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

Viewpoint

It Is in Our DNA: Bringing Electronic Health Records and Genomic Data Together for Precision Medicine (e55632)
Alan Robertson, Andrew Mallett, Zornitza Stark, Clair Sullivan. ................................................................. 2

Editorial

ChatGPT and Medicine: Together We Embrace the AI Renaissance (e52700)
Sean Hacking. ............................................................................................................................................. 10

Original Papers

Machine Learning Models for Prediction of Maternal Hemorrhage and Transfusion: Model Development Study (e52059)
Homa Ahmadzia, Alexa Dzienny, Mike Bopf, Jaclyn Phillips, Jerome Federspiel, Richard Amdur, Madeline Rice, Laritza Rodriguez. ...................... 15

The Roles of NOTCH3 p.R544C and Thrombophilia Genes in Vietnamese Patients With Ischemic Stroke: Study Involving a Hierarchical Cluster Analysis (e56884)
Huong Bui, Qu nh Ng Yu ng Th ng, Ho Cam Tu, Sinh Nguyen Phuong, Thuy Pham, Thu Vu, Huyen Nguyen Thi Thu, Lam Khanh Ho, Dung Nguyen Tien. ........................................................................... 26

Review

Assessing Privacy Vulnerabilities in Genetic Data Sets: Scoping Review (e54332)
Mara Thomas, Nuria Mackes, Asad Preuss-Dodhy, Thomas Wieland, Markus Bundschus. ................................................................. 50
It Is in Our DNA: Bringing Electronic Health Records and Genomic Data Together for Precision Medicine

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.

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).

**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>
<thead>
<tr>
<th>Table 1. Clinical applications of genomics.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Application</td>
</tr>
<tr>
<td>Diagnosis of genetic disease</td>
</tr>
<tr>
<td>Disease screening and early detection</td>
</tr>
<tr>
<td>Family planning</td>
</tr>
<tr>
<td>Cancer diagnosis, treatment, and monitoring</td>
</tr>
<tr>
<td>Infectious disease diagnosis characterization</td>
</tr>
<tr>
<td>Precision treatment and pharmacogenomics</td>
</tr>
</tbody>
</table>
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:

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 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).
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.

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

<table>
<thead>
<tr>
<th>Activity</th>
<th>Genetic+genomic testing</th>
<th>A potential model of genomics-enabled care</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generation of sequence data</td>
<td>DNA from the genes associated with the condition is sequenced when a test is ordered</td>
<td>Individual’s whole genome sequence is available from a prior interaction with the health care system.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A CDSS recommends if there is a benefit to generate complementary sequence data (eg, long read, transcriptomic, cell-free).</td>
</tr>
<tr>
<td>Analysis and interpretation of genetic data</td>
<td>Variants within the sequenced DNA are determined</td>
<td>A CDSS accesses the specific genes currently associated with condition from multiple high-quality, peer-reviewed resources.</td>
</tr>
<tr>
<td></td>
<td>The clinical significance of the variants is accessed</td>
<td>A CDSS recommends if genome data should be aligned to a new reference genome or use updated variant detection methods.</td>
</tr>
<tr>
<td>Clinical decisions and reporting</td>
<td>Clinician synthesizes genetic results, patient’s history, and clinical experience to make decision</td>
<td>Variants within the selected genes are determined.</td>
</tr>
<tr>
<td></td>
<td>A clinical report is generated</td>
<td>The clinical significance of the variants is accessed.</td>
</tr>
<tr>
<td></td>
<td>Report is uploaded to the EHR as a PDF</td>
<td>If successful, share causative variants with public repositories and related individuals. Make results accessible to other clinicians treating the individual (where appropriate).</td>
</tr>
<tr>
<td>Data storage</td>
<td>Raw sequence data and results stored in the laboratory system</td>
<td>If inconclusive, flag candidate variants of uncertain significance for automatic monitoring, monitor information associated with disease, and determine when the individual should be reanalyzed.</td>
</tr>
<tr>
<td></td>
<td>Note: external collaborators do not always provide raw-sequence data</td>
<td>Ensure all information is in a standardized interoperable and time-stamped format (ie, GA4GH or eMerge).</td>
</tr>
</tbody>
</table>

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.

aCDSS: clinical decision support system.
bEHR: electronic health record.
cLIM: 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


Abbreviations

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

ChatGPT and Medicine: Together We Embrace the AI Renaissance

Sean Hacking¹, MBBCh
NYU Langone, New York, NY, United States

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.
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 conclusion, 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.

