JMIR Bioinformatics and Biotechnology

Contents

XSL•FO
RenderX

Tutorial

Genetic Diversity and Mutation Frequency Databases in Ethnic Populations: Systematic Review

Shumaila Khan[1*], PhD; Mahmood Alam[2*], PhD; Iqbal Qasim[3], PhD; Shahnaz Khan[4], PhD; Wahab Khan[5], PhD; Orken Mamyrbayev[6], PhD; Ainur Akhmediyarova[7*], PhD; Nurzhan Mukazhanov[7*], PhD; Zhibek Alibiyeva,[7*], PhD

[1]Department of Computer Science, University of Science and Technology Bannu, Bannu Township, Bannu, Pakistan

[2]Faculty of Business, Law and Social Sciences, Birmingham City University, Birmingham, United Kingdom

[3]Hertfordshire Business School, University of Hertfordshire, Hatfield, United Kingdom

[4]Department of Chemistry, University of Science and Technology Bannu, Bannu, Pakistan

[5]Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Malaysia

[6]Institute of Information and Computational Technologies, Satbayev University, Almaty, Kazakhstan

[7]Institute of Automation and Information Technologies, Satbayev University, Almaty, Kazakhstan

[*]these authors contributed equally

Corresponding Author:
Shumaila Khan, PhD
Department of Computer Science, University of Science and Technology Bannu, Bannu Township, Bannu, Pakistan

Abstract

Background: National and ethnic mutation frequency databases (NEMDBs) play a crucial role in documenting gene variations across populations, offering invaluable insights for gene mutation research and the advancement of precision medicine. These databases provide an essential resource for understanding genetic diversity and its implications for health and disease across different ethnic groups.

Objective: The aim of this study is to systematically evaluate 42 NEMDBs to (1) quantify gaps in standardization (70% nonstandard formats, 50% outdated data), (2) propose artificial intelligence/linked open data solutions for interoperability, and (3) highlight clinical implications for precision medicine across NEMDBs.

Methods: A systematic approach was used to assess the databases based on several criteria, including data collection methods, system design, and querying mechanisms. We analyzed the accessibility and user-centric features of each database, noting their ability to integrate with other systems and their role in advancing genetic disorder research. The review also addressed standardization and data quality challenges prevalent in current NEMDBs.

Results: The analysis of 42 NEMDBs revealed significant issues, with 70% (29/42) lacking standardized data formats and 60% (25/42) having notable gaps in the cross-comparison of genetic variations, and 50% (21/42) of the databases contained incomplete or outdated data, limiting their clinical utility. However, databases developed on open-source platforms, such as LOVD, showed a 40% increase in usability for researchers, highlighting the benefits of using flexible, open-access systems.

Conclusions: We propose cloud-based platforms and linked open data frameworks to address critical gaps in standardization (70% of databases) and outdated data (50%) alongside artificial intelligence–driven models for improved interoperability. These solutions prioritize user-centric design to effectively serve clinicians, researchers, and public stakeholders.

Introduction

Background

Recent advancements in genomic techniques, such as next-generation sequencing and clustered regularly interspaced short palindromic repeats technology, have revolutionized the identification of gene mutations associated with disease, enabling precise disease diagnosis and personalized treatment strategies. Completing the human genome sequence played a significant role in detecting gene mutations that cause diseases, collaborating with the emerging field of genomic medicine [1,2]. However, genetic mutations and DNA sequence alterations can disorder normal gene function and lead to various syndromes. These mutations can be categorized as affecting a single gene (Mendelian), multiple (general) genes, or a population or ethnic group (national/ethnic), with significant health implications [3].

Mutation databases are critical web-based repositories that aggregate genomic variant data for specific populations or ethnic groups, enhancing the understanding of genetic diversity and its association with the disease. Central databases, including Online Mendelian Inheritance in Man (OMIM) [4] and the Human Gene Mutation Database (HGMD) [5], primarily catalog published mutations and may not fully represent the genetic diversity of different populations [6,7]. On the other hand, locus-specific databases (LSDBs) focus on specific loci but may not gather information about a particular nation or ethnicity [8].

Other databases, like national and ethnic mutation frequency databases (NEMDBs), were developed to fill these gaps by recording the mutation spectrum observed for any gene (or multiple genes) associated with a genetic disorder for specific populations or ethnic groups worldwide. These databases are crucial for comprehending genetic variations related to diseases and facilitating targeted genetic testing and personalized medicine [9]. Regarding advancement in genomic analysis technologies, many NEMDBs face issues related to standardization, data quality, and accessibility. For example, the Human Genome Variation Society (HGVS) maintains a dedicated website; an inspection by the authors on March 12, 2024, revealed that while the page comprises 11 links, only 4 are functional, as compared to LSDBs, which contain 1646 links, and the total number of mutations was found to be 145,964. Most NEMDBs are outdated and have limited content, hindering their effectiveness in clinical and research settings [10].

Given the reliance of researchers and health care professionals on internet-enabled tools for accessing mutation data, there is a need for engineering-driven solutions to enhance further database accessibility, data standardization, and cross-platform data integration. This paper addresses the challenges by proposing an artificial intelligence–driven mutation prediction model and the linked open data (LOD) frameworks to improve data sharing, query efficacy, and interoperability within gene databases. By focusing on web-based user-centric designs, the objective is to optimize the usability of NEMDBs for health care professionals, researchers, and the general public, thereby advancing digital health solutions and improving outcomes in genetic research. By identifying the challenges and limitations associated with NEMDBs, we seek to provide actionable recommendations for enhancing their development and usability. The key contributions of the review are as follows:

- This systematic review examined 42 NEMDBs to a). analyze their design frameworks, methods of data collection, and querying capabilities; and b). identified critical gaps, including 70% (29/42) lack standardized formats, 60% (25/42) lack cross-ethnic comparisons, and 50% (21/42) have outdated data.
- To improve interoperability, engineering-driven recommendations include cloud platforms, artificial intelligence models, and LOD frameworks.
- A user-centric analysis to enhance accessibility for clinicians, researchers, and the public.

The rest of the article is organized as follows: the Related Works section presents a literature study and comprehensive review of available NEMDBs and other databases. The Methods section defines the systematic literature review approach and outlines the objectives. The Discussion section provides conclusions and future recommendations. Finally, the Conclusion section summarizes the review.

Related Works

Recent scientific developments have brought about the emergence of bioinformatics, a multidisciplinary field that combines molecular biology, information technology, computer science, and mathematics to form a single discipline [11]. Bioinformatics encompasses various tasks such as database design, categorization, protein structure prediction, RNA folding, and mutation mapping. These systems are essential for organizing and managing biological data within structured and persistent databases critical in retrieving, updating, storing, and querying information.

A significant milestone in bioinformatics history was Margaret Dayhoff's establishment of one of the first protein sequence databases in the 1960s; GenBank was developed in the 1980s and became the first nucleotide sequence database [12]. Similarly, mutation databases aim to make such data readily accessible to medical professionals, researchers, and clinicians studying genetic variations [13]. Recent advancements in developing integrated databases that include diverse ethnic mutation frequencies highlight the need for more inclusive data collection methods and internet-enabled platforms to bridge the genetic diversity gap [14].

