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. 3

Corrigenda and Addenda

Correction: Mutations of SARS-CoV-2 Structural Proteins in the Alpha, Beta, Gamma, and Delta Variants: Bioinformatics Analysis (e64915)
Saima Khetran, Roma Mustafa. 11

Editorial

ChatGPT and Medicine: Together We Embrace the AI Renaissance (e52700)
Sean Hacking. 13

Original Papers

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 Nguy n Th Ph ng, Ho Cam Tu, Sinh Nguyen Phuong, Thuy Pham, Thu Vu, Huyen Nguyen Thi Thu, Lam Khanh Ho, Dung Nguyen Tien. 18

Deep Learning–Based Identification of Tissue of Origin for Carcinomas of Unknown Primary Using MicroRNA Expression: Algorithm Development and Validation (e56538)
Ananya Raghu, Anisha Raghu, Jillian Wise. 60

XSL•FO
RenderX

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. 72

Review

Assessing Privacy Vulnerabilities in Genetic Data Sets: Scoping Review (e54332)

Mara Thomas, Nuria Mackes, Asad Preuss-Dodhy, Thomas Wieland, Markus Bundschus. 42

Viewpoint

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

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

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

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

Abstract

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

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

KEYWORDS

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

XSL•FO
RenderX

Introduction

Digital Health Care Systems Are Transforming Health Care

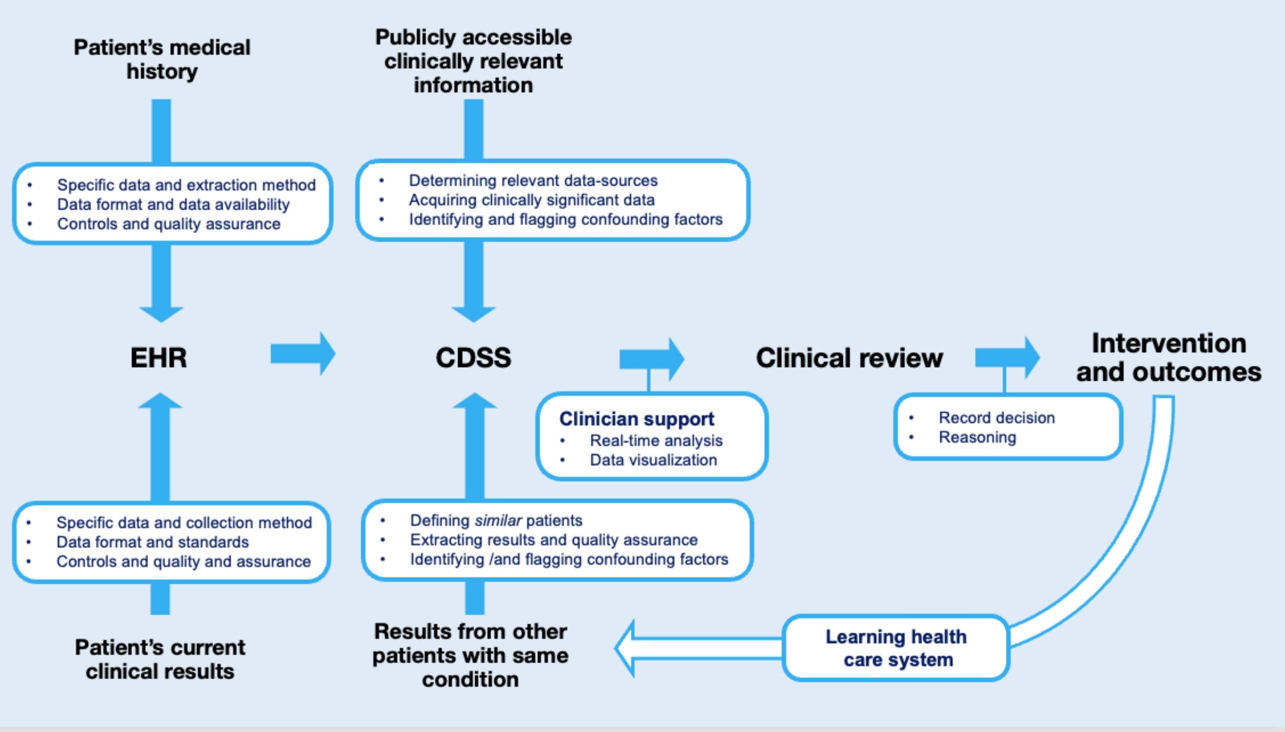
The adoption of electronic health records (EHRs) is transforming health care [1-4]. This digital infrastructure allows health services to store a patient's complete medical history and collect additional observations and results in real time. Having this information in a standardized, readily accessible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

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

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

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

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


Patient's medical history
Publicly accessible clinically relevant information
• Specific data and extraction method
• Data format and data availability
• Controls and quality assurance
• Determining relevant data-sources
• Acquiring clinically significant data
• Identifying and flagging confounding factors
EHR
CDSS
Clinical review
Intervention and outcomes
Clinician support
• Real-time analysis
• Data visualization
• Record decision
• Reasoning
• Specific data and collection method
• Data format and standards
• Controls and quality and assurance
• Defining *similar* patients
• Extracting results and quality assurance
• Identifying /and flagging confounding factors
Patient's current clinical results
Results from other patients with same condition
Learning health care system

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

XSL•FO
RenderX

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

Genomic Data Are an Important Element of Precision Medicine

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

Table 1. Clinical applications of genomics.

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

XSL•FO
RenderX

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

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

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

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

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

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

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

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

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

Textbox 1. A clinical vignette.

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

Moving From Prescriptive to Precision Medicine

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

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

XSL•FO
RenderX

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

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

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

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

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

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

[a] CDSS: clinical decision support system.

[b] EHR: electronic health record.

[c] LIM: Laboratory Information Management System.

XSL•FO RenderX

Conclusions

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

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

Authors' Contributions

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

Conflicts of Interest

AJR is the founder and director of ClearSKY Genomics.

References

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

XSL•FO
RenderX

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

XSL•FO
RenderX

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

Abbreviations

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

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

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

©Alan J Robertson, Andrew J Mallett, Zornitza Stark, Clair Sullivan. Originally published in JMIR Bioinformatics and Biotechnology (https://bioinform.jmir.org), 13.06.2024. 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.

XSL•FO
RenderX

Corrigenda and Addenda

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

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

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

[*] all authors contributed equally

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

Related Article:

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

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

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

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

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

This has been changed as follows:

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

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

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

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

XSL•FO
RenderX

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

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

©Saima Rehman Khetran, Roma Mustafa. Originally published in JMIR Bioinformatics and Biotechnology (https://bioinform.jmir.org), 05.08.2024. 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.

XSL•FO
RenderX

Editorial

ChatGPT and Medicine: Together We Embrace the AI Renaissance

Sean Hacking[1], MBBCh

NYU Langone, New York, NY, United States

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

Abstract

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

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

KEYWORDS

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

Introduction

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

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

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

Viewpoint of the Physician

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

XSL•FO
RenderX

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

XSL•FO
RenderX

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

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

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

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

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

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

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

Conclusion

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

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

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

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

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

XSL•FO
RenderX

Acknowledgments

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

Conflicts of Interest

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

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

References

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

XSL•FO
RenderX

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

Abbreviations

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

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

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

©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.

XSL•FO
RenderX

Original Paper

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

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

[1]Department of Biochemistry, Thai Nguyen University of Medicine and Pharmacy, Thai Nguyen, Vietnam
[2]Department of Immunology Molecular Genetic, Thainguyen National Hospital, Thai Nguyen, Vietnam
[3]Department of Clinical Pharmacy, Thai Nguyen University of Medicine and Pharmacy, Thai Nguyen, Vietnam
[4]Center of Gene and Protein Research, Hanoi Medical University, Hanoi, Vietnam
[5]Institute of Virology, School of Medicine, Technical University of Munich, Munich, Germany
[6]Department of Rehabilitation, Thai Nguyen University of Medicine and Pharmacy, Thai Nguyen, Vietnam
[7]Department of Internal Medicine, Thai Nguyen University of Medicine and Pharmacy, Thai Nguyen, Vietnam
[8]Department of Telecomunication, Hung Yen University of Technology and Education, Hung Yen, Vietnam
[*]these authors contributed equally

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

Abstract

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

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

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

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

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

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

KEYWORDS

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

Introduction

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

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

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

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

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

Methods

Study Design

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

XSL•FO
RenderX

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

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

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

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

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

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

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

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

Genetic Testing

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

Ethical Considerations

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

Statistical Analysis and HCA

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

XSL•FO
RenderX

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

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

Results

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

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

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

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

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

XSL•FO
RenderX

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

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

XSL•FO
RenderX

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

Table 3. Symptoms at admission.

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

XSL•FO
RenderX

Table 4. Important variables in this study.

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

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

XSL•FO
RenderX

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

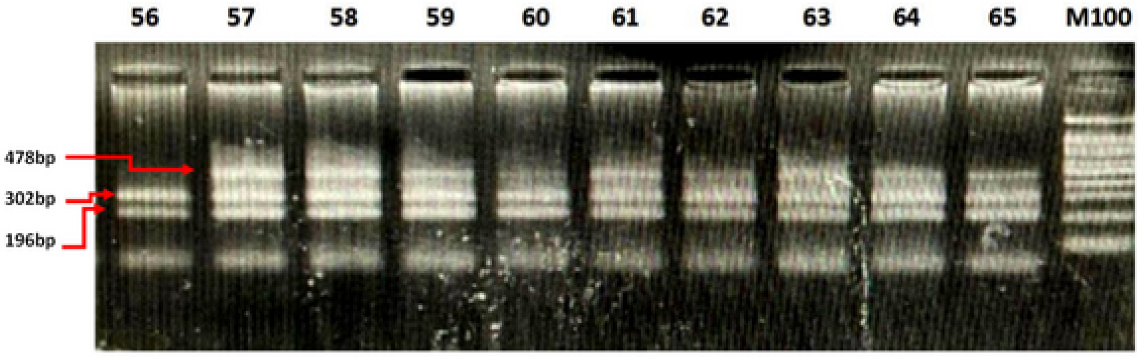


Figure 2. Identification of the FV-Leiden variant by real-time polymerase chain reaction. (A) Wildtype; (B) Heterozygous. RFU: relative fluorescence units.

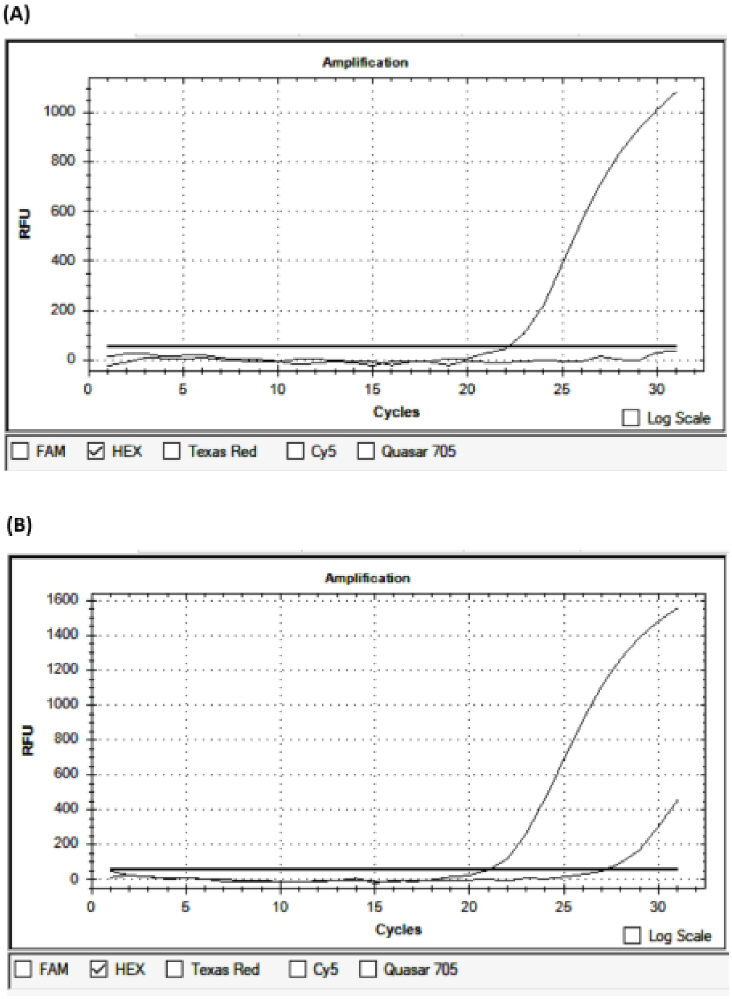
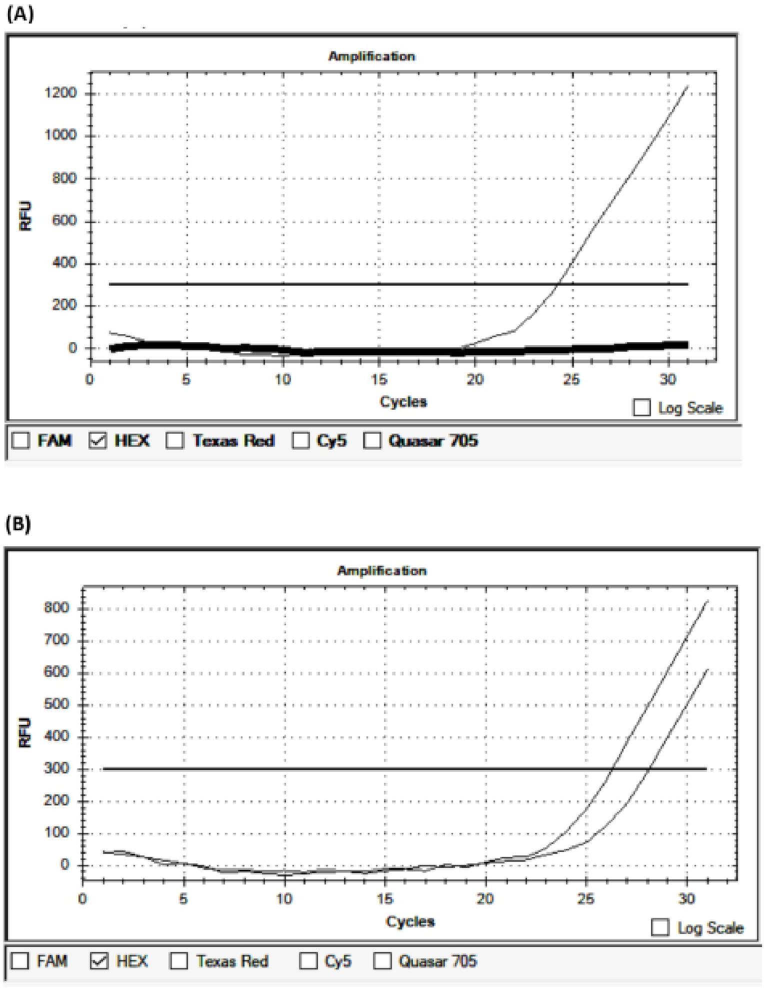