References

Abbreviations

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

©Sean Hacking. Originally published in JMIR Bioinformatics and Biotechnology (https://bioinform.jmir.org), 07.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Bioinformatics and Biotechnology, is properly cited. The complete bibliographic information, a link to the original publication on https://bioinform.jmir.org/, as well as this copyright and license information must be included.
Original Paper

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

Homa Khorrami Ahmadzia1,2*, MD, MPH; Alexa C Dzienny3, BS; Mike Bopf4*, MS; Jaclyn M Phillips1, MD; Jeffrey Federspiel5*, MD, PhD; Richard Amdur6, PhD; Madeline Murguia Rice7, PhD; Laritza Rodriguez4, MD, PhD

1Division of Maternal-Fetal Medicine, Department of Obstetrics and Gynecology, George Washington University, Washington, DC, United States
2Division of Maternal-Fetal Medicine, Department of Obstetrics and Gynecology, Inova Health System, Falls Church, VA, United States
3The George Washington University School of Medicine and Health Sciences, Washington, DC, United States
4Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, Bethesda, MD, United States
5Division of Maternal-Fetal Medicine, Department of Obstetrics and Gynecology, Duke University, Durham, NC, United States
6Medical Faculty Associates, The George Washington University School of Medicine and Health Sciences, Washington, DC, United States
7George 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

https://bioinform.jmir.org/2024/1/e52059
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/m²). 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 (e.g., 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.

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

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

<sup>a</sup>ROC-AUC: area under the receiver operating characteristic curve.

<sup>b</sup>PR-AUC: area under the precision-recall curve.

<sup>c</sup>MCC: Matthews correlation coefficient.

<sup>d</sup>F2: modified F-score skewed toward recall.

<sup>e</sup>GB: gradient boosting.

<sup>f</sup>RF: random forest.

<sup>g</sup>Emb: learned embedding.

<sup>h</sup>MLP: multilayer perceptron.

<sup>i</sup>TFIM: TensorFlow imbalanced.

<sup>j</sup>SVM: support vector machine.

<sup>k</sup>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.

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

\(^a\)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.

\(^e\)F\(^2\): 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.
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 a 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 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.

Multimedia Appendix 2
Overall Patient Characteristics.

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 perceptron, nTensorflow imbalanced, oSupport vector machines, plogistic regression.

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 perceptron, nTensorflow imbalanced, oSupport vector machines, plogistic regression.

References


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:

URL: https://bioinform.jmir.org/2024/1/e52059
doi:10.2196/52059
PMID:
The Roles of NOTCH3 p.R544C and Thrombophilia Genes in Vietnamese Patients With Ischemic Stroke: Study Involving a Hierarchical Cluster Analysis

Huong Thi Thu Bui¹,²*, MD, PhD; Quỳnh Nguyễn ธ fn Th P nhg³*, MD; Ho Cam Tu⁴,⁵*, MSci; Sinh Nguyên Phuong⁶, MD, PhD; Thuỷ Thi Phạm¹, MD, PhD; Thu Vù⁴, Sp1; Huyền Nguyên Thi Thu⁷, PhD; Lam Khánh Ho⁸, PhD; Dung Nguyên Tien⁹, MD, PhD

¹Department of Biochemistry, Thai Nguyen University of Medicine and Pharmacy, Thai Nguyen, Vietnam
²Department of Immunology Molecular Genetic, Thainguyen National Hospital, Thai Nguyen, Vietnam
³Department of Clinical Pharmacy, Thai Nguyen University of Medicine and Pharmacy, Thai Nguyen, Vietnam
⁴Center of Gene and Protein Research, Hanoi Medical University, Hanoi, Vietnam
⁵Institute of Virology, School of Medicine, Technical University of Munich, Munich, Germany
⁶Department of Rehabilitation, Thai Nguyen University of Medicine and Pharmacy, Thai Nguyen, Vietnam
⁷Department of Internal Medicine, Thai Nguyen University of Medicine and Pharmacy, Thai Nguyen, Vietnam
⁸Department 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 \( (P_{\text{Fisher}} < .05) \). 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.
**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. Hinca 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 hypertensity-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:

\[
\text{Margin of Error} = Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}
\]