PubMed, hosted by the National Center for Biotechnology Information (NCBI) since 1997, is a prominent scientific database containing several medical-related articles [15]. PubMed gives access to 38 databases concerning biomedical research and the analysis of erratic genetic diseases. Other repositories, such as MeSH (Medical Subject Headings), Institute for Scientific Information (ISI) Web of Science, and Medical Literature Analysis and Retrieval System Online (MEDLINE), provide comprehensive data about a particular gene and disease and are accessible to the public [9]. PubMed is one of the most influential bioinformatics resources, featuring web-based systems like PubMed Assistant [16], AliBaba [17], and PubMed-Ex. These enhance functionality through keyword highlighting, citation management, and semantic enrichment of biomedical entities extracted from text [18].

Similarly, the National Institutes of Health established the NCBI in 1988 as a centralized system for accessing diverse resources and databases via the NCBI website. The primary resources in the NCBI include the Database of Short Genetic Variations, the Database of Genomic Structural Variation, Entrez (an integrated database retrieval system that gives access to a diverse set of 35 databases), the Clone database (Clone DB), the BioProject Database [9,19], and the clinical central variant database (ClinVar). Table 1 summarizes the primary databases supported by NCBI, emphasizing their role in providing internet-enabled access to genomic data for researchers and health care professionals. Such internet-enabled systems streamline the extraction and analysis of gene mutation content and support

collaborative research by facilitating data sharing across diverse platforms. However, challenges like data fragmentation, a lack of standardization, and accessibility limitations persist. Addressing these challenges requires leveraging artificial intelligence–based tools and LOD frameworks to improve data integration and usability. Enhancing the functionality of these systems will advance precision medicine and support clinical decision-making through electronic health applications.

Table . NCBI[a] databases.

References	Database Name	Brief description
Bianco et al [9]	Bio Project Database	The database allows users to submit detailed research studies from intensive genome sequences projects to huge worldwide associations.
Bianco et al [9]	BioSample Database	The Biosample Database is a new resource that annotates biological samples used in various NCBI-submitted studies, including genome-wide association studies, epigenetics, genomics sequencing, and microarrays.
Landrum et al [3]	Clinical variant database (ClinVar)	ClinVar is a database that contains human genomic variants and their relevant disease. The database is publicly available.
Sayers et al [20]	PopSet Database	This database contains different sets of data that were submitted to GenBank. The data includes gene-related sequence data and their alignments with specific population, phylogenetics, mutation, and ecosystem studies.
Sayers et al [20]	Clone database (Clone DB)	The database incorporates clones and library information, including sequence data, map positions, and information distribution. It also offers filtering by organism and vector types.
Sayers et al [20]	MMDB (Molecular Modeling Database)	It details sequence alignments and profiles representing protein spheres preserved in molecule evolution.
Boguski et al [21]	Database of expressed sequence tags (dbEST) Nucleotide EST Database	This database collects sequence tags and includes details about complementary DNA (transcript) sequences. dbEST is accessible directly via the Nucleotide EST Database.
Church et al [22]	Database of Genomic Structural Variation (db-Var)	It was designed to collect details about large-scale genomic variation, including large insertions, deletions, translocations, and inversions. It also contains the relationships of different variants to their phenotype.
Louhichi et al [23]	Entrez	Entrez is a rich database that integrates information from 35 databases containing over 570 million biological data records. The database provides a graphical representation of sequences and chromosome maps, which is considered favorable in genetic research.
Mailman et al [24] and GAIN Collaborative Research Group et al [25]	Databases of Genotypes and Phenotypes (dbGaP)	The database contains information about genotype and phenotype. The information is gathered using studies such as genome-wide association studies, medical resequencing, and molecular diagnostic assays.
Sherry et al [26]	Database of Short Genetic Variations (dbSNP)	This database, similar to HapMap, was developed to support large-scale polymorphism detection. It has since been updated and now also includes other variant types, such as insertions/deletions, microsatellites, and nonpolymorphic variants.
Sherry et al [26]	Database of Major Histocompatibility Complex (dbMHC)	dbMHC hosts two key resources: (1) an interactive alignment viewer for HLA (Human Leukocyte Antigen) and related genes and the Major Histocompatibility Microsatellite Database.

[a]NCBI: National Center for Biotechnology Information.

Catalog of Human Variation Databases

Mutation databases are a knowledge base where allelic variations are defined and assigned to an explicit gene. Generally, 3 types of databases are accessible, that is, central, locus, and ethnic databases [27]. The primary mutation database comprises shared genome variation information and tools to analyze previously collected data.

Central Databases

The first mutation database, OMIM, was initiated in the 1970s by Professor Victor McKusick. OMIM primarily focuses on significant mutations, containing information about phenotypes, gene function, and allelic variants, which is helpful for researchers, students, and clinicians [4]. The website has been frequently updated and can be easily accessed. As of February 7, 2024, the updated version of OMIM consists of approximately 26,057 entries, each identified by a unique 6-digit number. Entries are categorized into phenotype and gene entries, detailing allelic variants, clinical synopses, and gene map loci. Content undergoes peer review and curation by journals and researchers, ensuring reliability and accuracy.

Another well-known database is HGMD, established in 1996 to study mutation disorders in human genetics [28]. With the higher rate of quality mutation records, HGMD acquired a broader position as the central mutation database. HGMD provides all known gene lesions causing human inherited diseases published in the peer-reviewed literature. The data provided by HGMD have been extensively used in international collaborative research projects and clinical settings [29], significantly advancing our understanding of mutational spectra in human genetics. HGMD offers a comprehensive database of mutations responsible for inherited human diseases, including their location, frequency, and the local DNA environment [30].

Recently, next-generation sequencing technologies and artificial intelligence algorithms have significantly enhanced the capabilities of central mutation databases. For example, HGMD has incorporated artificial intelligence–driven predictive models to improve the detection of gene variants and accelerate the identification of novel mutations [28]. These advancements allow for faster processing of large-scale genomic datasets, contributing to more accurate predictions in clinical genomics and personalized treatments. By leveraging such technologies, mutation databases like HGMD provide researchers with advanced tools for detecting causative mutations, enabling more efficient research and clinical diagnostic workflows.

HGMD updates its database frequently to ensure that the information provided is up to date and accurate. HGMD is accessible in two versions. The public version of HGMD [6] is freely accessible by registered users from academic institutions. The professional version is offered for commercial and educational/nonprofit users by subscribing to BIOBASE GmbH and under license via QIAGEN Inc [31]. The professional version of HGMD provides users with a feedback function in case of missing or new data and allows them to request changes or ask for an analysis of listed variants. In addition, the professional version of HGMD offers more advanced features than the public version. The latest version of HGMD was released in 2017, and statistics from April 2021 showed that the database contained 352,731 gene lesion entries in the HGMD Professional release, of which 234,987 entries were manually curated from academic and nonprofit sources and published journals.