Figure 3. Identification of the *FV-H1299R* variant by real-time polymerase chain reaction. (A) Wildtype; (B) Heterozygous. RFU: relative fluorescence units.



(A)

☐ Log Scale

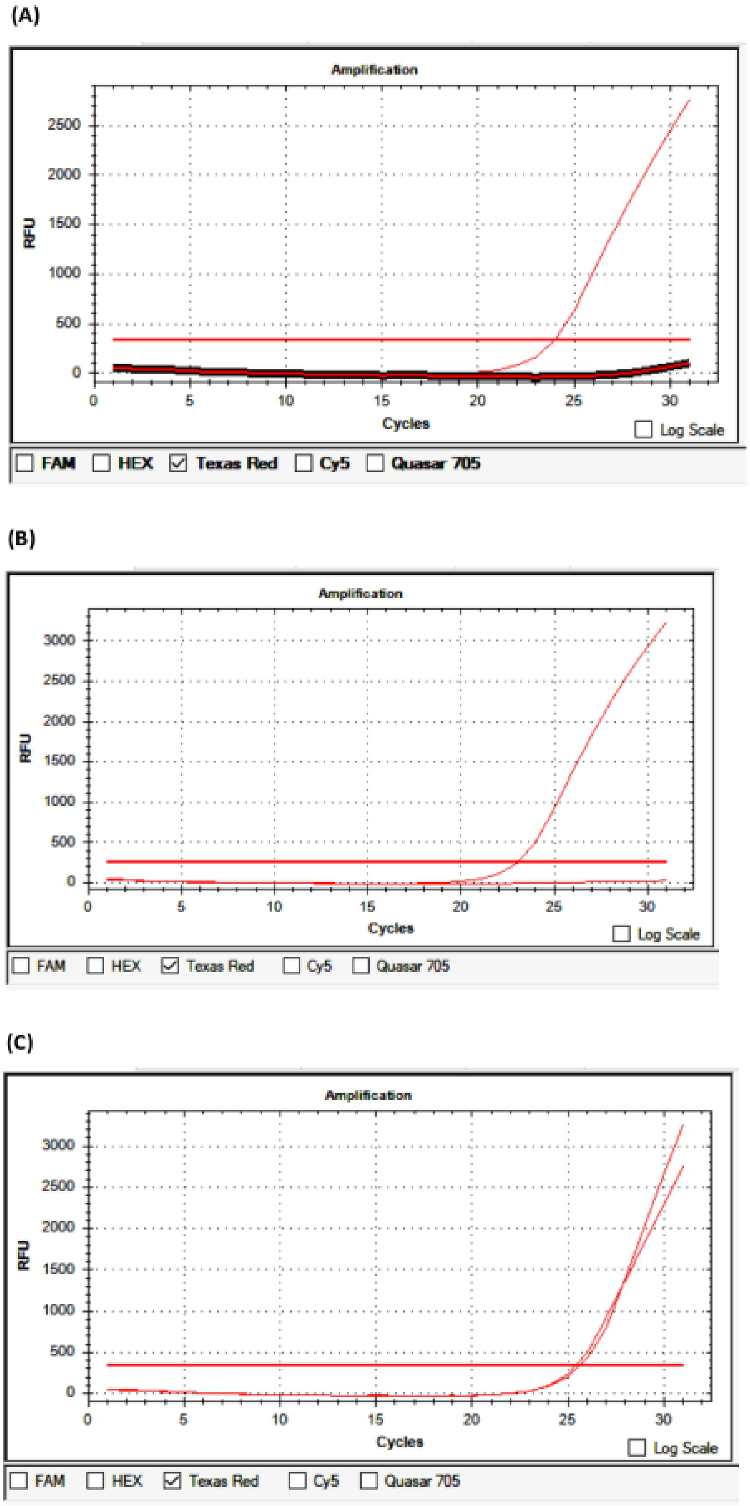
☐ FAM ☑ HEX ☐ Texas Red ☐ Cy5 ☐ Quasar 705

(B)

☐ Log Scale

☐ FAM ☑ HEX ☐ Texas Red ☐ Cy5 ☐ Quasar 705

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



XSL•FO
RenderX

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

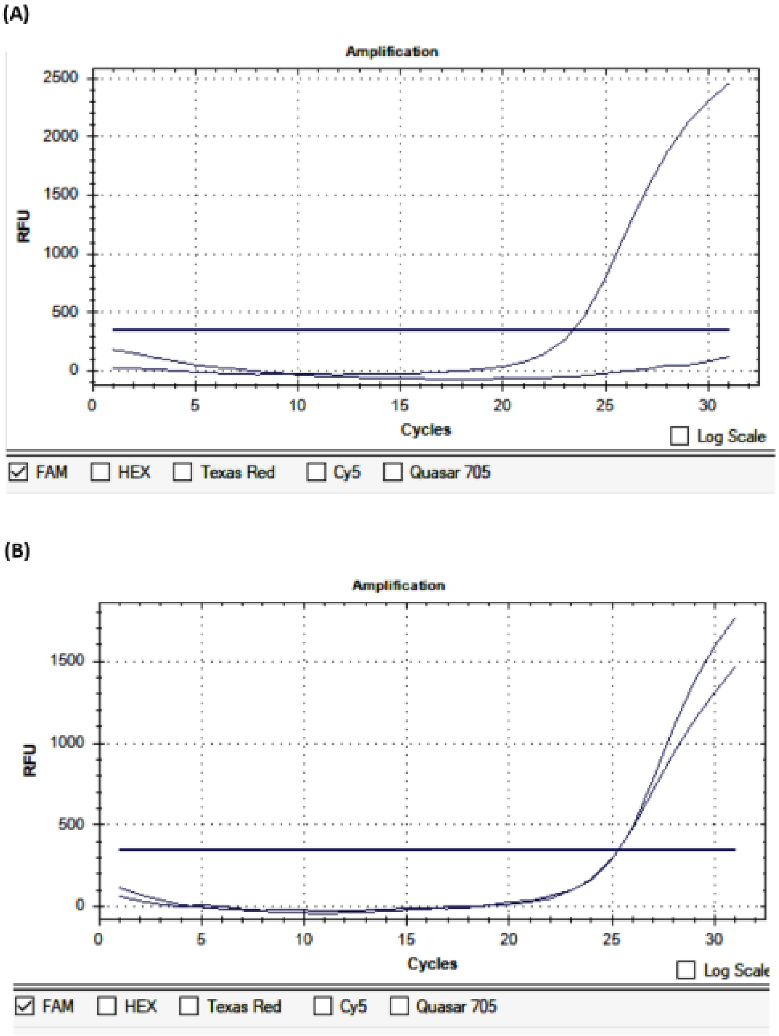


Figure 6, Table 1, and Table S1 in Multimedia Appendix 1 provide an overall view of gene prevalence and correlations in both negative and positive genes. We confirmed the presence of significant correlations of *NOTCH3* p.R544C, *FV-H1299R*, *MTHFR-C677T*, *MTHFR-A1298C*, *FII Prothrombin*, *FV-Cambridge*, *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).

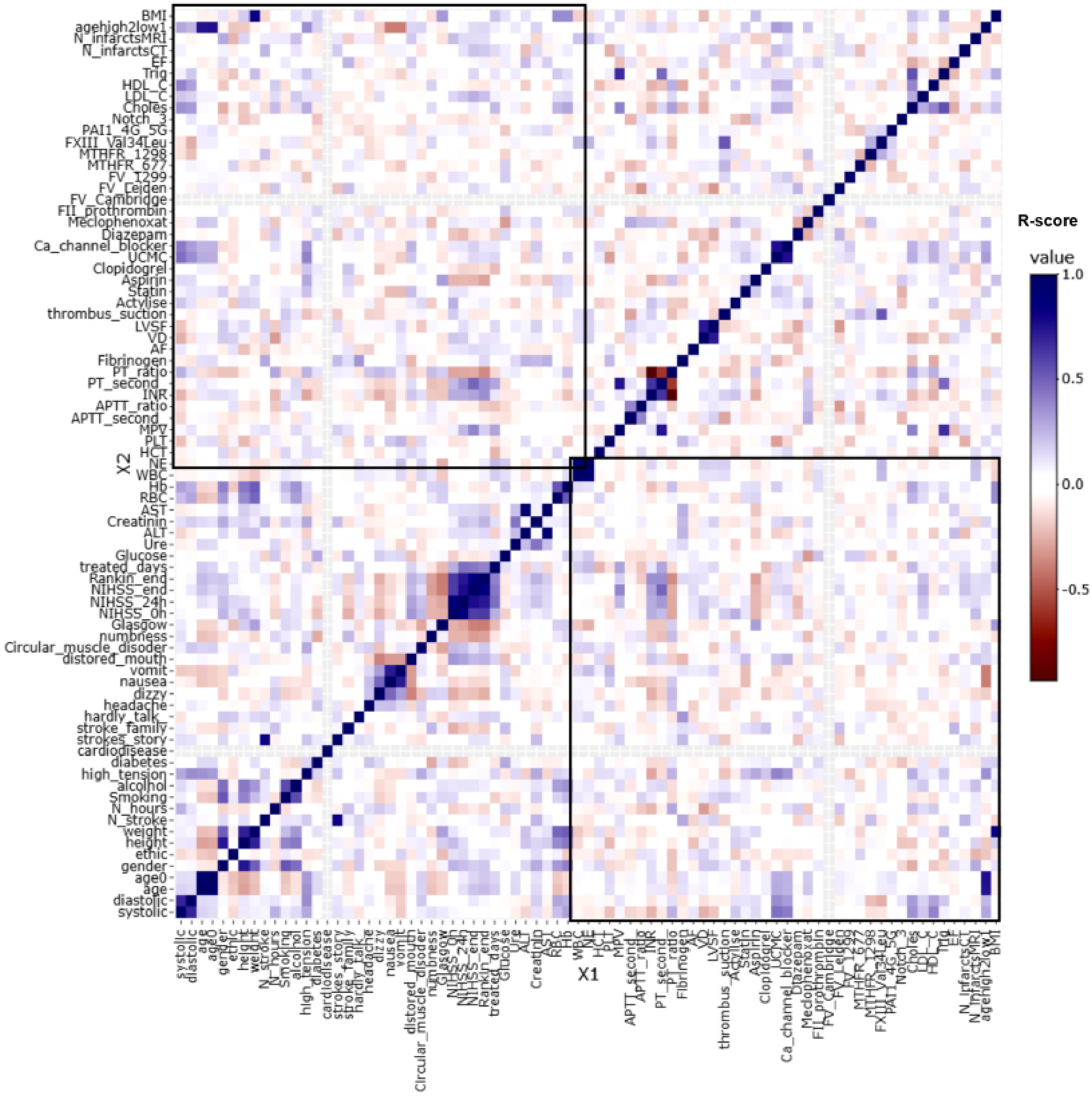
In the clustering step, dendrograms were built based on the clustering metric “Euclidean,” and we selected “average” as the most appropriate linkage model, which had the best correlation between cophenetic distances and the original distance data (Table 5).

