]
53. Teran L, Pincay J, Wallimann-Helmer I, Portmann E. A literature review on digital ethics from a humanistic and sustainable perspective. 2021 Presented at: Proceedings of the 14th International Conference on Theory and Practice of Electronic Governance; 2012 Jan 12; New York p. 57-64. [doi: 10.1145/3494193.3494295]
54. Ho A. Deep ethical learning: taking the interplay of human and artificial intelligence seriously. Hastings Cent Rep 2019;49(1):36-39. [doi: 10.1002/hast.977] [Medline: 30790317]
55. Varlamov OO, Chuvikov DA, Adamova LE, Petrov MA, Zabolotskaya IK, Zhilina TN. Logical, philosophical and ethical aspects of AI in medicine. Int J Mach Learn Comput 2019;9(6):868. [doi: 10.18178/ijmlc.2019.9.6.885]
56. Huang RS, Lu KJQ, Meaney C, Kemppainen J, Punnett A, Leung F. Assessment of resident and AI chatbot performance on the University of Toronto family medicine residency progress test: comparative study. JMIR Med Educ 2023;9:e50514 [FREE Full text] [doi: 10.2196/50514] [Medline: 37725411]
57. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the chat-GPT model. Res Sq 2023:1-17 [FREE Full text] [doi: 10.21203/rs.3.rs-2566942/v1] [Medline: 36909565]
58. Ostherr K. Artificial intelligence and medical humanities. J Med Humanit 2022;43(2):211-232 [FREE Full text] [doi: 10.1007/s10912-020-09636-4] [Medline: 32654043]
59. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intell 2019;1(9):389-399. [doi: 10.1038/s42256-019-0088-2]
60. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023;6:1169595 [FREE Full text] [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
61. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on large language models (LLMs). NPJ Digit Med 2024;7(1):183 [FREE Full text] [doi: 10.1038/s41746-024-01157-x] [Medline: 38977771]
62. Sharma P. Chatbots in medical research: advantages and limitations of artificial intelligence-enabled writing with a focus on ChatGPT as an author. Clin Nucl Med 2023;48(9):838-839. [doi: 10.1097/RLU.0000000000004665] [Medline: 37083827]