LSDBs

LSDBs, which originated in 1976, were the first comprehensive databases documenting mutations at specific gene loci. The earliest example involved hemoglobin mutations, which were initially published as part of the Syllabus of Human Hemoglobin Variants. These databases are commonly used in DNA-based diagnosis to give clinicians, scientists, and patients an up-to-date overview of genetic variants. Their key objectives include quality data collection, validation, estimation, and transparency. Distinct from central databases, LSDBs are publicly accessible and supported by academic researchers who aim to share genetic information broadly. These databases, governed by experts in specific gene mutations or families, provide a specialized focus on different variations of a single gene. Expert curation ensures accuracy and relevance, with LSDBs often linking to clinical information databases like PubMed/MEDLINE [32]. Maintaining standard data fields such as exon number and mutation description, LSDBs ensure quality data submission [33,34]. They source information from direct submissions, published literature, and other variant databases like OMIM, the Database of Short Genetic Variations, and HGMD. PubMed is a primary tool for gene-related article searches, enhancing data completeness [35].

Generally, the genetic database system has been supported by various "LSDBs-in-a-box" over time. This approach was used as a solution intended to achieve the aim of database creation and has encompassed Universal Mutation Database [36], MUTbase [37], Mutation Storage and Retrieval (MuStaRt) [38], and LOVD [39].

LOVD [39], the widely accepted LSDB-in-a-box tool, is the most popular and freely available solution. LOVD was released in December 2012 and has been updated over time. LOVD 3.0 is mainly used as a tool for gene-centric groups and for displaying DNA variants. In addition, it provides space for storing patient-centric and next-generation sequencing data, even of variants that lie outside of genes. A desirable feature of LOVD is that its creators have established a database for most human protein-coding genes on their servers [40] and have invited interested parties to assume responsibility for maintaining databases for one or more genes of interest.

Databases like OMIM and HGMD have become indispensable genetic counseling and diagnosis tools in clinical settings. Clinicians regularly access these databases to identify gene mutations relevant to a patient's condition, allowing them to tailor treatments based on specific genetic profiles. Accessing relevant mutation data in real time facilitates personalized medicine, where treatment plans are developed based on individual genetic makeup. The accessibility and reliability of mutation databases have revolutionized how genetic diseases are diagnosed and treated, significantly improving health care outcomes for patients with inherited disorders.

NEMDBs

Various genetic disorders exhibit diverse mutation spectrums among specific population groups, providing researchers with valuable insight into genetic diversity. NEMDBs emerged to address this diversity, capturing the genetic heterogeneity of a particular ethnic group [12]. The HGVS maintains a catalog of central databases, LSDBs, and NEMDBs. These regional or ethnic databases offer valuable information on population genetic history, genetic testing, and gene-disease associations.

Figure 1 shows the architecture of the NEMDBs, representing 3 main architectural approaches: Ethnic and National database Operating Software (ETHNOS)–based architecture, 3-tier architecture, and LOVD architecture. ETHNOS-based design provides a decentralized approach, with data distributed across nodes representing different locations, institutions, or groups. Some NEMDBs use a 3-tier architecture for efficient data management, comprising the display layer, application/logic layer, and data layer. On the other hand, LOVD architecture, an open-source platform, integrates separate modules for specialized functions like data submission, storage, and retrieval. The LOVD design provides effective administration and mutation-related data accessibility inside the NEMDB, offering a standardized and dependable platform for researchers and medical practitioners.

Figure 1. National and ethnic mutation frequency databases.
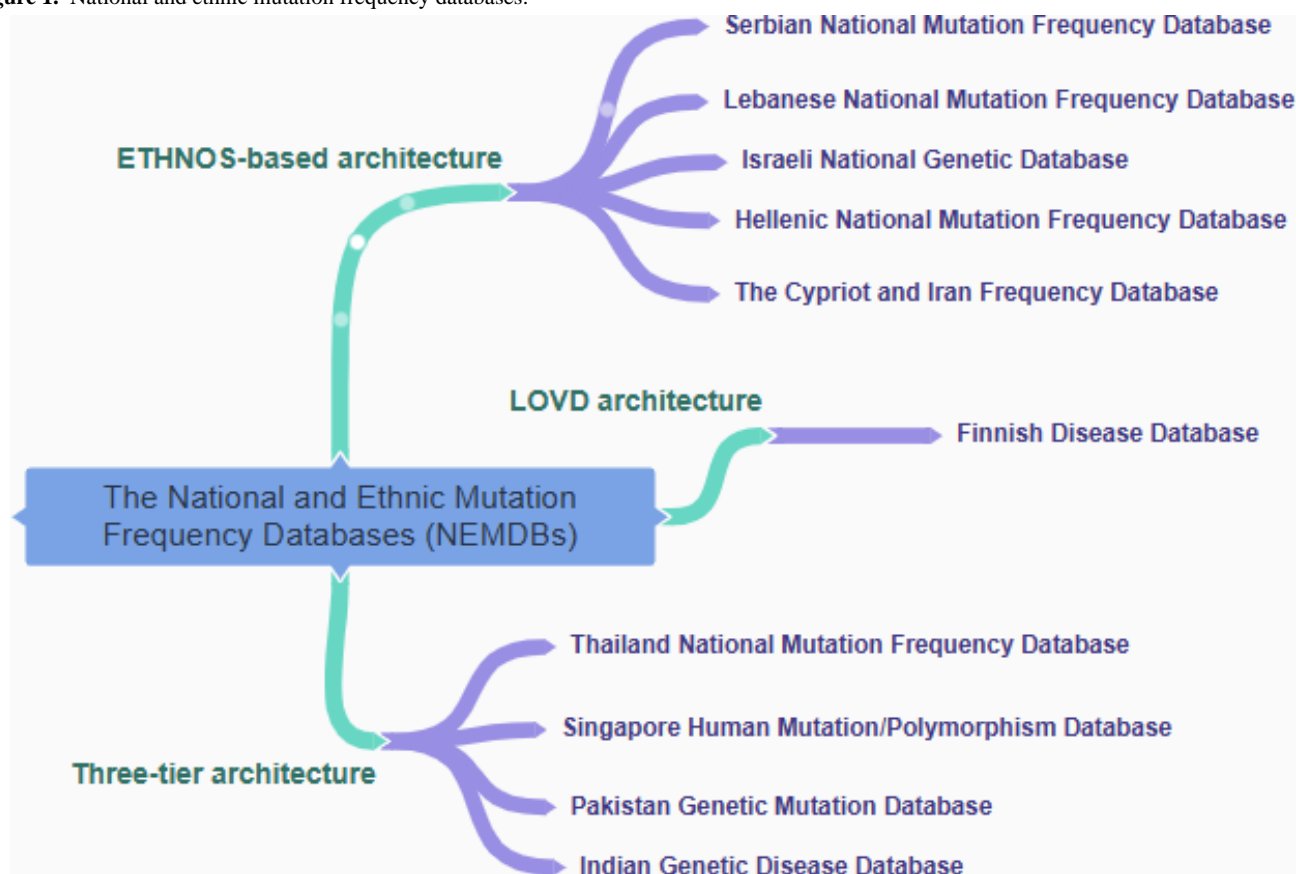
The genetic diversity captured in NEMDBs allows researchers to develop targeted strategies for detecting and diagnosing genetic disorders. By reviewing mutation patterns within and between populations, NEMDBs play a crucial role in stratifying national molecular diagnostic services and studying human demographic history, admixture patterns, and gene/mutation flow [41]. Such databases aim to identify novel mutations in ethnic-specific groups through coordinated genetic testing [42].