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

XSL•FO
RenderX

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

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



XSL•FO
RenderX

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

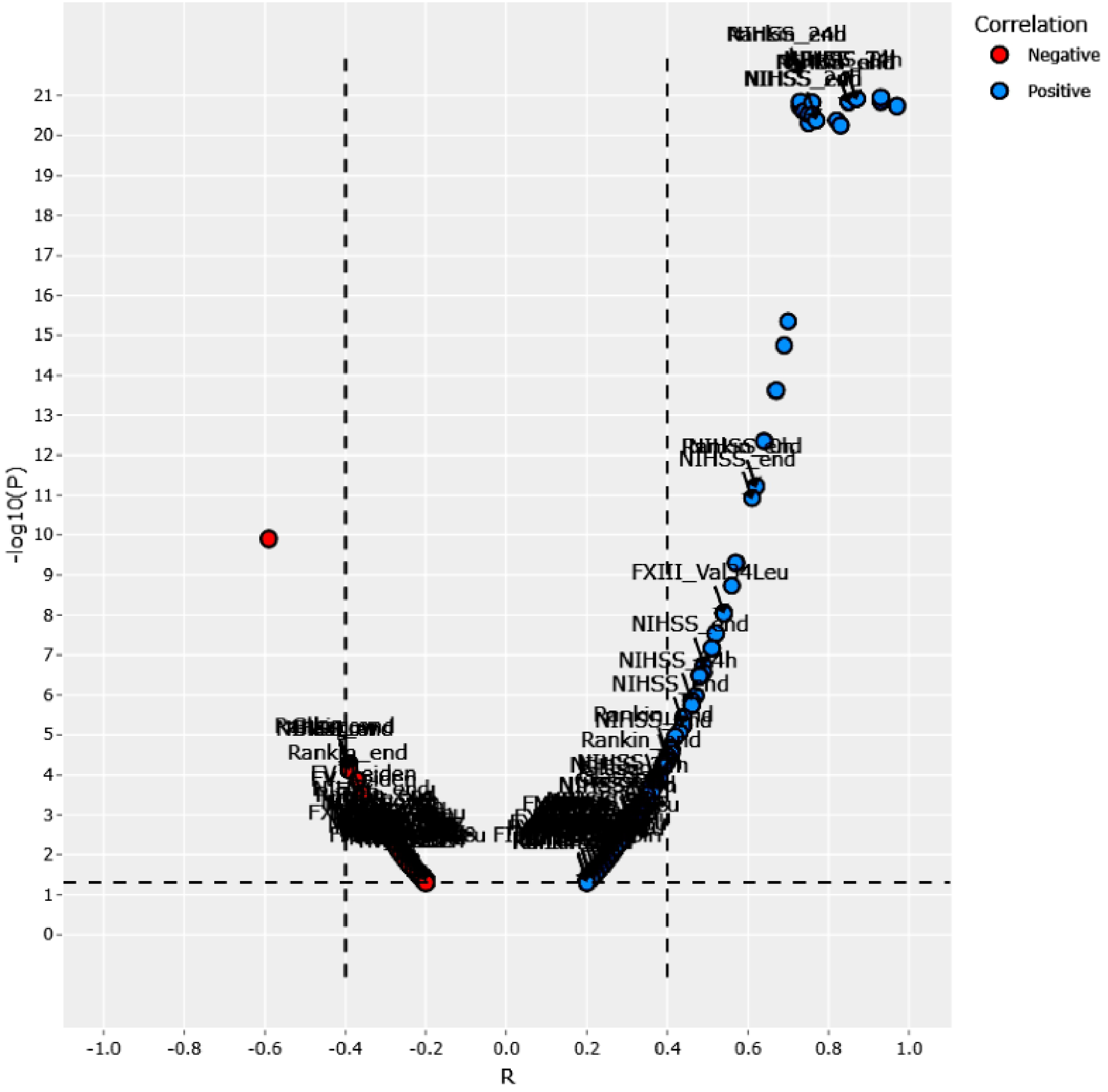
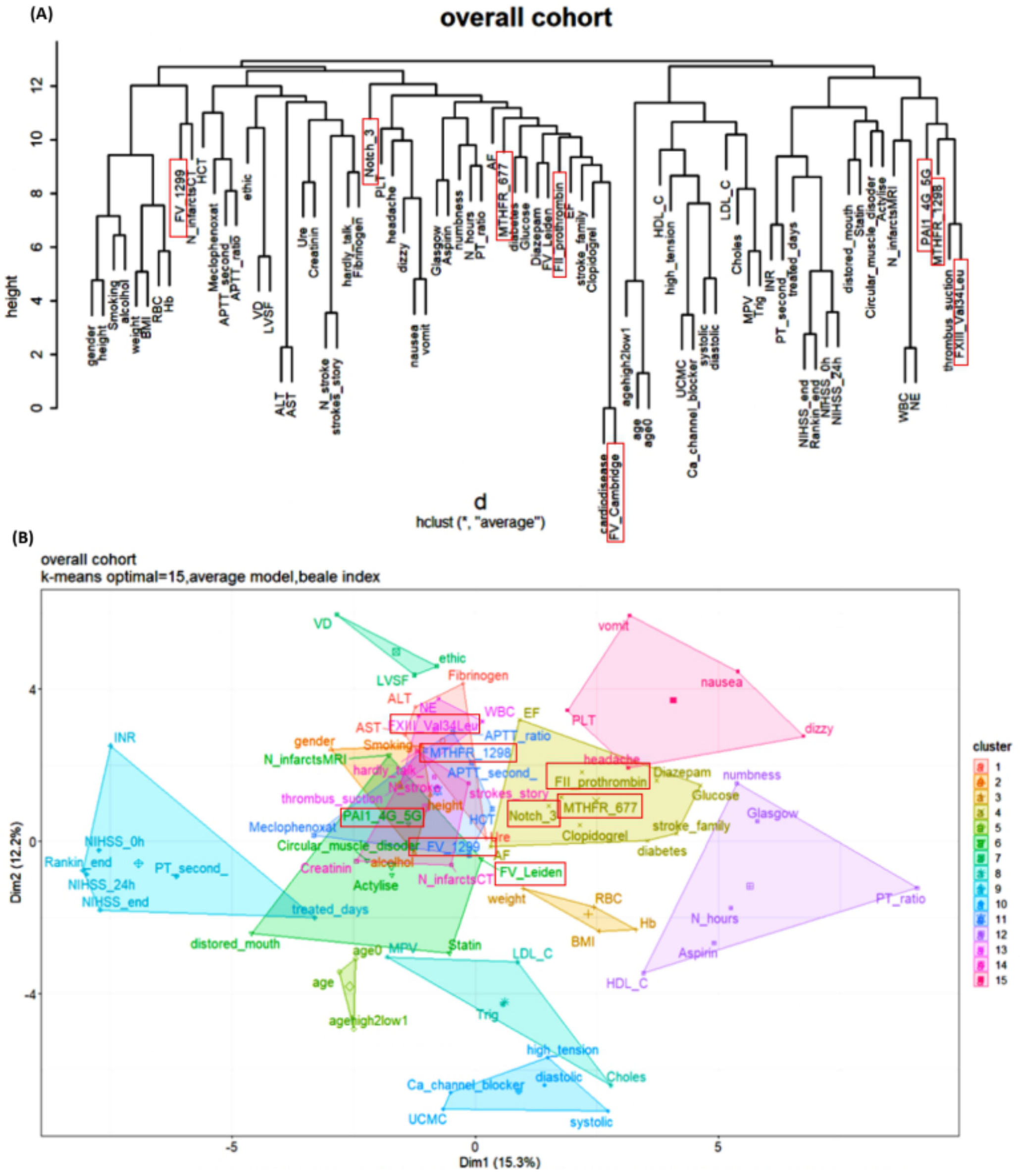


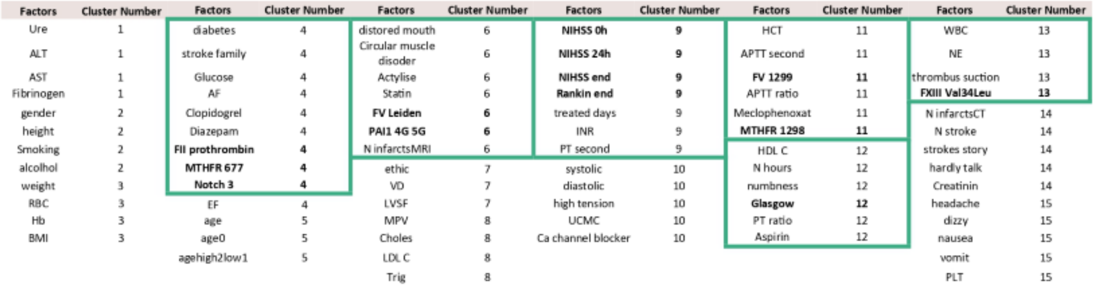
Table 5. Correlation between cophenetic distances and the original distance data.

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

XSL•FO
RenderX

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





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

XSL•FO
RenderX

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.

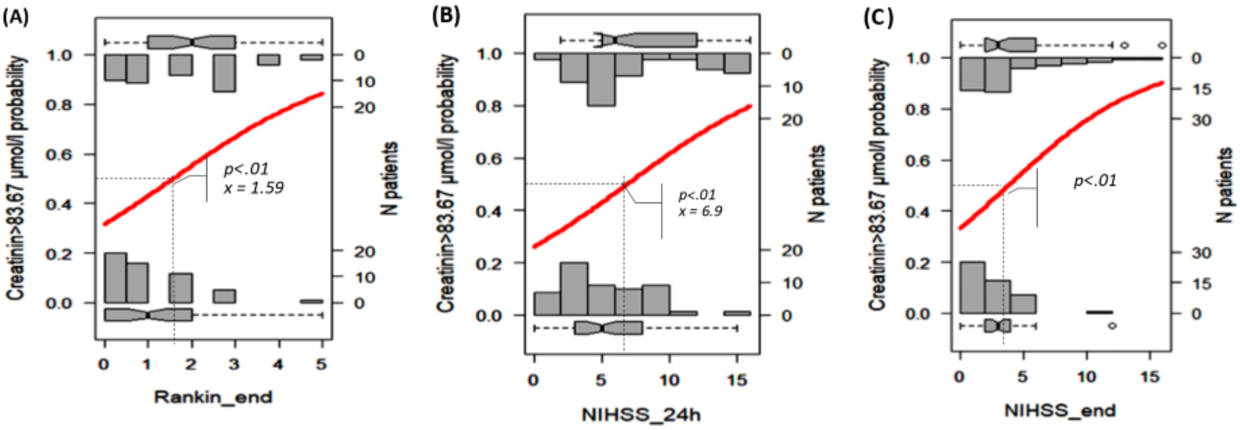
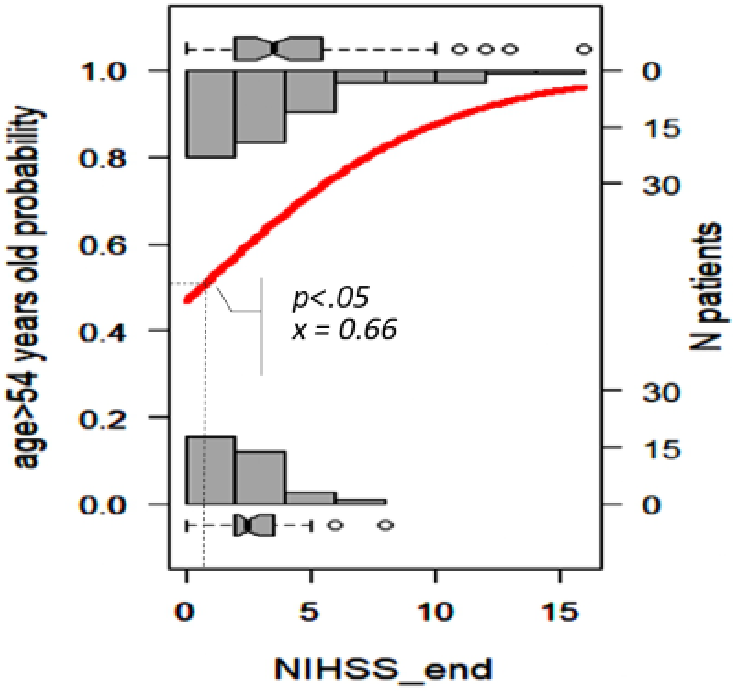


Figure 10. Significant cutpoint 50% of the National Institutes of Health Stroke Scale (NIHSS) score at discharge for patient age >54 years.