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 \( \Delta \) and \( \sigma \) (\( d = \Delta / \sigma \)), where \( \sigma \) is the standard deviation and \( \Delta \) 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:

\[
\text{Sample Size} = \frac{Z_{\alpha/2}^2 \times \sigma^2}{d^2}
\]

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

Thus, screening all risk factors may have a medium or higher level of influence on the phenotype (\( P<0.05 \) indicates statistical significance) (Table 1).

Table 1. Two-sample \( t \) test power calculation results.

<table>
<thead>
<tr>
<th>Sample size for each group</th>
<th>Cohen ( d )</th>
</tr>
</thead>
<tbody>
<tr>
<td>99.08</td>
<td>0.4</td>
</tr>
<tr>
<td>63.76</td>
<td>0.5</td>
</tr>
<tr>
<td>44.58</td>
<td>0.6</td>
</tr>
<tr>
<td>33.02</td>
<td>0.7</td>
</tr>
</tbody>
</table>

aThe 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 throbophilia 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′-GTGGGGTGGAGTGGAAGTGG (F1) and 5′-GAGCAGTCTCCACGTGCA (R1) for the C allele, and 5′-TTGAGGGACGCTGTGTGATC (F2) and 5′-CTAGATGCAACATCCCAAAACC (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 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/HĐĐ-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 ($P_{untw}$) 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 [FAM], 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


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/m². Of the 100 patients, 3 had a BMI of <18.5 kg/m², 56 had a BMI of 18.5-22.9 kg/m², 27 had a BMI of 23-24.9 kg/m², and 14 had a BMI of 25-29.9 kg/m². 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 (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.