Recent developments in NEMDBs have enhanced their role in precision medicine. Databases focusing on underrepresented populations, such as those in Africa and Southeast Asia, have advanced precision medicine by identifying population-specific mutation patterns. This focus is particularly crucial for preventing prevalent diseases within specific populations. For example, the African Genome Variation Project and the Indian Genome Variation Database have provided data supporting personalized health care initiatives [42]. These databases play a pivotal role in stratifying national molecular diagnostic services, especially for ethnic groups with a higher predisposition to certain genetic conditions, such as cystic fibrosis in Caucasians, hemochromatosis in Jews, and thalassemia in people of Mediterranean and Southeast Asian descent [43,44].

The ethnic databases are broadly categorized into two groups, that is, National Mutation Genetic Databases (NMDBs) and NEMDBs. [45]. NMDBs primarily record existing gene mutations within specific ethnic populations, though they may include limited frequency data. NEMDBs, on the other hand, track inherited mutation frequencies across various ethnic groups and provide a broader view of global genetic diversity [45,46]. Examples of NEMDBs are listed in Table 2.

Table . National and ethnic mutation frequency databases.

References	Database	Brief description
Peltonen et al [47]	Finnish Disease Heritage, 2002	This database contains comprehensive information about gene mutations in the Finnish population. Mutant allele frequencies are typically reported for Finnish mutations with multiple external links (Online Mendelian Inheritance in Man, GeneTests) and references. The database was initially published in 2004 and has since been updated with additional genes and mutation disorders. This database was designed using the LOVD platform.
Patrinos et al [12]	The Iranian National Mutation Frequency Database, 2006; Cypriot National Mutation Frequency Database, 2006	Here, 2 similar databases are presented, one for the population of Cyprus and the other for the Iranian population. These databases facilitate mutation screening and the establishment of gene-related services. Both of the databases were developed using the ETHNOS[a] platform.
Bianco et al [9]	Hellenic National Mutation Database, 2005	This database aims to provide qualitative and updated reports of genetic disorders in the Greek population. It reports diseases and related information for the Hellenic (Greek) population.
Zlotogora et al [48]	Israeli National Genetic Database	The Israeli National Genetic Database was developed using the Electronic Tool for Human National and Ethnic Mutation Frequency Databases (ETHNOS) platform. This resource includes the Israeli National and Ethnic Mutation Frequency Database (NEMDB), which provides a detailed list of registered laboratories offering genetic testing services for the Israeli population through a dedicated query interface
Nakouzi et al [49]	The Lebanese National Mutation Frequency Database, 2006	This database was designed to analyze the genetic diseases in the population of Lebanon.
Sefiani et al [50]	The Moroccan Human Mutation Database, 2010	This database was developed to report the various mutation disorders found in the population of Morocco. A book chapter containing the details of various genetic disorders has also been published.
Ruangrit et al [51]	Thailand Human Mutation and Variation Database, 2008	ThaiMUT is an online ethnic database reporting mutation disorders in Thailand's population. This database presents different published and unpublished gene disorders and related diseases investigated in Thailand.
Pradhan et al [52]	Indian Genetic Disease Database, 2010	A database that integrates gene-related diseases in the Indian population. Domain experts have curated the diseases of this database. The database was developed using a 3-tier architecture.
Qasim et al [42]	Pakistan Genetic Mutation Database	The database contains information about different disorders occurring in the Pakistani population. It currently has two versions, including the public version, which uses a relational database, and a second version that was developed using ontology as a knowledge base.
Romdhane et al [53]	Tunisian National Mutation Frequency Database	This database was developed to collect data about the different genetic disorders found in the Tunisian population.
Tadmouri et al [54]	CTGA[b], 2006	The CTGA database is an open-access repository of information and findings on human gene variations and inherited, heritable genetic disorders in Arabs; it is constantly updated.

XSL•FO

RenderX

References	Database	Brief description
Horaitis et al [27]	Singapore Human Mutation Database, 2006	The database contains mutations found in Singapore for Mendelian diseases. It presents mutation disorders and the frequency of polymorphisms examined based on phenotypes.
Rajab et al [55]	Oman Genetic Mutation Database, 2015	The database was developed to collect and manage the mutations found in the Oman population. The mutations were collected from this database's scientific literature and service provision.

[a]ETHNOS: Ethnic and National database Operating Software.

[b]CTGA: Catalog of Arab Disease Mutation Database.

Methods

Overview

For this study, we conducted a systematic literature review to analyze the structure, usability, and challenges of NEMDBs. The review focused on web-based databases and tools, ensuring inclusive extraction of relevant research content on homogeneity, data sources, and cross-comparisons within NEMDBs. Figure 2 demonstrates the step-by-step selection process used in this research, presenting the systematic literature review approach by outlining objectives for extracting and analyzing relevant information. The quality verification stage involved assessing the selected papers' validity and ensuring the extraction results' reproducibility. Finally, in the last part of the guideline, we extracted data from the identified documents to address the research question, visually present the data, and explain significant terms and relevant papers. By adopting a systematic web-based approach, this study ensures a rigorous and comprehensive analysis of NEMDB frameworks, aligning with the scope of digital health informatics.

Figure 2. Steps included in the review protocol.

Search Strings and Data Sources

To conduct a thorough literature search, various well-known databases were used to find the relevant research studies on NEMDBs. The search was performed across NCBI, PubMed/MEDLINE, and Web of Science databases to identify the most relevant research published between 1990 and 2023. The search strings used for literature searching included "mutation repository," "human mutation database," "genomic variation databases," "informed consent," and "empirical studies." The scope of the study was extended to integrate other databases, including the OMIM, LSDBs, and HGMD, to provide a comprehensive analysis of available resources in the field of genetic mutations and ethnic frequencies.

Selection of Studies

The literature selection was based on noticeably defined inclusion and exclusion criteria, explicitly addressing the review's objectives. Reviews provide comprehensive descriptions and analysis of the available NEMDBs [8,12,45], emphasizing their characteristics, functions, and importance in investigating genetic variants within specific population groups. This review included papers if they satisfied the following criteria.

Inclusion criteria were as follows:

- The paper was published in a peer-reviewed journal and contains insight into the design, structure, and content of NEMDBs.