XSL•FO
RenderX

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.

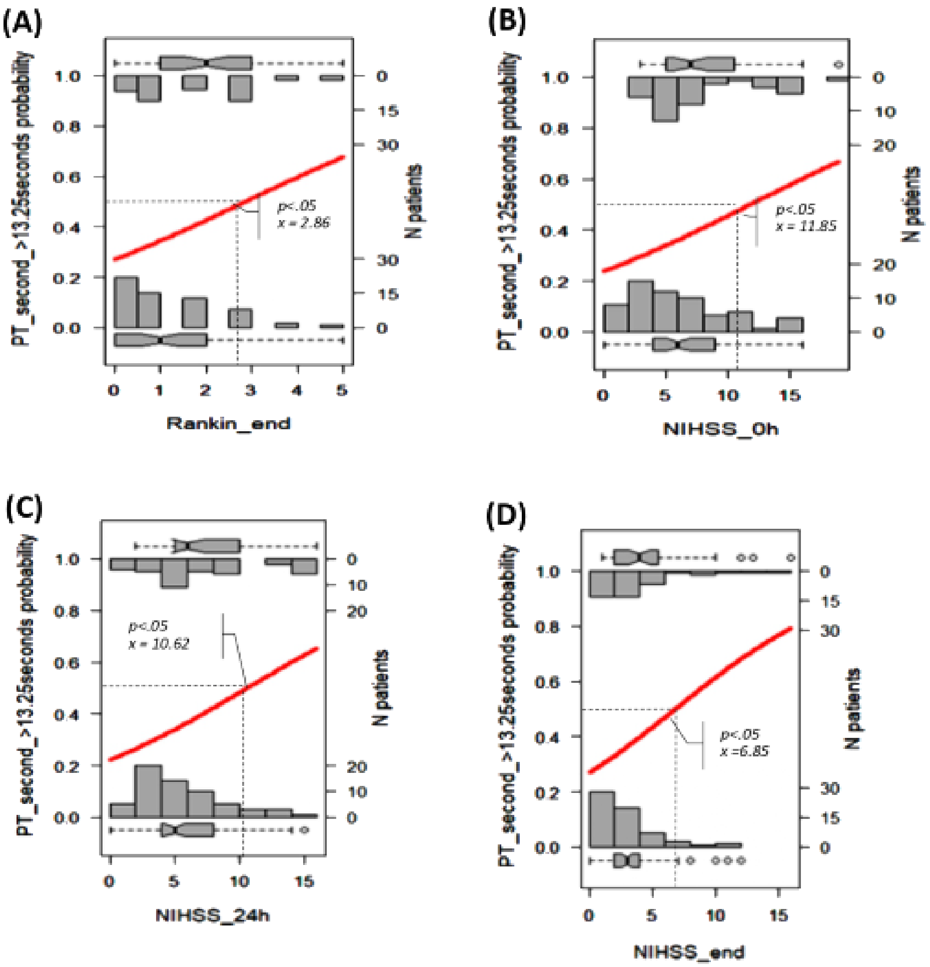


Figure 12. Significant cutpoint 50% of the modified Rankin scale score at discharge (A) and the National Institutes of Health Stroke Scale (NIHSS) scores at admission (B), after 24 hours (C), and at discharge (D) for prothrombin (PT) ratio >99.

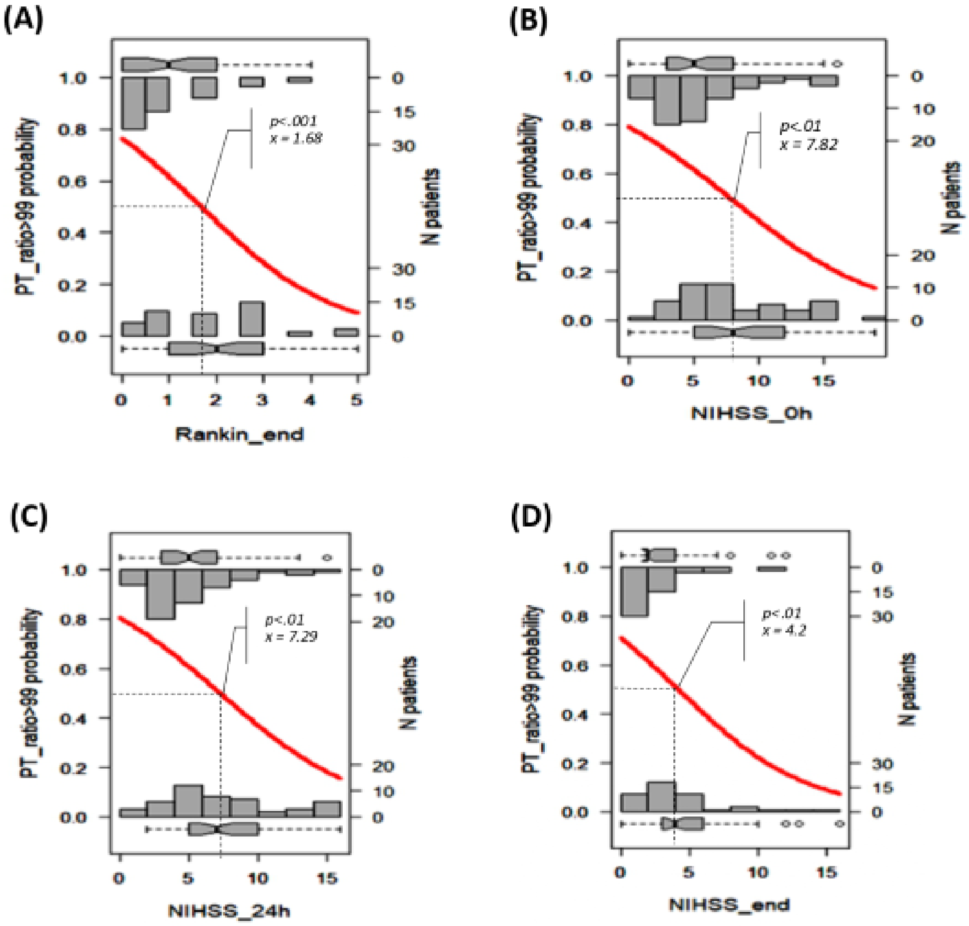
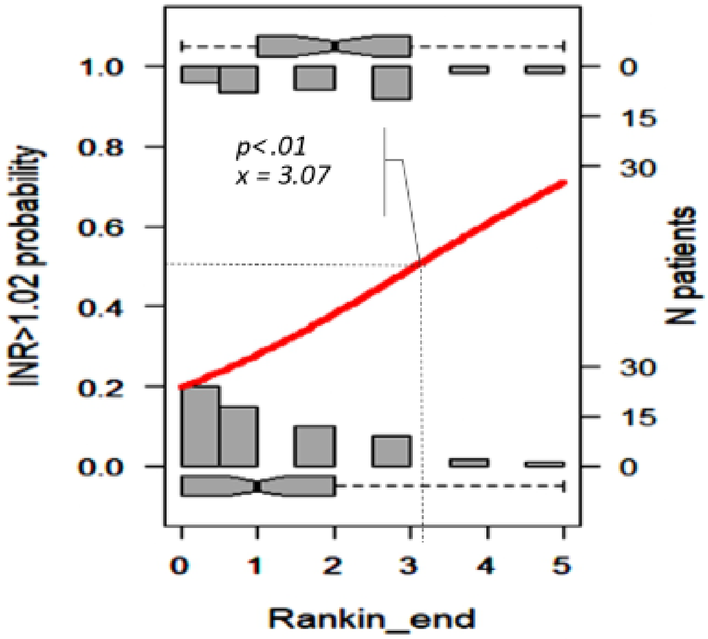


Figure 13. Significant cutpoint 50% of the modified Rankin scale score at discharge for international normalized ratio (INR) >1.02.



XSL•FO
RenderX

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

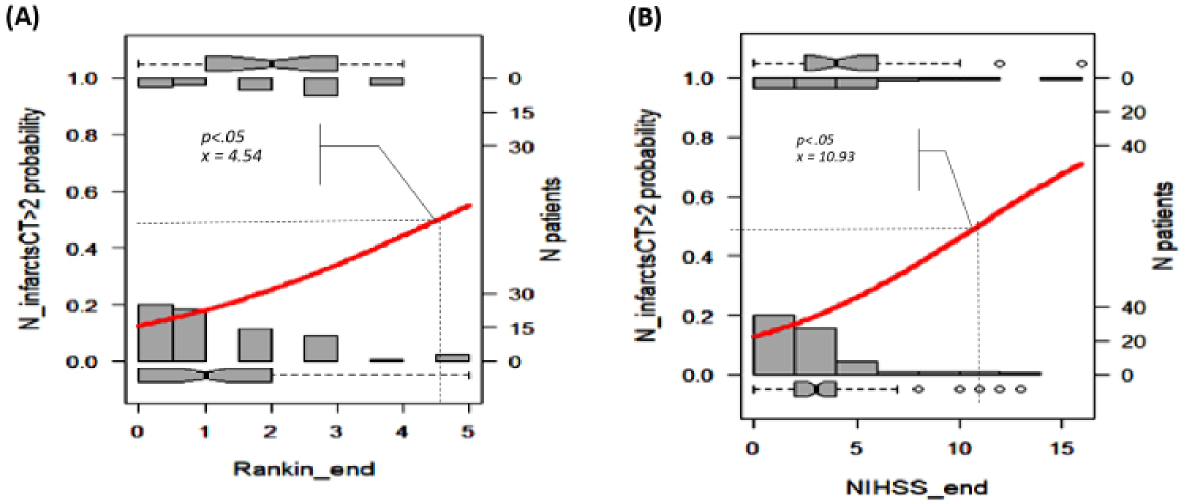


Figure 15. Significant cutpoint 50% of the National Institutes of Health Stroke Scale (NIHSS) score at admission for patient height >161 cm.

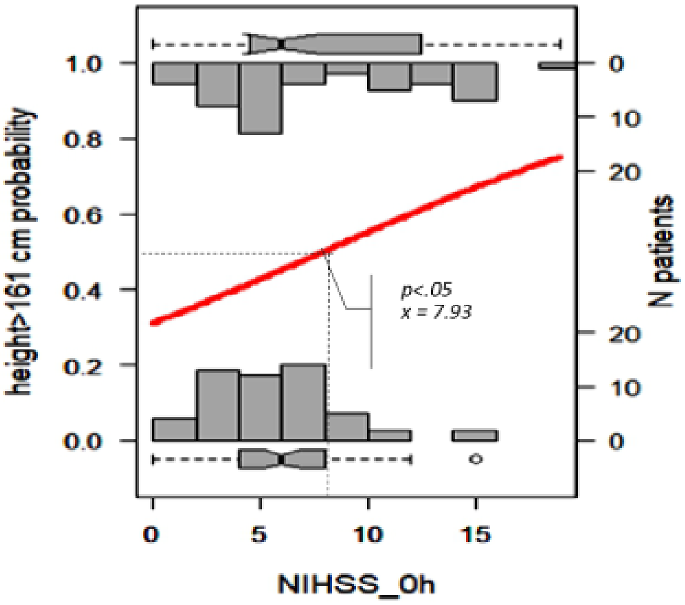
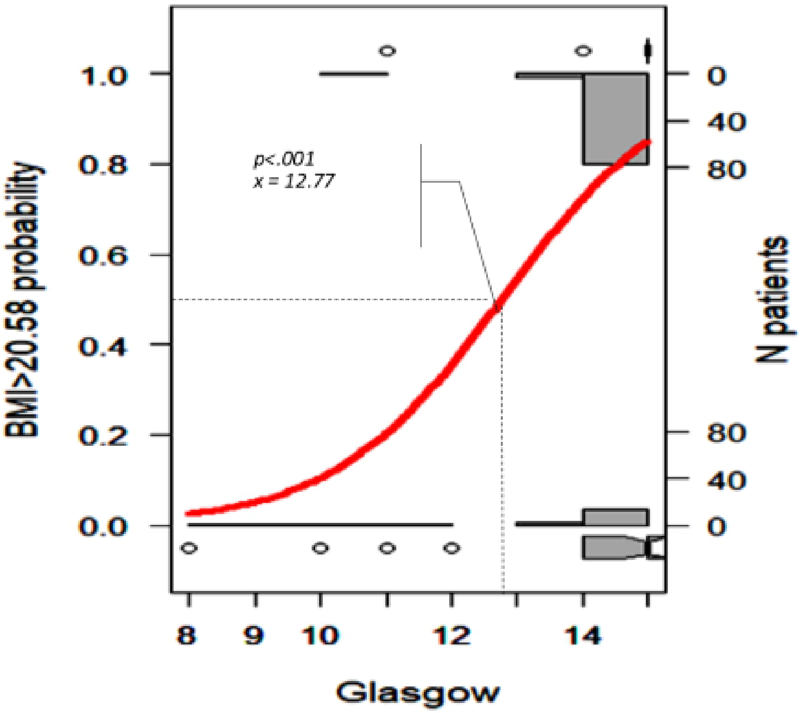


Figure 16. Significant cutpoint 50% of the Glasgow Coma Scale score for patient BMI >20.58 kg/m2.