<table>
<thead>
<tr>
<th>Factors</th>
<th>Value (N=100)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>62 (62)</td>
</tr>
<tr>
<td>Female</td>
<td>38 (38)</td>
</tr>
<tr>
<td><strong>Age group (years), n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>24-49</td>
<td>22 (22)</td>
</tr>
<tr>
<td>50-59</td>
<td>23 (23)</td>
</tr>
<tr>
<td>60-69</td>
<td>37 (37)</td>
</tr>
<tr>
<td>70-91</td>
<td>18 (18)</td>
</tr>
<tr>
<td><strong>Age (years)</strong></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>60.14 (12.63)</td>
</tr>
<tr>
<td>Minimum-maximum</td>
<td>24-91</td>
</tr>
<tr>
<td><strong>BMI group (kg/m²), n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>&lt;18.5</td>
<td>3 (3)</td>
</tr>
<tr>
<td>18.5-22.9</td>
<td>56 (56)</td>
</tr>
<tr>
<td>23.0-24.9</td>
<td>27 (27)</td>
</tr>
<tr>
<td>25.0-29.9</td>
<td>14 (14)</td>
</tr>
<tr>
<td><strong>BMI (kg/m²)</strong></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>22.62 (2.49)</td>
</tr>
<tr>
<td>Minimum-maximum</td>
<td>12.4-29.4</td>
</tr>
<tr>
<td><strong>Ethnic group, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Kinh</td>
<td>75 (75)</td>
</tr>
<tr>
<td>Tay</td>
<td>25 (25)</td>
</tr>
<tr>
<td><strong>Smoking history, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>31 (31)</td>
<td></td>
</tr>
<tr>
<td><strong>Alcohol consumption, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>29 (29)</td>
<td></td>
</tr>
<tr>
<td><strong>Blood pressure, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>70 (70)</td>
<td></td>
</tr>
<tr>
<td><strong>Diabetes, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>20 (20)</td>
<td></td>
</tr>
<tr>
<td><strong>Brain stroke, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>35 (35)</td>
<td></td>
</tr>
<tr>
<td><strong>Brain stroke cases in the family, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>44 (44)</td>
<td></td>
</tr>
<tr>
<td><strong>PAI1 4G/5G status, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Wildtype</td>
<td>24 (24)</td>
</tr>
<tr>
<td>Heterozygous</td>
<td>44 (44)</td>
</tr>
<tr>
<td>Homozygous</td>
<td>32 (32)</td>
</tr>
<tr>
<td><strong>FV 1299 status, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Wildtype</td>
<td>96 (96)</td>
</tr>
<tr>
<td>Heterozygous</td>
<td>4 (4)</td>
</tr>
<tr>
<td>Homozygous</td>
<td>0 (0)</td>
</tr>
<tr>
<td><strong>FV-Cambridge status, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Wildtype</td>
<td>100 (100)</td>
</tr>
<tr>
<td>Heterozygous</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Homozygous</td>
<td>0 (0)</td>
</tr>
<tr>
<td><strong>MTHFR 1298 status, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Wildtype</td>
<td></td>
</tr>
<tr>
<td>Heterozygous</td>
<td></td>
</tr>
<tr>
<td>Homozygous</td>
<td></td>
</tr>
<tr>
<td>Factors</td>
<td>Value (N=100)</td>
</tr>
<tr>
<td>-------------------------</td>
<td>---------------</td>
</tr>
<tr>
<td>FII Prothrombin status, n (%)</td>
<td></td>
</tr>
<tr>
<td>Wildtype</td>
<td>98 (98)</td>
</tr>
<tr>
<td>Heterozygous</td>
<td>7 (7)</td>
</tr>
<tr>
<td>Homozygous</td>
<td>0 (0)</td>
</tr>
<tr>
<td>FV-Leiden status, n (%)</td>
<td></td>
</tr>
<tr>
<td>Wildtype</td>
<td>93 (93)</td>
</tr>
<tr>
<td>Heterozygous</td>
<td>7 (7)</td>
</tr>
<tr>
<td>Homozygous</td>
<td>0 (0)</td>
</tr>
<tr>
<td>MTHFR 677 status, n (%)</td>
<td></td>
</tr>
<tr>
<td>Wildtype</td>
<td>55 (55)</td>
</tr>
<tr>
<td>Heterozygous</td>
<td>37 (37)</td>
</tr>
<tr>
<td>Homozygous</td>
<td>8 (8)</td>
</tr>
<tr>
<td>FXIII Val34Leu status, n (%)</td>
<td></td>
</tr>
<tr>
<td>Wildtype</td>
<td>98 (98)</td>
</tr>
<tr>
<td>Heterozygous</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Homozygous</td>
<td>1 (1)</td>
</tr>
<tr>
<td>NOTCH3 status, n (%)</td>
<td></td>
</tr>
<tr>
<td>Wildtype</td>
<td>6 (6)</td>
</tr>
<tr>
<td>Heterozygous</td>
<td>91 (91)</td>
</tr>
<tr>
<td>Homozygous</td>
<td>3 (3)</td>
</tr>
</tbody>
</table>

Table 3. Symptoms at admission.