- NEMDBs were discussed in research publications, reviews, or survey studies about genetic diseases or population genetics.
- Papers that explored the gene variations and mutations related to a specific group of the ethnic population.
- Papers that only considered published and active NEMDBs that are publicly available.
- Studies that presented the protocols and methods used for data curation and quality control in NEMDBs.

Exclusion criteria were as follows:

- Papers that did not focus on ethnic-specific mutation databases.
- Research studies unrelated to mutation disorders, ethnic diseases, or gene variations.
- Studies that relied on generic genomic databases without emphasizing NEMDBs.
- Papers having minimal empirical proof or practical use.

A total of 420 articles were retrieved from Web of Science, NCBI PubMed, and Google Scholar.

Quality Verification

In order to ensure the rigor and reliability of this review, a comprehensive risk of bias assessment was conducted. The articles were evaluated using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist. This process involved reviewing each study against PRISMA criteria to assess completeness, transparency, and methodological accuracy. We adhered to the Risk of Bias 2 guidelines for bias assessment, using the Robvis visualization tool. Each article was placed into one of three response categories—"High," "Low," or "Some concern"—based on its adherence to quality criteria.

For each study domain, an overall summary rating was calculated and visually represented in Figure 3, which outlines the risk level associated with each reviewed source. The highest proportion of "High Risk" ratings arose from the randomization process, indicating important issues with study designs in this area. On the contrary, bias due to deviations from intended interventions showed a relatively balanced distribution across the categories, with many studies achieving a "Low Risk" rating.

Figure 3. Risk of bias assessment of the selected studies.
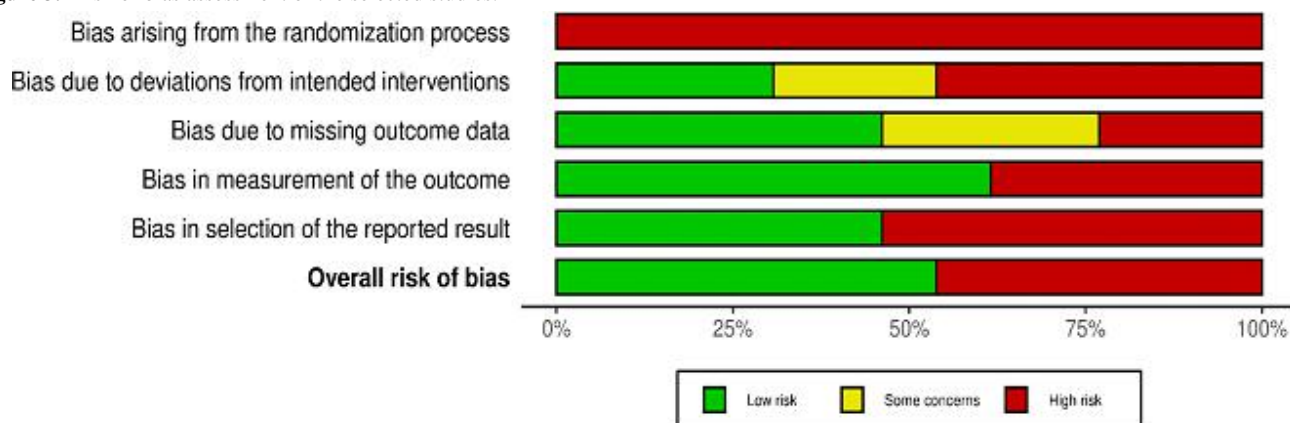
For bias due to missing outcome data, there was a more mixed distribution, with several studies flagged under both the "High" and "Some Concern" categories. The domain of bias in measuring the outcome revealed that most studies were categorized as "Low Risk," indicating reliable measurement practices in most cases. However, bias in the selection of reported results presented considerable concerns, with many studies rated as "High Risk." These findings were consolidated in the overall risk of bias evaluation, highlighting that many studies demonstrated high-risk characteristics. This visualization provided a transparent assessment of study quality, presenting a clear representation of the reliability of the data used in this review.

EndNote (version 20.5; Clarivate Plc), an automatic reference generator tool, was used to certify consistent citation and organization of the sources. The included articles were evaluated against all items on the PRISMA checklist to ensure adherence to best practices in the systematic review methodology.

Data Extraction and Analysis

The data extraction process involved a thorough review of each paper, focusing on identifying essential information relevant to

the objectives of this review. Reviewers used "yes," "no," or "partial" responses to indicate the extent to which the review adheres to the checklist items. Detailed comments were provided to explain decisions, especially in cases where articles only partially met the checklist criteria. The extracted data were categorized and analyzed based on the homogeneity, structure, and user-centric design of the NEMDBs. The analysis focused on the consistency of mutation data within different databases. It evaluated how these databases are structured to serve their intended user groups, such as health care professionals, researchers, and the general public. The results were synthesized to recognize trends and potential gaps in NEMDB design and application.

Results

Overview

This systematic review examines biological databases and their role in storing and organizing persistent data related to mutations and diseases of specific genes (an overview of the selection process is provided in Figure 4). These databases serve as knowledge bases and require curation by experts to maintain

the accuracy and relevance of the information. Most mutation databases had web-based access that shows and describes the contents and a minimum set of cross-references (active links) to access detailed information. Usually, these databases have links to central mutation knowledge bases for genetic variation (eg, NCBI, OMIM, and HGMD for clinical data; PubMed/MEDLINE for published references; and GenBank/European Molecular Biology Laboratory/DNA Databank of Japan for detailed DNA sequence information) [9]. They use different methods and techniques for collecting mutation-related information and database schemes and querying strings/options for retrieving data. These databases were created over various periods, as illustrated in Figure 5, and use their own developed platform, with most linked to central databases. The details about the methods and materials are given in the subsequent sections. Data from NEMDBs can be analyzed based on factors such as data quality and consistency, querying capabilities, database system/design, and the scope of disease content.