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

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

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

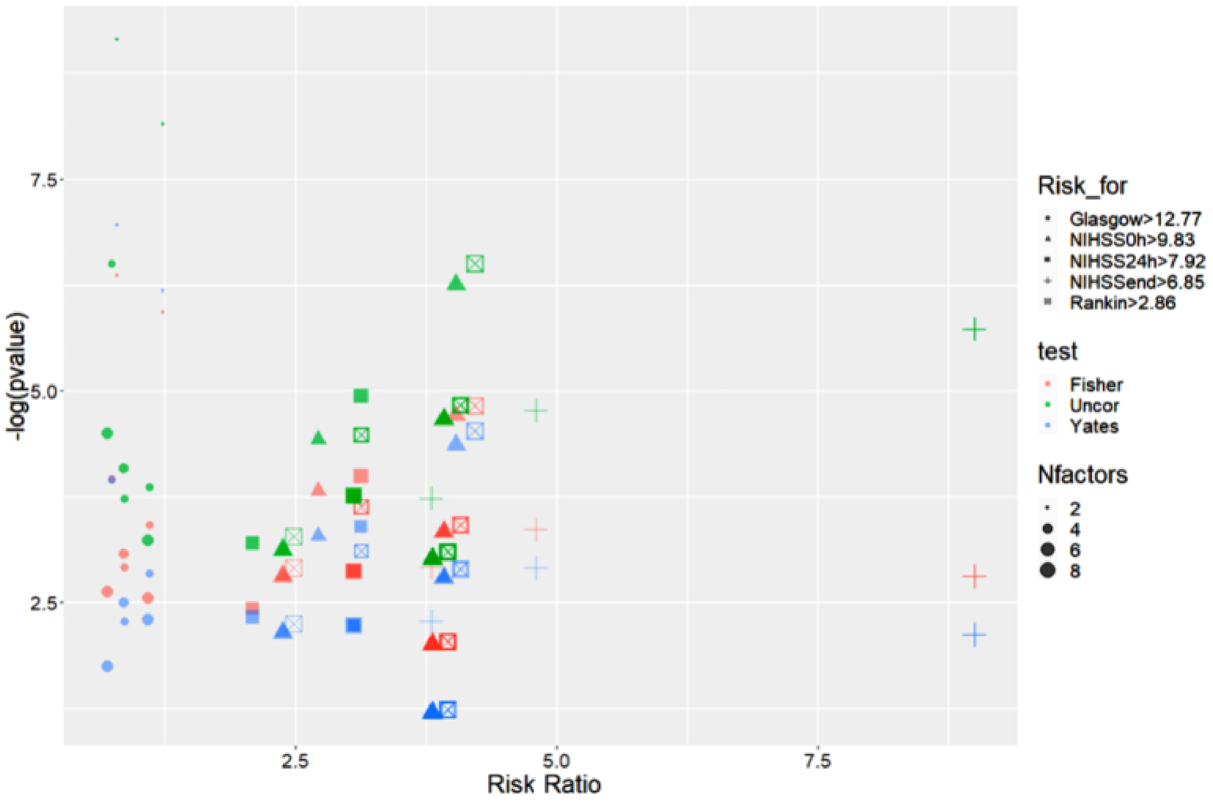
The NIHSS quantifies the impairment caused by stroke and aids in planning post-acute care disposition, although it has been intended to assess differences in interventions in clinical trials. A NIHSS score of 0 indicates no stroke symptoms, 1-4 indicates minor stroke, 5-15 indicates moderate stroke,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

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

XSL•FO
RenderX

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

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



Discussion

Principal Findings

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

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

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

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

XSL•FO
RenderX

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

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

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

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

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

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

Limitations

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

Conclusions

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

XSL•FO
RenderX

Acknowledgments

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

Data Availability

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

Authors' Contributions

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

Conflicts of Interest

None declared.

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

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

References

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

XSL•FO
RenderX

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

Abbreviations

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

XSL•FO
RenderX

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

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

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

©Huong Thi Thu Bui, Quỳnh Nguy n Th Ph ng, Ho Cam Tu, Sinh Nguyen Phuong, Thuy Thi Pham, Thu Vu, Huyen Nguyen Thi Thu, Lam Khanh Ho, Dung Nguyen Tien. 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.

XSL•FO
RenderX

Review

Assessing Privacy Vulnerabilities in Genetic Data Sets: Scoping Review

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

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

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

Abstract

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

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

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

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

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

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

KEYWORDS

genetic privacy; privacy; data anonymization; reidentification

Introduction

Privacy Risks of Genetic Data

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

XSL•FO
RenderX

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

XSL•FO
RenderX

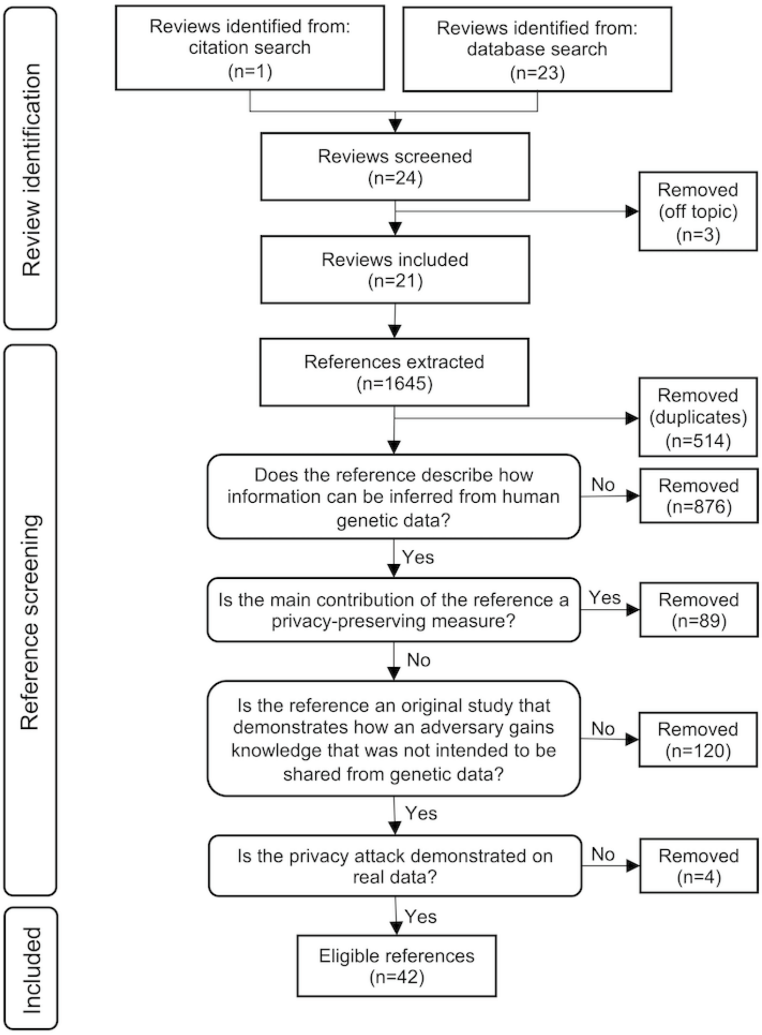
of genetic data find little guidance on performing an assessment on genetic data that considers these factors. While several genomic privacy metrics have been proposed, the great majority focus on evaluating SNPs only [34] and neglect other known privacy-critical aspects of genetic data as well as aspects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

Methods

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

XSL•FO
RenderX

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


Review identification
Reviews identified from: citation search (n=1)
Reviews identified from: database search (n=23)
Reviews screened (n=24)
Removed (off topic) (n=3)
Reviews included (n=21)
Reference screening
References extracted (n=1645)
Removed (duplicates) (n=514)
Does the reference describe how information can be inferred from human genetic data?
No
Removed (n=876)
Yes
Is the main contribution of the reference a privacy-preserving measure?
Yes
Removed (n=89)
No
Is the reference an original study that demonstrates how an adversary gains knowledge that was not intended to be shared from genetic data?
No
Removed (n=120)
Yes
Is the privacy attack demonstrated on real data?
No
Removed (n=4)
Yes
Included
Eligible references (n=42)

Results

A Comprehensive Overview of Privacy Risks in Genetic Data Sets

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

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

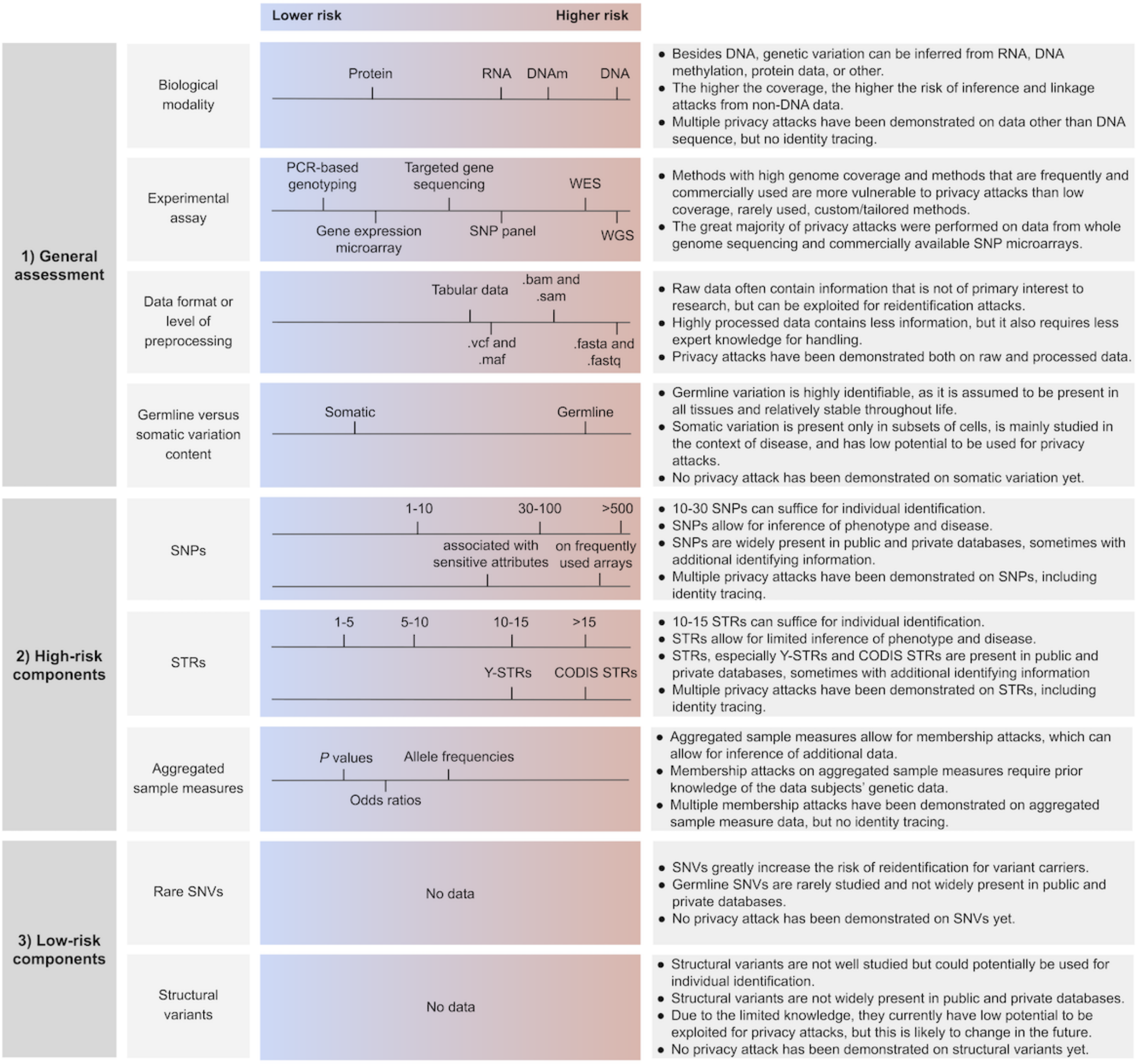
We summarize our findings in an overview figure, which lists the 9 features and their relevance for privacy. While it is challenging to define clear risk thresholds, there is a recognized need for practical guidance and orientation. To address this, we