<table>
<thead>
<tr>
<th>Symptom</th>
<th>Value (N=100), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocal issue</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>5 (5)</td>
</tr>
<tr>
<td>Broca type</td>
<td>86 (86)</td>
</tr>
<tr>
<td>Wernicke type</td>
<td>9 (9)</td>
</tr>
<tr>
<td>Headache</td>
<td>49 (49)</td>
</tr>
<tr>
<td>Dizziness</td>
<td>27 (27)</td>
</tr>
<tr>
<td>Nausea or vomiting</td>
<td>8 (8)</td>
</tr>
<tr>
<td>Mouth distortion</td>
<td>72 (72)</td>
</tr>
<tr>
<td>Circular muscle disorder</td>
<td>21 (21)</td>
</tr>
<tr>
<td>Numbness</td>
<td>41 (41)</td>
</tr>
<tr>
<td>Movement paralysis</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>3 (3)</td>
</tr>
<tr>
<td>Paralysis of the right half of the body</td>
<td>52 (52)</td>
</tr>
<tr>
<td>Paralysis of the left half of the body</td>
<td>39 (39)</td>
</tr>
<tr>
<td>Paralysis of the whole body</td>
<td>6 (6)</td>
</tr>
</tbody>
</table>
Table 4. Important variables in this study.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value (N=100)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age (years)</strong></td>
<td></td>
</tr>
<tr>
<td>Minimum-maximum</td>
<td>24-91</td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>60.14 (12.63)</td>
</tr>
<tr>
<td><strong>BMI (kg/m²)</strong></td>
<td></td>
</tr>
<tr>
<td>Minimum-maximum</td>
<td>12.4-29.4</td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>22.62 (2.49)</td>
</tr>
<tr>
<td><strong>Time to hospitalization (h)</strong></td>
<td></td>
</tr>
<tr>
<td>Minimum-maximum</td>
<td>1-120</td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>10.94 (15.98)</td>
</tr>
<tr>
<td><strong>Time to hospitalization groups, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>&lt;4.5 h</td>
<td>33 (33)</td>
</tr>
<tr>
<td>4.6-6 h</td>
<td>26 (26)</td>
</tr>
<tr>
<td>&gt;6 h</td>
<td>41 (41)</td>
</tr>
<tr>
<td><strong>Systolic blood pressure (mmHg)</strong></td>
<td></td>
</tr>
<tr>
<td>Minimum-maximum</td>
<td>90-210</td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>148.6 (23.66)</td>
</tr>
<tr>
<td><strong>Diastolic blood pressure (mmHg)</strong></td>
<td></td>
</tr>
<tr>
<td>Minimum-maximum</td>
<td>60-120</td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>88.06 (9.5)</td>
</tr>
<tr>
<td><strong>Glasgow Coma Scale score</strong></td>
<td></td>
</tr>
<tr>
<td>Minimum-maximum</td>
<td>8-15</td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>14.72 (1.06)</td>
</tr>
<tr>
<td><strong>NIHSSa score</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Admission</strong></td>
<td></td>
</tr>
<tr>
<td>Minimum-maximum</td>
<td>0-19</td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>7.14 (4.33)</td>
</tr>
<tr>
<td><strong>After 24 h</strong></td>
<td></td>
</tr>
<tr>
<td>Minimum-maximum</td>
<td>0-16</td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>6.71 (4.26)</td>
</tr>
<tr>
<td><strong>Discharge</strong></td>
<td></td>
</tr>
<tr>
<td>Minimum-maximum</td>
<td>0-16</td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>3.73 (3.87)</td>
</tr>
<tr>
<td><strong>Modified Rankin scale score at discharge</strong></td>
<td></td>
</tr>
<tr>
<td>Minimum-maximum</td>
<td>0-5</td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>1.52 (1.35)</td>
</tr>
<tr>
<td><strong>Duration of inpatient treatment at the hospital (days)</strong></td>
<td></td>
</tr>
<tr>
<td>Minimum-maximum</td>
<td>1-23</td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>10.11 (4.33)</td>
</tr>
</tbody>
</table>

aNIHSS: 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 Vai34Leu 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.

<table>
<thead>
<tr>
<th>Linkage mode</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ward.D</td>
<td>0.515</td>
</tr>
<tr>
<td>Ward.D2</td>
<td>0.623</td>
</tr>
<tr>
<td>Single</td>
<td>0.806</td>
</tr>
<tr>
<td>Complete</td>
<td>0.523</td>
</tr>
<tr>
<td>Average</td>
<td>0.813</td>
</tr>
<tr>
<td>McQuitty</td>
<td>0.694</td>
</tr>
<tr>
<td>Median</td>
<td>0.750</td>
</tr>
<tr>
<td>Centroid</td>
<td>0.797</td>
</tr>
</tbody>
</table>
Figure 8. Results of hierarchical cluster analysis on the overall data set. (A) Dendrogram; (B) Principal component analysis map.
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 $P_{\text{water}}$, $P_{\text{uncor}}$, and $P_{\text{fisher}}$. 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/m² as well as NOTCH3 heterozygous mutation, MTHFR-C677T, and FV-Prothrombin than in the other groups (RR=1.23, 95% CI 0.99-1.54; $P_{\text{fisher}}=2.68\times10^{-3}$). This risk was 20% lower in the group of patients with BMI less than 20.8 kg/m² and with MTHFR-A1298C and FV-H1299R wildtype variants than in the other groups (RR=0.79, 95% CI 0.61-1.01; $P_{\text{fisher}}=1.72\times10^{-3}$).