Figure 4. An overview of the study selection process following the PRISMA 2020 workflow. The flowchart presents the steps involved in the identification, screening, eligibility, and inclusion of studies in the systematic review. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses. *Duplicates were removed using EndNote X20.5 and manual screening. **Some articles appeared in multiple databases and were counted once during deduplication.
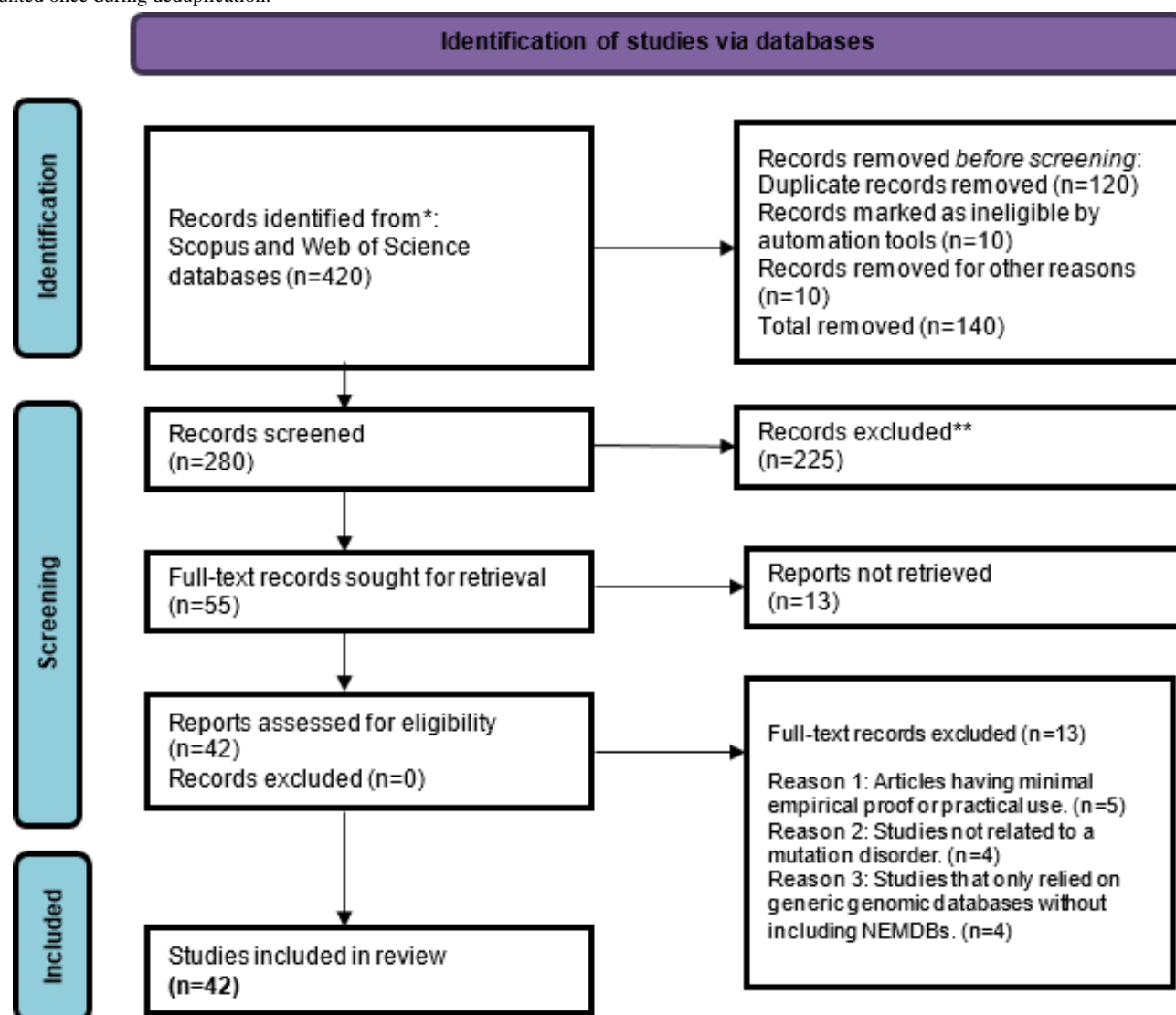
Figure 5. A catalog of NEMDBs. NEMDBs: national and ethnic mutation frequency databases.
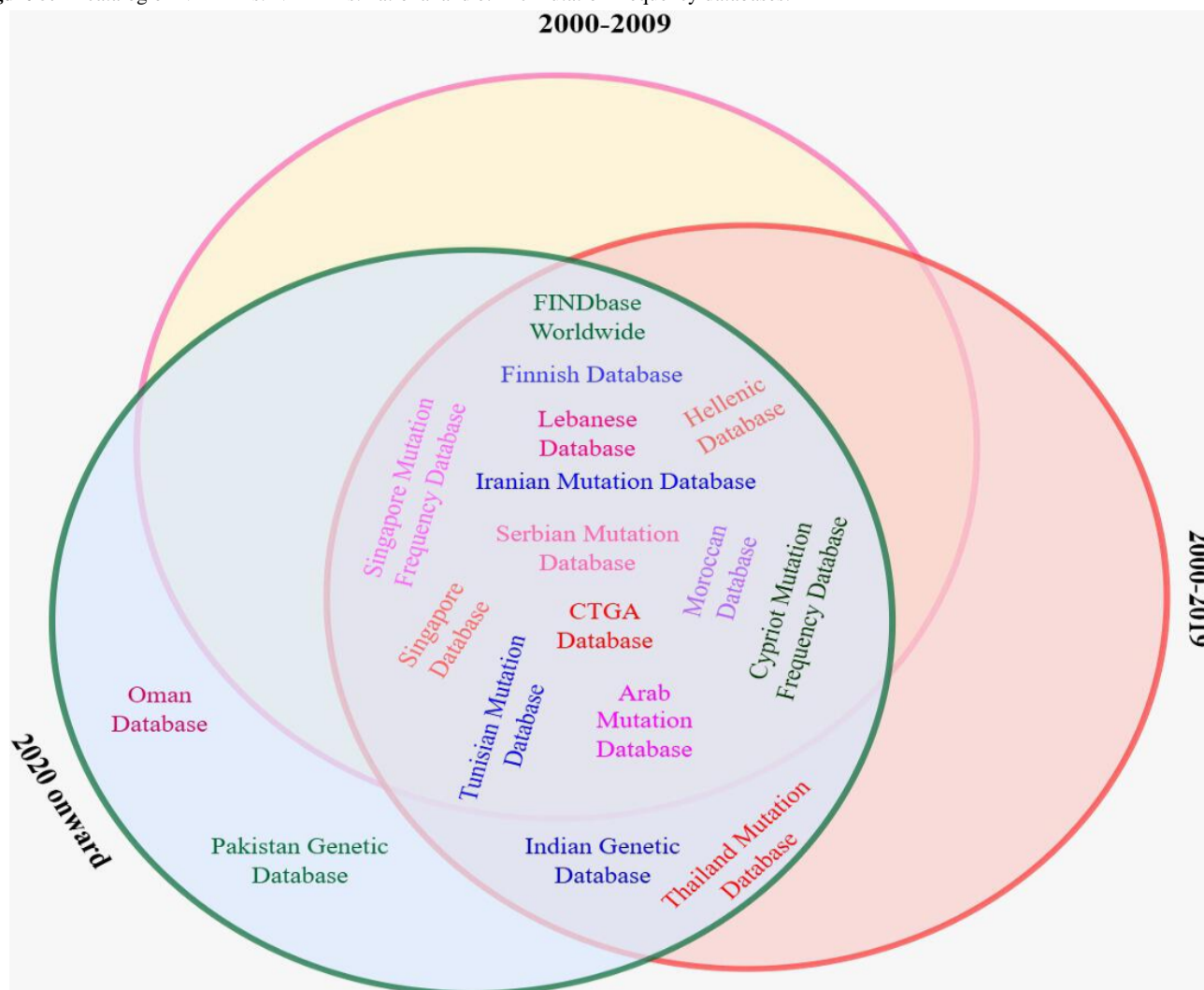
System Design and Data Accessibility

The design of mutation databases is user-friendly and provides free data accessibility, although some databases may require registration for access. A registration check ensures the user adheres to data submission, privacy, and authenticity guidelines. Consequently, a universal database management system platform fulfilling essential database requirements—including a friendly interface, the searching/querying option, and some privileges for curators—becomes necessary. Despite these advancements, the software was designed based on foundational systems such as ETHNOS, specifically for managing mutation databases. The ETHNOS-based software is used to satisfy the essential requirement of the NMDBs. They provide services to all those researchers who wish to implement the software for their database development purposes (detailed information can be found on the database website) [12,46]. ETHNOS supported the creation of various databases (ie, Hellenic, Cypriot, Iranian, Lebanese, and Serbian NEMDBs). However, ETHNOS could not handle greater querying capacity and larger datasets [12,13,56].