XSL•FO
RenderX

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

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

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



		Lower risk → Higher risk	
1) General assessment	Biological modality	Protein RNA DNAm DNA	Besides DNA, genetic variation can be inferred from RNA, DNA methylation, protein data, or other. The higher the coverage, the higher the risk of inference and linkage attacks from non-DNA data. Multiple privacy attacks have been demonstrated on data other than DNA sequence, but no identity tracing.
	Experimental assay	PCR-based genotyping Gene expression microarray Targeted gene sequencing SNP panel WES WGS	Methods with high genome coverage and methods that are frequently and commercially used are more vulnerable to privacy attacks than low coverage, rarely used, custom/tailored methods. The great majority of privacy attacks were performed on data from whole genome sequencing and commercially available SNP microarrays.
	Data format or level of preprocessing	Tabular data .vcf and .maf .bam and .sam .fasta and .fastq	Raw data often contain information that is not of primary interest to research, but can be exploited for reidentification attacks. Highly processed data contains less information, but it also requires less expert knowledge for handling. Privacy attacks have been demonstrated both on raw and processed data.
	Germline versus somatic variation content	Somatic Germline	Germline variation is highly identifiable, as it is assumed to be present in all tissues and relatively stable throughout life. Somatic variation is present only in subsets of cells, is mainly studied in the context of disease, and has low potential to be used for privacy attacks. No privacy attack has been demonstrated on somatic variation yet.
2) High-risk components	SNPs	1-10 30-100 >500 associated with sensitive attributes on frequently used arrays	10-30 SNPs can suffice 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	1-5 5-10 10-15 >15 Y-STRs CODIS STRs	10-15 STRs can suffice 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	*P* values Odds ratios Allele frequencies	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 measure data, but no identity tracing.
3) Low-risk components	Rare SNVs	No data	SNVs greatly increase the risk of reidentification for variant carriers. Germline SNVs are rarely studied and not widely present in public and private databases. No privacy attack has been demonstrated on SNVs yet.
	Structural variants	No data	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.

Evidence of Privacy Risks in Genetic Data

Part 1. General Assessment

Biological Modality

While most privacy attacks have been demonstrated on DNA sequence data, other types of molecular data (eg, DNA methylation data or data derived from RNA) are also considered genetic data under General Data Protection Regulation, can also be identifiable, and have also been exploited for attacks [58-67].

Attacks on these types of data are performed mainly by 3 mechanisms. The first mechanism is direct extraction of DNA sequence from raw or low-processed data. This is possible, because even if not of primary interest, DNA sequence information is often a by-product of gene expression or DNA methylation studies [68-70]. For example, Gürsoy et al [70] demonstrated how genetic variants can be called from raw RNA sequencing data. The second mechanism is inference of DNA sequence, for example, through known associations of genetic sequence and gene expression or other modalities. For example,

XSL•FO
RenderX

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

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

The guiding questions in this context are as follows:

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

Experimental Assay

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

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

The guiding questions in this context are as follows:

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

XSL•FO
RenderX

Data Format or Level of Processing

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

The guiding question in this context is as follows:

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

Germline Versus Somatic Variation Content

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

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

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

XSL•FO
RenderX

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

The guiding questions in this context are as follows:

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

Part 2. High-Risk Components

SNPs

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

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

The guiding questions in this context are as follows:

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

XSL•FO
RenderX

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

XSL•FO
RenderX

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

XSL•FO
RenderX

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

The guiding questions in this context are as follows:

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

Discussion

Limitations

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

Conclusions

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

Acknowledgments

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

Conflicts of Interest

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

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

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

References

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

XSL•FO
RenderX

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

XSL•FO
RenderX

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

XSL•FO
RenderX

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

XSL•FO
RenderX

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

XSL•FO
RenderX

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

XSL•FO
RenderX

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

Abbreviations

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

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

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

©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

XSL•FO
RenderX

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.

XSL•FO
RenderX

Original Paper

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

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

[1]Quarry Lane School, San Ramon, CA, United States
[2]Department of Biology and Biomedical Sciences, Salve Regina University, Newport, RI, United States
[3]Broad Institute of MIT and Harvard, Cambridge, MA, United States
[4]Pre-College Programs, Tufts University, Medford, MA, United States
[*]these authors contributed equally

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

Abstract

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

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

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

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

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

XSL•FO
RenderX

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

KEYWORDS

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

Introduction

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

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

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

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

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

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

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

XSL•FO
RenderX

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

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

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

Methods

Data Sets

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

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

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

XSL•FO
RenderX

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

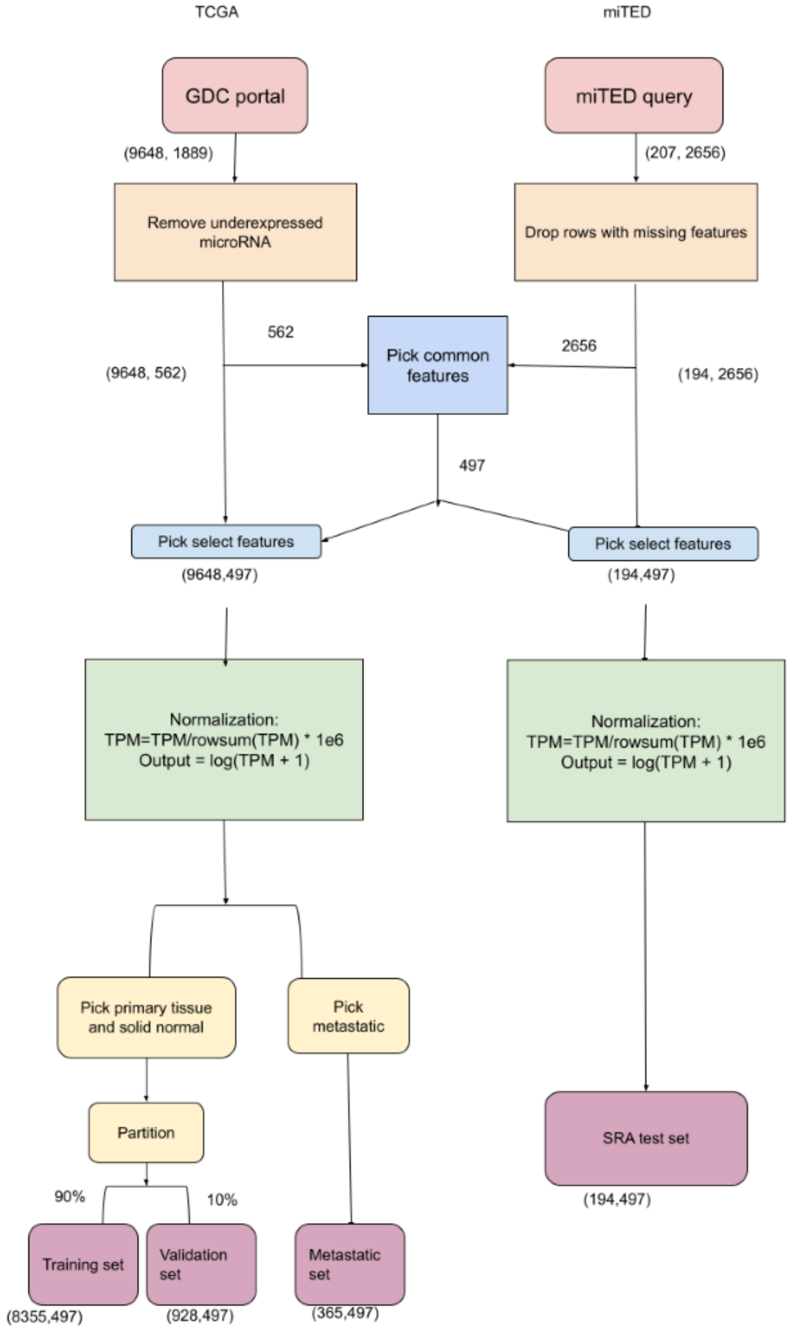
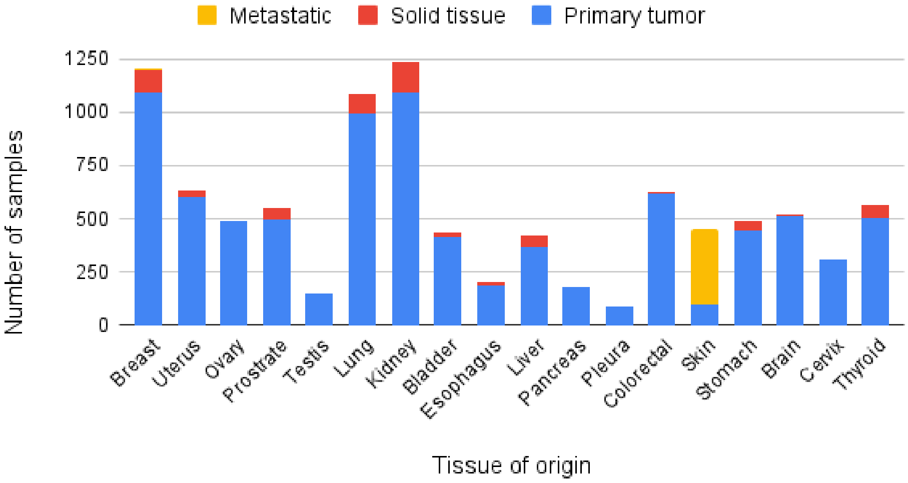


Figure 2. The Cancer Genome Atlas (TCGA) data set distribution across tissue of origin (TOO). Distribution of the different cancer samples in the TCGA data set that are from the primary tumor site, solid tissue, or metastatic. Note that metastatic samples primarily corresponded to the skin as the TOO.



Training Procedure

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

Ethical Considerations

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

Results

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

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

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

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

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

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

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

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

XSL•FO
RenderX

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

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

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

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

[a] SRA: Sequence Read Archive.

[b] N/A: not applicable.

XSL•FO
RenderX

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

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

[a]TCGA: The Cancer Genome Atlas.

[b]N/A: not applicable.

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

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

[a]SRA: Sequence Read Archive.

[b]TCGA: The Cancer Genome Atlas.

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

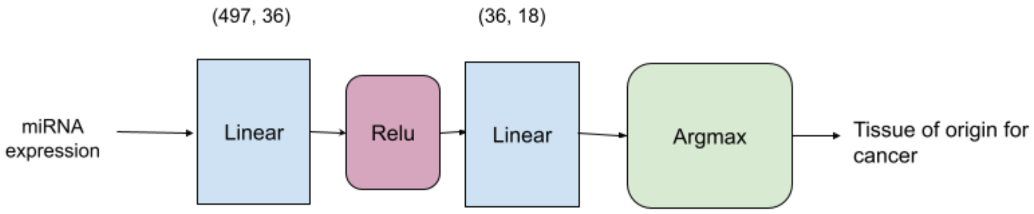
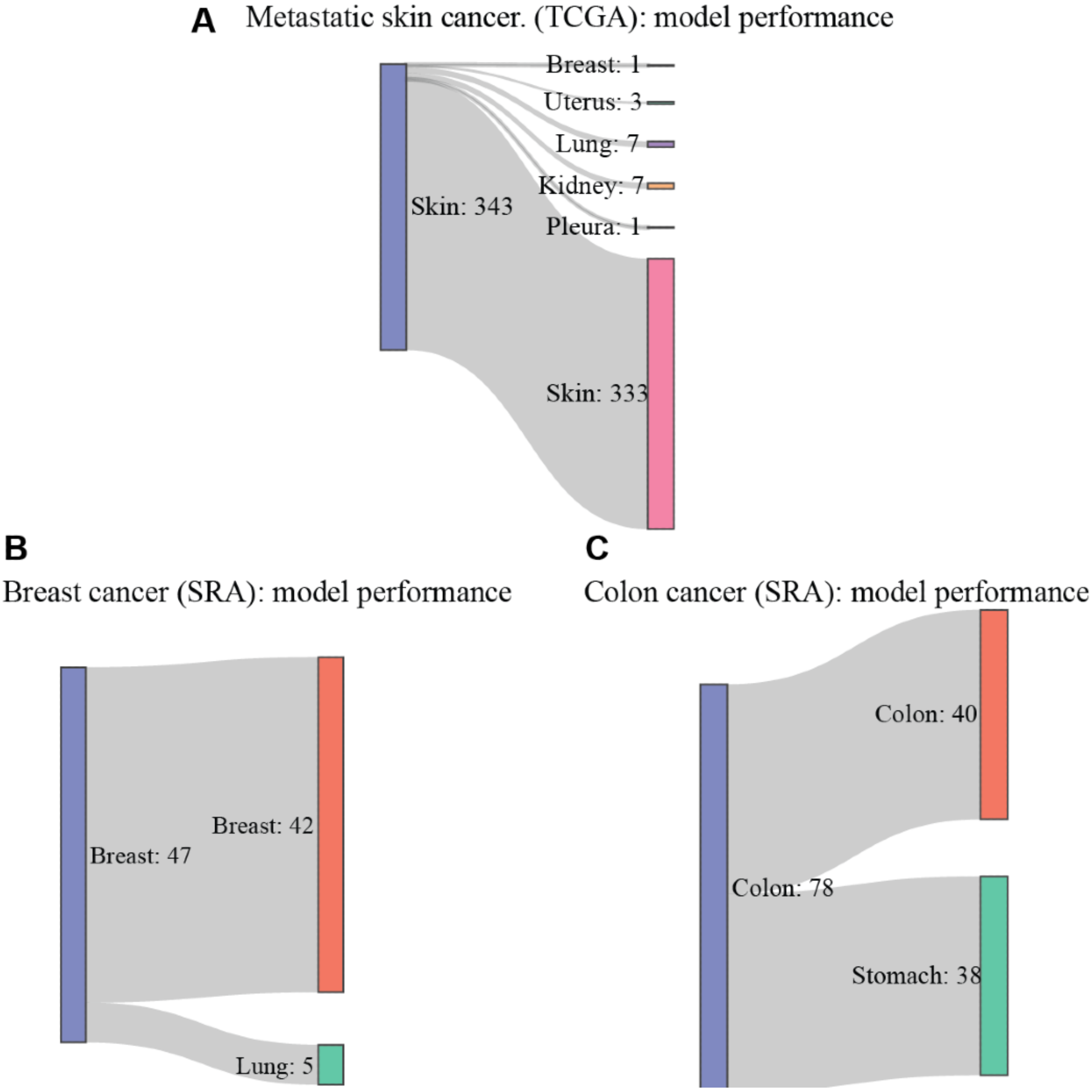