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 $\leq 13.25$ seconds, PT ratio $\leq 99$, 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; $P_{\text{fisher}}=2.19\times10^{-2}$ and $8.81\times10^{-2}$, 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 $\leq 13.25$ seconds, PT ratio $\leq 99$, 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; $P_{\text{fisher}}=3.47\times10^{-2}$).

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 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=2.64×10^{-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 determine the optimal thresholds of several continuous factors (creatine, 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/m² are considered to have a healthy weight. Our calculated baseline BMI was 20.85 kg/m², 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**

Data found that 3 of the 9 gene variants had significant RR.s. 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


15. Interactive Graph 3 (created by Hocamtu). Plotly Chart Studio. URL: https://chart-studio.plotly.com/~hocamtu/13/#/ [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]


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]


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

https://bioinform.jmir.org/2024/1/e56884

JMIR Bioinform Biotech 2024 | vol. 5 | p.48

(page number not for citation purposes)
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
Assessing Privacy Vulnerabilities in Genetic Data Sets: Scoping Review

Mara Thomas1, PhD; Nuria Mackes2, PhD; Asad Preuss-Dodhy3, PhD; Thomas Wieland4, PhD; Markus Bundschus3, PhD
1F. Hoffmann-La Roche AG, Basel, Switzerland
2xValue GmbH, Ratingen, Germany
3Roche Diagnostics GmbH, Penzberg, Germany
4Foundation Medicine GmbH, Penzberg, Germany

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.

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 (e.g., 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 (i.e., 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 (i.e., 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.
**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 (e.g., 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.

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. Due to the limited knowledge, they currently have low potential to be exploited for privacy attacks, but this is likely to change in the future.

No privacy attack has been demonstrated on somatic variation yet.

SNPs

- 10-30 SNPs can suffer for individual identification.
- SNPs allow for inference of phenotype and disease.
- SNPs are widely present in public and private databases, sometimes with additional identifying information.
- Multiple privacy attacks have been demonstrated on SNPs, including identity tracing.

STRs

- 10-15 STRs can suffer for individual identification.
- STRs allow for limited inference of phenotype and disease.
- STRs, especially Y-STRs and CODIS STRs are present in public and private databases, sometimes with additional identifying information.
- Multiple privacy attacks have been demonstrated on STRs, including identity tracing.

Aggregated sample measures

- Aggregated sample measures allow for membership attacks, which can allow for inference of additional data.
- Membership attacks on aggregated sample measures require prior knowledge of the data subjects' genetic data.
- Multiple membership attacks have been demonstrated on aggregated sample measures data, but no identity tracing.

Rare SNVs

- SNVs greatly increase the risk of reidentification for variant carriers.
- Germ line SNVs are rarely studied and not widely present in public and private databases.
- No privacy attack has been demonstrated on SNVs yet.

Structural variants

- Structural variants are not well studied but could potentially be used for individual identification.
- Structural variants are not widely present in public and private databases.
- Due to the limited knowledge, they currently have low potential to be exploited for privacy attacks, but this is likely to change in the future.
- No privacy attack has been demonstrated on structural variants yet.
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, protein levels, or other modalities (eg, age [72], cigarette smoking, and alcohol consumption [59] from DNA methylation).

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 services, or 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_0000023.9: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 or 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 (e.g., diseases and physical traits)?
- Could additional DNA sequence information be inferred from the data (e.g., 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**


https://bioinform.jmir.org/2024/1/e54332 JMIR Bioinform Biotechnol 2024 | vol. 5 | e54332 | p.61 (page number not for citation purposes)


84. Regalado A. More than 26 million people have taken an at-home ancestry test. MIT Technology Review. 2019 Feb 11. [URL: https://www.sciencemag.org/content/293/5536/1299]


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

©Mara Thomas, Nuria Mackes, Asad Preuss-Dodhy, Thomas Wieland, Markus Bundschus. Originally published in JMIR Bioinformatics and Biotechnology (https://bioinform.jmir.org), 27.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Bioinformatics and
Biotechnology, is properly cited. The complete bibliographic information, a link to the original publication on https://bioinform.jmir.org/, as well as this copyright and license information must be included.