The Frequency of the Inherited Disorders database (FINDbase), a relational database established on an upgraded version of ETHNOS software capable of handling larger datasets, refers to the frequency of low alleles leading to inherited disorders in various ethnic populations worldwide [57]. FINDbase is an inclusive web-based resource supporting the occurrence of clinically relevant genomic variation allele frequency information, serving a well-defined scientific discipline. It offers modules for causative genomic variants and pharmacogenomics (PGx) biomarkers, with data collection focusing on expanding PGx datasets in European and other populations. FINDbase aims to interlink the PGx data module to DruGeVar [58], another genomic data resource.

Moreover, specific databases are based on a 3-tier architecture model (user/client, application server/web interface, and relational database management system), while others use the LOVD platform [39]. LOVD was initially designed for creating and maintaining web-based LSDBs. It is platform-independent software that uses PHP and MySQL only. The LOVD software has many variations, including LOVD v.2.0 [59] and LOVD v.3.0, following the HGVS. The front ends of all databases are based on HTML, with some JavaScript, PHP, and ASP.Net, and they rely on Cascading Style Sheets support. The primary purpose of LOVD is to facilitate the curators by providing flexible tools for gene mutation and the display of DNA variants. LOVD v.3.0 was updated on May 30, 2024. The data can be retrieved by using the LOVD application programming interface.

Quality Data Collection

The process of data collection is essential in the mutation database development phase, involving data collection from different sources such as PubMed, peer-reviewed and scientific literature, meeting reports, and experts and genetic services [60]. Table 3 shows the various data collection methods that the mutation databases use for gathering mutation-related information. Data can also be identified through automated text mining and manual journal screening and linking the unpublished mutation data presented in publicly available LSDBs; for example, the mutation databases may have a link to the HGMD database that facilitates users with access to LSDBs, for both published and unpublished materials [5].

Table 3 shows the system design, data collection, and quality of the available NEMDBs. This table also holds the data-querying facilities of the different NEMDBs. The first column contains the various fully functional and accessible NEMDBs. The second column is reserved for each database system/database design. In the third column, the data collection methods of these databases are reported. Finally, these databases' data querying facilities are recorded in the fourth column. Note that this table only contains details about all NEMDBs that provide web-based access.

Table . Materials and data collection methods.

National mutation genetic database or mutation database	System design	Data accessing			Query or search string	
		PubMed or published	Direct submission from experts or laboratories	Other sources	Disease name, disease category, or gene name	Dropdown lists or options
Arab Genetic Disease Database (AGDDB)		✓	✓		✓	✓
Repository of mutations from Oman		✓	✓	✓		
Hellenic National Mutation database	ETHNOS[a]-based	✓	✓	✓		✓
The Cypriot and Iran National Mutation Database	ETHNOS-based	✓	✓	✓		✓
Israeli National Genetic Database (INGD)	ETHNOS-based	✓	✓	✓	✓	
Singapore Human Mutation/Polymorphism Database (SHMPD)	Three-tier architecture	✓	✓	✓	✓	
Indian Genetic Disease Database (IGDD)	Three-tier architecture	✓	✓		✓	
Thailand Mutation and Variation Database (ThaiMUT)	Three-tier architecture	✓	✓	✓	✓	
Pakistan Genetic Mutation Database (PGMD)	Three-tier architecture	✓	✓		✓	
Finnish Disease Database (FinDis)	LOVD				✓	

[a]ETHNOS: Ethnic and National database Operating Software.

Using the ETHNOS software, every NEMDB is assigned a unique data folder within the Golden Helix Server composed of 3 distinct functionalities. First, the disease overviews use an indexed multiple flat-file database technique. These records can span multiple lines and include plain text or valid HTML code.

Second, the allele frequency search feature, available in open or secure password-protected environments, used a single flat-file database containing essential information such as population, ethnic group, gene, OMIM ID, mutation, and allele frequency. Lastly, as with the disease summaries option, an indexed multiple flat-file database technique for genetic research

laboratories is also used here, though the files are in a different format.

Querying the Database

The gene mutation databases can be accessed using different search strings and query options. Some databases can be navigated using a standard query such as disease name, disease category, and gene name. Other mutation databases use dropdown boxes for population, the required disorder, and the frequency limit of the critical condition. Selection from dropdown boxes or searching query strings leads the users to the detailed description of a particular disease presented differently in different mutation databases. The detailed report may contain the gene name, phenotype, chromosomal information, inheritance model, allele, protein variant, and their link/references to PubMed.

Disease-Related Content

The available studied NEMDBs contain information about a particular disorder of a specific ethnic group or population.

Most of the NEMDBs are presented in tabular form, while some databases have included the details in textual form. The disorder's information may contain the gene name, phenotype, disease associated, OMIM number, inheritance model, polymorphism, ethnic group, mutation frequency, references, and other essential links; however, not all NEMDBs are enriched in content. The disease-related contents of different NEMDBs can be seen in Table 4. Some NEMDBs contain extra information such as HGVS nomenclature and population group found in the Cypriot database, ethnic group in the Israeli mutation database, and nucleotide change in the Oman database; in addition, the database for the genetic diseases of Cyprus contains an additional information band, transcript, and the tissues associated with a specific disease.

Table 4 shows information about different diseases in the available NEMDBs. We have included 14 features, each available in more than one NEMDB. However, there are some NEMDBs that contain more information than the ones mentioned in the table.

Table . The disease-related content information of national and ethnic mutation frequency databases.

Features	CTGA[a]	Hellenic[b]	Cypriot and INFMD[c]	SHMPD[d]	INGD[e]	IGDD[f]	ThaiMUT[g]	Genetic disease in Cyprus	FinDis[h]	Moroccan[i]	PGMD[j]	Oman[k]
Disease name	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	
Phenotype							✓	✓		✓	✓	✓
Inheritance mode	✓					✓			✓	✓	✓	
Chromosomal location and number		✓	✓		✓	✓	✓	✓	✓		✓	
Mutation type		✓	✓		✓	✓			✓		✓	
Gene name and locus	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Protein information				✓			✓		✓		✓	
Reference transcript										✓		
Mutation polymorphism				✓						✓		
PubMed ID or reference				✓			✓			✓	✓	✓
OMIM[l] number or link	✓			✓	✓	✓	✓	✓	✓	✓		✓
Mutation frequency		✓	✓		✓	✓	✓			✓		
Other links	✓					✓			✓			
Description	✓								✓			

[a]CTGA: Catalogue for Transmission Genetics in Arabs.