Figure 4. Sankey plot for deep learning model on Sequence Read Archive (SRA) and The Cancer Genome Atlas (TCGA) test data sets. (A) On the TCGA data set, our deep learning model is able to correctly classify 333 out of 343 metastatic skin cancer samples, demonstrating high accuracy. (B) On the SRA test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



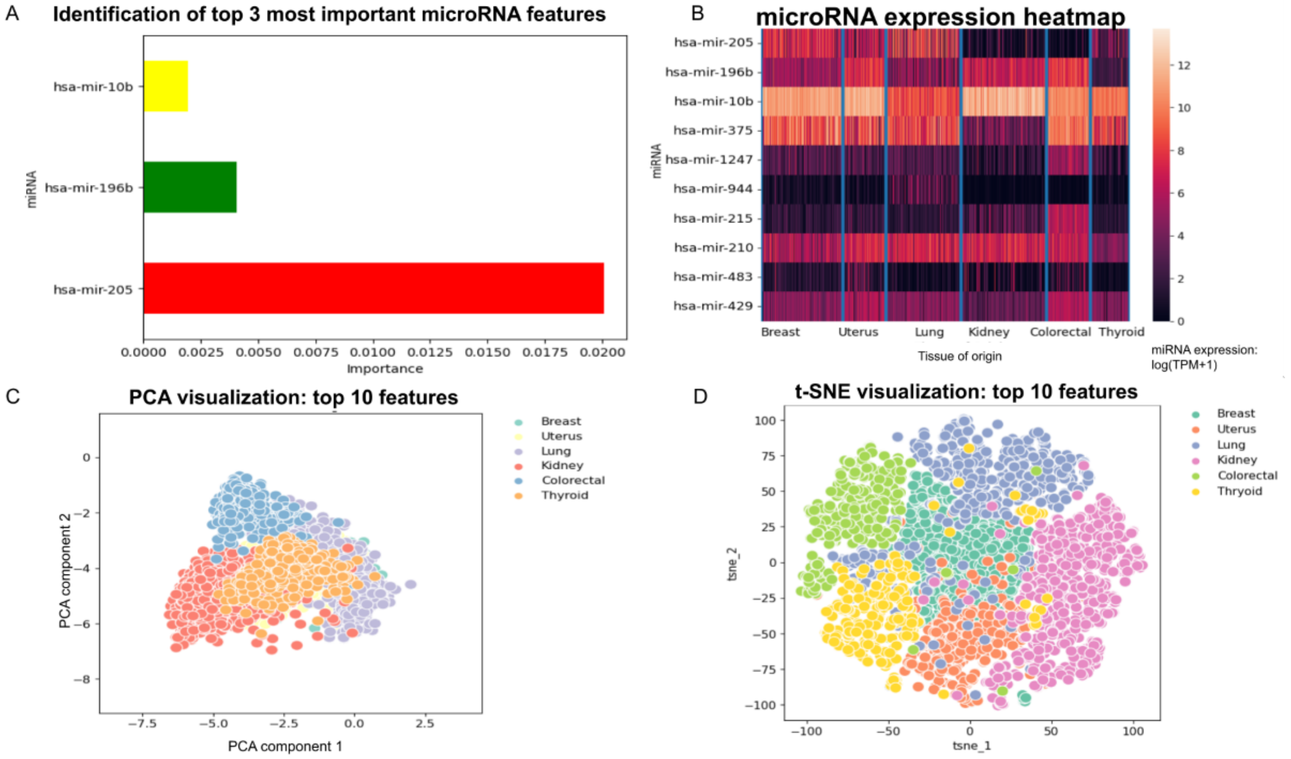
These results confirm our hypotheses and show that we were able to predict the TOO with high accuracy using deep learning. Furthermore, our findings demonstrated that deep learning techniques significantly increase the accuracy in comparison to decision tree, logistic regression, and random forest models.

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

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

XSL•FO
RenderX

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



Discussion

Principal Findings

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

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

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

Limitations

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

XSL•FO
RenderX

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

Conclusions

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

Acknowledgments

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

Data Availability

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

Conflicts of Interest

None declared.

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

References

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

XSL•FO
RenderX

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

XSL•FO
RenderX

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

Abbreviations

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

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

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

©Ananya Raghu, Anisha Raghu, Jillian F Wise. Originally published in JMIR Bioinformatics and Biotechnology (https://bioinform.jmir.org), 24.07.2024. 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.

XSL•FO
RenderX

Original Paper

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

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

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

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

Abstract

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

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

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

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

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

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

XSL•FO
RenderX

KEYWORDS

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

Introduction

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

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

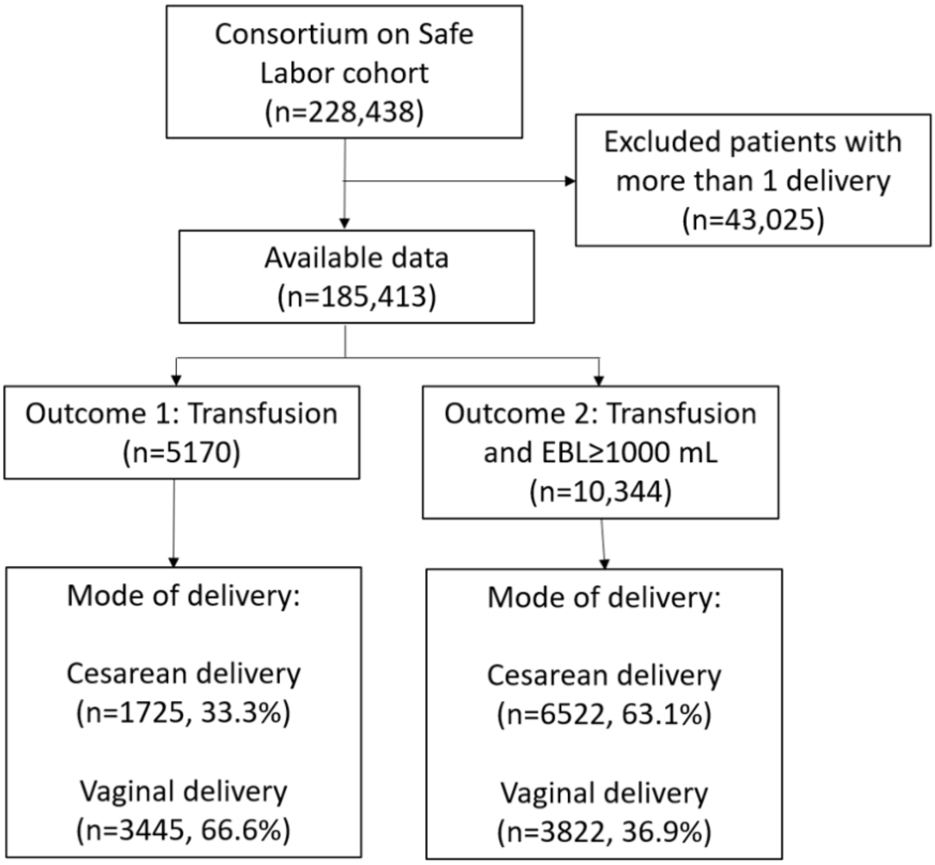
Methods

Data Collection

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

XSL•FO
RenderX

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


Consortium on Safe Labor cohort (n=228,438)
Excluded patients with more than 1 delivery (n=43,025)
Available data (n=185,413)
Outcome 1: Transfusion (n=5170)
Outcome 2: Transfusion and EBL≥1000 mL (n=10,344)
Mode of delivery:
Cesarean delivery (n=1725, 33.3%)
Vaginal delivery (n=3445, 66.6%)
Mode of delivery:
Cesarean delivery (n=6522, 63.1%)
Vaginal delivery (n=3822, 36.9%)

Missing Data

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

Feature Selection

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

Outcomes

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

Data Analysis

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

XSL•FO
RenderX

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

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

Ethical Considerations

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

Results

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

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

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

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

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

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

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

[d] MCC: Matthews correlation coefficient.

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

[f] GB: gradient boosting.

[g] RF: random forest.

[h] Emb: learned embedding.

[i] MLP: multilayer perceptron.

[j] TFIM: TensorFlow imbalanced.

[k] SVM: support vector machine.

[l] LR: logistic regression.

XSL•FO
RenderX

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

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

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

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

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

[d] MCC: Matthews correlation coefficient.

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

[f] GB: gradient boosting.

[g] RF: random forest.

[h] Emb: learned embedding.

[i] MLP: multilayer perceptron.

[j] TFIM: TensorFlow imbalanced.

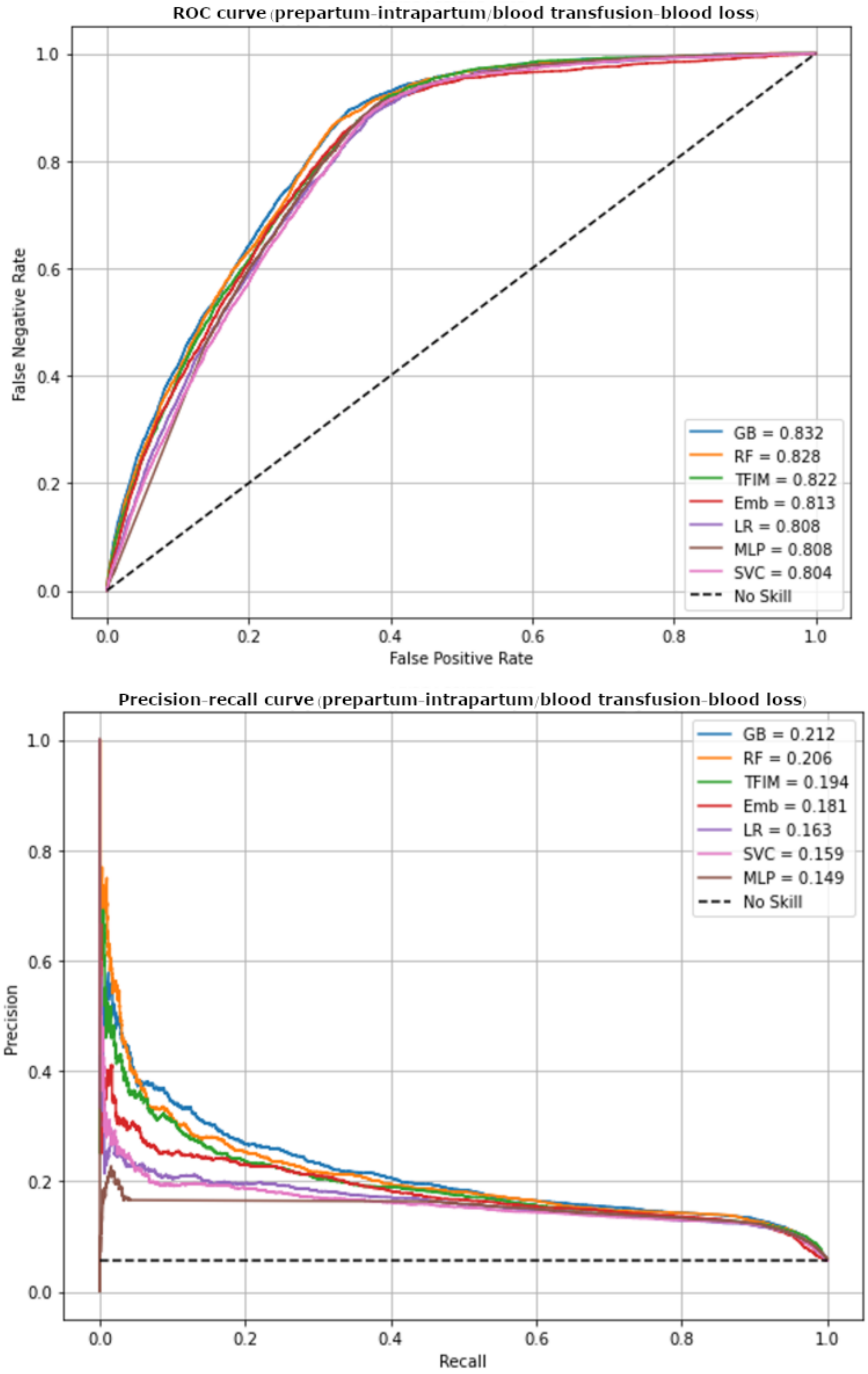
[k] SVM: support vector machine.