63. Goodman RS, Patrinely JR, Stone CA, Zimmerman E, Donald RR, Chang SS, et al. Accuracy and reliability of chatbot responses to physician questions. JAMA Netw Open 2023 Oct 02;6(10):e2336483 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.36483] [Medline: 37782499]
64. Oca MC, Meller L, Wilson K, Parikh A, McCoy AO, Chang J, et al. Bias and inaccuracy in AI chatbot ophthalmologist recommendations. Cureus 2023;15(9):e45911 [FREE Full text] [doi: 10.7759/cureus.45911] [Medline: 37885556]
65. Leslie D. Understanding artificial intelligence ethics and safety. arXiv Preprint posted online Jun 11, 2019. [doi: 10.48550/arXiv.1906.05684]
66. Au Yeung J, Kraljevic Z, Luintel A, Balston A, Idowu E, Dobson RJ, et al. AI chatbots not yet ready for clinical use. Front Digit Health 2023;5:1161098 [FREE Full text] [doi: 10.3389/fdgth.2023.1161098] [Medline: 37122812]
67. May R, Denecke K. Security, privacy, and healthcare-related conversational agents: a scoping review. Inform Health Soc Care 2022;47(2):194-210. [doi: 10.1080/17538157.2021.1983578] [Medline: 34617857]
68. Mijwil M, Aljanabi M, Ali AH. ChatGPT: exploring the role of cybersecurity in the protection of medical information. Mesopotamian J Cybersecur 2023:18-21. [doi: 10.58496/mjcs/2023/004]
69. Sheth A, Shekarpour S, Yip HY. Extending patient-chatbot experience with internet-of-things and background knowledge: case studies with healthcare applications. IEEE Intell Syst 2019;34(4):24-30 [FREE Full text] [doi: 10.1109/mis.2019.2905748] [Medline: 34690576]
70. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of artificial intelligence chatbot responses to top searched queries about cancer. JAMA Oncol 2023;9(10):1437-1440. [doi: 10.1001/jamaoncol.2023.2947] [Medline: 37615960]
71. Swick RK. The accuracy of artificial intelligence (AI) chatbots in telemedicine. J S C Acad Sci 2021;19(2):17.
72. Bélisle-Pipon JC, Monteferrante E, Roy MC, Couture V. Artificial intelligence ethics has a black box problem. AI Soc 2022;38(4):1507-1522. [doi: 10.1007/s00146-021-01380-0]
73. Yang J, Chen YL, Por LY, Ku CS. A systematic literature review of information security in chatbots. Appl Sci 2023;13(11):6355. [doi: 10.3390/app13116355]
74. Liebrenz M, Schleifer R, Buadze A, Bhugra D, Smith A. Generating scholarly content with ChatGPT: ethical challenges for medical publishing. Lancet Digit Health 2023;5(3):e105-e106 [FREE Full text] [doi: 10.1016/S2589-7500(23)00019-5] [Medline: 36754725]
75. Loh E. ChatGPT and generative AI chatbots: challenges and opportunities for science, medicine and medical leaders. BMJ Lead 2023:000797. [doi: 10.1136/leader-2023-000797] [Medline: 37192124]
76. Devaram S. Empathic chatbot: emotional intelligence for empathic chatbot: emotional intelligence for mental health well-being. arXiv Preprint posted online Dec 15, 2020. [doi: 10.48550/arXiv.2012.09130]
77. Walker HL, Ghani S, Kuemmerli C, Nebiker CA, Müller BP, Raptis DA, et al. Reliability of medical information provided by ChatGPT: assessment against clinical guidelines and patient information quality instrument. J Med Internet Res 2023;25:e47479 [FREE Full text] [doi: 10.2196/47479] [Medline: 37389908]
78. Pryss R, Kraft R, Baumeister H, Winkler J, Probst T, Reichert M, et al. Using chatbots to support medical and psychological treatment procedures: challenges, opportunities, technologies, reference architecture. In: Baumeister H, Montag C, editors. Digital Phenotyping and Mobile Sensing: New Developments in Psychoinformatics. Cham: Springer; 2019:249-260.
79. Parviainen J, Rantala J. Chatbot breakthrough in the 2020s? An ethical reflection on the trend of automated consultations in health care. Med Health Care Philos 2022;25(1):61-71 [FREE Full text] [doi: 10.1007/s11019-021-10049-w] [Medline: 34480711]
80. Zeng G, Yang W, Ju Z, Yang Y, Wang S, Zhang R, et al. MedDialog: large-scale medical dialogue datasets. 2020 Presented at: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020 Nov 10; San Diego p. 9241-9250. [doi: 10.18653/v1/2020.emnlp-main.743]
81. Chakraborty S, Paul H, Ghatak S, Pandey SK, Kumar A, Singh KU, et al. An AI-based medical chatbot model for infectious disease prediction. IEEE Access 2022;10:128469-128483. [doi: 10.1109/access.2022.3227208]
82. Xue J, Zhang B, Zhao Y, Zhang Q, Zheng C, Jiang J, et al. Evaluation of the current state of chatbots for digital health: scoping review. J Med Internet Res 2023;25:e47217 [FREE Full text] [doi: 10.2196/47217] [Medline: 38113097]
83. Omar RA. Unabashed bias: how health-care organizations can significantly reduce bias in the face of unaccountable AI. Denv L Rev 2020;98:807.
84. Adam H, Balagopalan A, Alsentzer E, Christia F, Ghassemi M. Mitigating the impact of biased artificial intelligence in emergency decision-making. Commun Med (Lond) 2022;2(1):149 [FREE Full text] [doi: 10.1038/s43856-022-00214-4] [Medline: 36414774]
85. Denecke K, Abd-Alrazaq A, Househ M, Warren J. Evaluation metrics for health chatbots: a delphi study. Methods Inf Med 2021;60(5-06):171-179. [doi: 10.1055/s-0041-1736664] [Medline: 34719011]
86. Radziwill NM, Benton MC. Evaluating quality of chatbots and intelligent conversational agents. arXiv Preprint posted online Apr 15, 2017. [doi: 10.48550/arXiv.1704.04579]
87. Hauglid MK, Mahler T. Doctor chatbot: the EU s regulatory prescription for generative medical AI. Oslo Law Review 2023;10(1):1-23. [doi: 10.18261/olr.10.1.1]