[l] LR: logistic regression.

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

XSL•FO
RenderX

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


ROC curve (prepartum-intrapartum/blood transfusion-blood loss)

Precision-recall curve (prepartum-intrapartum/blood transfusion-blood loss)

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.

XSL•FO
RenderX

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.

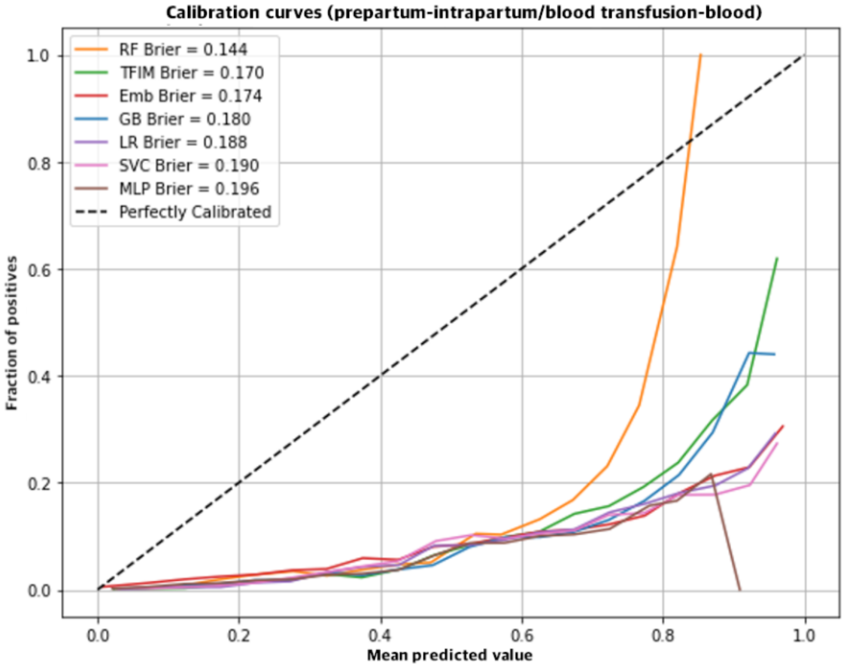
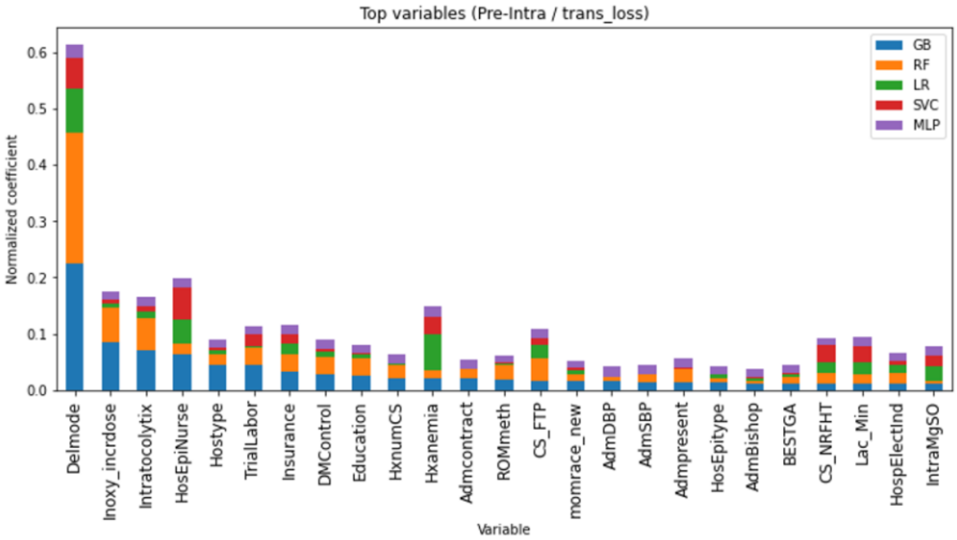


Figure 4 displays the top 25 predictive variables included for model development using antepartum and intrapartum features for the prediction of the transfusion-PPH composite. As the machine learning GB model was the best performing model overall, the variables in Figure 4 are in order of variable importance 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.



XSL•FO
RenderX

Discussion

Principal Findings

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

Clinical Implications

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

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

Research Implications

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

Strengths and Limitations

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

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

XSL•FO
RenderX

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

Conclusions

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

Acknowledgments

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

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

Conflicts of Interest

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

Multimedia Appendix 1

All antepartum and intrapartum variables were included for analysis for feature selection.

[DOCX File , 21 KB - bioinform_v5i1e52059_app1.docx]

Multimedia Appendix 2

Overall Patient Characteristics.

[DOCX File , 14 KB - bioinform_v5i1e52059_app2.docx]

Multimedia Appendix 3

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

[DOCX File , 14 KB - bioinform_v5i1e52059_app3.docx]

Multimedia Appendix 4

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

[DOCX File , 14 KB - bioinform_v5i1e52059_app4.docx]

References

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

XSL•FO
RenderX

2. Neggers YH. Trends in maternal mortality in the United States. Reprod Toxicol 2016 Sep;64:72-76. [doi: 10.1016/j.reprotox.2016.04.001] [Medline: 27063184]
3. Pregnancy Mortality Surveillance System. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/reproductivehealth/maternal-mortality/pregnancy-mortality-surveillance-system.htm [accessed 2021-01-09]
4. Collier AY, Molina RL. Maternal mortality in the United States: updates on trends, causes, and solutions. Neoreviews 2019 Oct;20(10):e561-e574 [FREE Full text] [doi: 10.1542/neo.20-10-e561] [Medline: 31575778]
5. Shields LE, Wiesner S, Fulton J, Pelletreau B. Comprehensive maternal hemorrhage protocols reduce the use of blood products and improve patient safety. Am J Obstet Gynecol 2015 Mar;212(3):272-280. [doi: 10.1016/j.ajog.2014.07.012] [Medline: 25025944]
6. Al-Zirqi I, Vangen S, Forsen L, Stray-Pedersen B. Prevalence and risk factors of severe obstetric haemorrhage. BJOG 2008 Sep;115(10):1265-1272. [doi: 10.1111/j.1471-0528.2008.01859.x] [Medline: 18715412]
7. Bateman BT, Berman MF, Riley LE, Leffert LR. The epidemiology of postpartum hemorrhage in a large, nationwide sample of deliveries. Anesth Analg 2010 May 01;110(5):1368-1373. [doi: 10.1213/ANE.0b013e3181d74898] [Medline: 20237047]
8. Lee H, Yoon SB, Yang S, Kim WH, Ryu H, Jung C, et al. Prediction of acute kidney injury after liver transplantation: machine learning approaches vs. logistic regression model. J Clin Med 2018 Nov 08;7(11) [FREE Full text] [doi: 10.3390/jcm7110428] [Medline: 30413107]
9. Deo RC. Machine learning in medicine. Circulation 2015 Nov 17;132(20):1920-1930 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.115.001593] [Medline: 26572668]
10. Pressly MA, Parker RS, Waters JH, Beck SL, Jeyabalan A, Clermont G. Improvements and limitations in developing multivariate models of hemorrhage and transfusion risk for the obstetric population. Transfusion 2021 Feb;61(2):423-434. [doi: 10.1111/trf.16216] [Medline: 33305364]
11. Ruppel H, Liu VX, Gupta NR, Soltesz L, Escobar GJ. Validation of postpartum hemorrhage admission risk factor stratification in a large obstetrics population. Am J Perinatol 2021 Sep;38(11):1192-1200 [FREE Full text] [doi: 10.1055/s-0040-1712166] [Medline: 32455467]
12. Hussain S, Guarini C, Blosser C, Poole A. Obstetric hemorrhage outcomes by intrapartum risk stratification at a single tertiary care center. Cureus 2019 Dec 24;11(12):e6456 [FREE Full text] [doi: 10.7759/cureus.6456] [Medline: 32025387]
13. Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. N Engl J Med 2016 Sep 29;375(13):1216-1219 [FREE Full text] [doi: 10.1056/NEJMp1606181] [Medline: 27682033]
14. Rosati G. Dealing with Missing Data. Medium. 2019. URL: https://towardsdatascience.com/dealing-with-missing-data-17f8b5827664 [accessed 2023-12-11]
15. Cramér H. Mathematical Methods of Statistics (PMS-9). Princeton, NJ: Princeton University Press; 1999.
16. Ahmadzia HK, Phillips JM, Kleiman R, Gimovsky AC, Bathgate S, Luban NLC, et al. Hemorrhage risk assessment on admission: utility for prediction of maternal morbidity. Am J Perinatol 2021 Sep;38(11):1126-1133 [FREE Full text] [doi: 10.1055/s-0040-1710501] [Medline: 32446252]
17. Bingham D, Melsop K, Main E. CMQCC Obstetric Hemorrhage Hospital Level Implementation Guide. The California Maternal Quality Care Collaborative (CMQCC). 2010. URL: https://www.cmqcc.org/resource/1489/download [accessed 2023-12-11]
18. Postpartum hemorrhage (PPH) risk assessment table 1. The AWHONN postpartum hemorrhage project. URL: https://cdn-links.lww.com/permalink/aog/b/aog_134_6_2019_10_06_kawakita_19-1065_sdc2.pdf [accessed 2021-01-07]
19. Dilla A, Waters J, Yazer M. Clinical validation of risk stratification criteria for peripartum hemorrhage. Obstet Gynecol 2013 Jul;122(1):120-126. [doi: 10.1097/AOG.0b013e3182941c78] [Medline: 23743452]
20. Kawakita T, Mokhtari N, Huang JC, Landy HJ. Evaluation of risk-assessment tools for severe postpartum hemorrhage in women undergoing cesarean delivery. Obstet Gynecol 2019 Dec;134(6):1308-1316. [doi: 10.1097/AOG.0000000000003574] [Medline: 31764744]
21. Venkatesh K, Strauss R, Grotegut C, Heine RP, Chescheir NC, Stringer JSA, et al. Machine learning and statistical models to predict postpartum hemorrhage. Obstet Gynecol 2020 Apr;135(4):935-944 [FREE Full text] [doi: 10.1097/AOG.0000000000003759] [Medline: 32168227]
22. Meyer SR, Carver A, Joo H, Venkatesh KK, Jelovsek JE, Klumpner TT, et al. External validation of postpartum hemorrhage prediction models using electronic health record data. Am J Perinatol 2022 Mar 02. [doi: 10.1055/a-1745-1348] [Medline: 35045573]
23. Leonard S, Kennedy C, Carmichael S, Lyell D, Main E. An expanded obstetric comorbidity scoring system for predicting severe maternal morbidity. Obstet Gynecol 2020 Sep;136(3):440-449 [FREE Full text] [doi: 10.1097/AOG.0000000000004022] [Medline: 32769656]
24. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015 Mar 4;10(3):e0118432 [FREE Full text] [doi: 10.1371/journal.pone.0118432] [Medline: 25738806]
25. Postpartum Hemorrhage. American College of Obstetricians and Gynecologists. 2017. URL: https://www.acog.org/en/clinical/clinical-guidance/practice [accessed 2021-07-11]

XSL•FO
RenderX

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

Abbreviations

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

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

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

©Homa Khorrami Ahmadzia, Alexa C Dzienny, Mike Bopf, Jaclyn M Phillips, Jerome Jeffrey Federspiel, Richard Amdur, Madeline Murguia Rice, Laritza Rodriguez. Originally published in JMIR Bioinformatics and Biotechnology (https://bioinform.jmir.org), 05.02.2024. 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.

XSL•FO
RenderX

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org

XSL•FO
RenderX