88. Vaishya R, Misra A, Vaish A. ChatGPT: is this version good for healthcare and research? Diabetes Metab Syndr 2023;17(4):102744. [doi: 10.1016/j.dsx.2023.102744] [Medline: 36989584]
89. Lindebaum D, Fleming P. ChatGPT undermines human reflexivity, scientific responsibility and responsible management research. Br J Manag 2023;35(2):566-575. [doi: 10.1111/1467-8551.12781]
90. Heaven WD. Geoffrey Hinton tells us why he's now scared of the tech he helped build. MIT Technology Review. 2023. URL: https://www.technologyreview.com/2023/05/02/1072528/geoffrey-hinton-google-why-scared-ai/ [accessed 2024-10-16]
91. Zhu L, Mou W, Lai Y, Lin J, Luo P. Language and cultural bias in AI: comparing the performance of large language models developed in different countries on traditional Chinese medicine highlights the need for localized models. J Transl Med 2024;22(1):319 [FREE Full text] [doi: 10.1186/s12967-024-05128-4] [Medline: 38553705]
92. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023;11(6):887 [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
93. Lee H, Kang J, Yeo J. Medical specialty recommendations by an artificial intelligence chatbot on a smartphone: development and deployment. J Med Internet Res 2021;23(5):e27460 [FREE Full text] [doi: 10.2196/27460] [Medline: 33882012]
94. Aljanabi M. ChatGPT: future directions and open possibilities. Mesopotamian J Cybersecur 2023:16-17.
95. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. J Med Internet Res 2023;25:e48009 [FREE Full text] [doi: 10.2196/48009] [Medline: 37566454]
96. Zenil H, Tegnér J, Abrahão FS, Lavin A, Kumar V, Frey JG, et al. The future of fundamental science led by generative closed-loop artificial intelligence. arXiv Preprint posted online Jul 9, 2023. [doi: 10.48550/arXiv.2307.07522]
97. Akinci D'Antonoli T, Stanzione A, Bluethgen C, Vernuccio F, Ugga L, Klontzas ME, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. Diagn Interv Radiol 2024;30(2):80-90 [FREE Full text] [doi: 10.4274/dir.2023.232417] [Medline: 37789676]
98. Schukow C, Smith SC, Landgrebe E, Parasuraman S, Folaranmi OO, Paner GP, et al. Application of ChatGPT in routine diagnostic pathology: promises, pitfalls, and potential future directions. Adv Anat Pathol 2024;31(1):15-21. [doi: 10.1097/PAP.0000000000000406] [Medline: 37501529]
99. Pearman S, Young E, Cranor LF. User-friendly yet rarely read: a case study on the redesign of an online HIPAA authorization. 2022 Presented at: Proceedings on Privacy Enhancing Technologies; 2022 July 01; Virginia p. 558-581. [doi: 10.56553/popets-2022-0086]
100. Frangoudes F, Hadjiaros M, Schiza E, Matsangidou M, Tsivitanidou O, Neokleous K. An overview of the use of chatbots in medical and healthcare education. : Springer International Publishing; 2021 Presented at: International Conference on Human-Computer Interaction; 2021 July 03; Cham p. 170-184. [doi: 10.1007/978-3-030-77943-6_11]
101. Nicolescu L, Tudorache MT. Human-computer interaction in customer service: the experience with AI chatbots—a systematic literature review. Electronics 2022;11(10):1579. [doi: 10.3390/electronics11101579]

Abbreviations

AI: artificial intelligence
BERT: Bidirectional Encoder Representations from Transformers
HIPAA: Health Insurance Portability and Accountability Act
LLM: large language model
NLP: natural language processing

Edited by E Uzun; submitted 16.07.24; peer-reviewed by J Lai, ST Arasteh; comments to author 22.08.24; revised version received 23.08.24; accepted 23.09.24; published 06.11.24.

Please cite as:
Chow JCL, Li K
Ethical Considerations in Human-Centered AI: Advancing Oncology Chatbots Through Large Language Models
JMIR Bioinform Biotech 2024;5:e64406
URL: https://bioinform.jmir.org/2024/1/e64406
doi:10.2196/64406
PMID:39321336

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