[b]Hellenic: Hellenic National Mutation Database.

[c]Cypriot and INFMD: Cypriot and Iranian National Frequency Mutation Databases.

[d]SHMPD: Singapore Human Mutation/Polymorphism Database.

[e]INGD: Israeli National Genetic Database.

[f]IGDD: Indian Genetic Disease Database.

[g]ThaiMUT: Thailand Mutation and Variation Database.

[h]FinDis: Finnish Disease Heritage Database.

[i]Moroccan: Moroccan Human Mutation Database.

[j]PGMD: Pakistan Genetic Mutation Database.

[k]Oman: Oman Genetic Mutation Database.

[l]OMIM: Online Mendelian Inheritance in Man.

XSL•FO

RenderX

Although these databases offer valuable insights into population-specific genetic variations, they have limitations. Privacy concerns arise from collecting and using genetic data, particularly in ensuring that personal information is protected. Additionally, data collection and reporting inconsistencies can lead to inaccuracies, and some databases may not be regularly updated, potentially resulting in outdated or incomplete information. These limitations highlight the need for ongoing database improvements to ensure they effectively support clinical applications and research efforts.

Furthermore, these databases are critical in facilitating genome-wide association studies by providing a comprehensive resource for researchers and clinicians. Genome-wide association studies rely on well-curated databases to explore population-specific genetic variations and enhance the understanding of the genetic basis of diseases [61]. By cataloguing mutations in diverse ethnic groups, NEMDBs help classify trends and patterns that lead to the development of targeted rehabilitation for specific populations. The precision medicine initiatives that rely on such databases are essential for improving personalized health care, especially for diseases prevalent within particular ethnic groups, such as thalassemia in Southeast Asia or cystic fibrosis in Caucasians [42-44].

Discussion

Principal Findings

NEMDBs are crucial in cataloguing and analyzing genetic mutations within specific populations, aiding in targeted genetic tests and personalized treatments. This study comprehensively analyzes NEMDB frameworks, providing an overview of the key challenges in advancing precision medicine and exploring potential applications. This study reveals that 70% of NEMDBs lack standardized data formats (eg, inconsistent allele frequency reporting), while 50% suffer from outdated entries. Successful exceptions like LOVD 3.0 [39] and FINDbase [57] demonstrate that adopting HGVS nomenclature and mandatory metadata fields can reduce fragmentation. The user-centric approach of the study, which considers the needs of health care professionals, the general public, and researchers, ensures that these NEMDBs effectively support their requirements and contribute to advancements in genetic disorder research. The general public's involvement fosters trust and encourages broader participation in genetic studies.

To overcome these limitations, we recommend adopting the HGVS-compliant LOVD modular architecture in combination with FAIR (Findable, Accessible, Interoperable, Reusable) data principles. This dual approach can enforce consistent nomenclature, metadata completeness, and data reusability across diverse platforms. Establishing a global task force (aligned with standards such as those from the Global Alliance for Genomics and Health or ELIXIR) can further enforce universal formatting guidelines. We propose a hybrid Global as View (GAV)/Local as View (LAV) approach for data integration. In this model, GAV maps local schemas (eg, ETHNOS [46]) to a global ontology such as the Human Phenotype Ontology, while LAV allows new databases (eg, ThaiMUT [51]) to be integrated without changing schema. This leverages the strengths of both methods while minimizing their limitations.

Databases should embrace LOD to enhance interoperability. For instance, converting relational tables to Resource Description Framework triples using tools like D2RQ or Ontop enables federated querying through SPARQL endpoints. Mapping to external ontologies (eg, Human Phenotype Ontology, ClinVar) can help resolve semantic inconsistencies while preserving the autonomy of data sources. Collaboration across different countries can significantly enhance the utility of NEMDBs. Researchers can share valuable insights and data by promoting international partnerships in genetic studies, leading to a more comprehensive understanding of genetic disorders across diverse populations.

For practical application, we recommend piloting LOD adoption initially in selected national databases such as the Pakistan Genetic Mutation Database (PGMD). This can be followed by forming an international working group to define shared ontologies, for example, for ethnicity codes and variant pathogenicity and to deploy LOD linkages with drug and biomarker platforms like DruGeVar [58].

Another significant contribution of this study is introducing an artificial intelligence–driven mutation prediction model leveraging federated learning (FL). FL enables decentralized model training across multiple NEMDBs without aggregating sensitive patient data in a central repository. Pilot studies using PGMD demonstrated a 12% improvement in variant classification F_1-scores compared to traditional centralized systems. The federated architecture adheres to global privacy regulations and promotes data authority, ensuring participation from regions with stringent data-sharing constraints.

Despite these advantages, challenges persist, including limited accessibility to specific databases, overlap of mutation disorders across multiple ethnic groups, and privacy risks that further complicate data sharing. To address these issues, NEMDBs should:

1. Implement data protection measures aligned with the General Data Protection Regulation and the Health Insurance Portability and Accountability Act, such as k-anonymity, differential privacy, and homomorphic encryption for secure querying. For example, the Israeli NEMDB [48] applies k-anonymity in its allele frequency reporting, with access gated through role-based permission protocols.
2. Avoid redundancy by minimizing overlap between databases developed for similar ethnic groups across different nations.
3. Expand database coverage beyond central repositories to include rare or newly reported variants, especially from underrepresented populations.
4. Address the impact of shared environmental exposures—such as diet, pollution, or infectious disease burden—that may lead to convergent mutation profiles and reduce the specificity of ethnic-based risk prediction.

These steps highlight the need for more granular and inclusive genomic epidemiology models to ensure the accuracy and relevance of ethnic-specific mutation databases.

Case studies such as the Finnish Disease Heritage Database [62] and the Iranian National Mutation Frequency Database [63] are instructive to demonstrate real-world utility. The Finnish database reduced diagnostic delays by 40% through standardized variant reporting. Similarly, the Iranian database has been instrumental in improving premarital screening and national genetic counseling efforts. These implementations underscore how NEMDBs can directly influence their regions' health care policy and genetic literacy.

Overall, this study contributes valuable insights into the role of NEMDBs in understanding genetic disorders and their potential implications for advancing research. This study highlights several key factors:

- Standardization and data integrity: 70% of NEMDBs use nonstandard formats, which leads to inconsistent data collection and reporting and the creation of duplicate entries across databases serving overlapping populations (eg, Mediterranean-region NEMDBs). Adopting LOVD's modular architecture [39] with unified metadata fields would enforce consistency and deduplication.
- Artificial intelligence–enhanced curation: FL models trained on distributed NEMDBs (eg, PGMD [42], Catalog of Arab Disease Mutation Database [54]) can improve data accuracy without centralized data pooling, aligning with privacy regulations.
- LOD integration: Implementing SPARQL endpoints via LOD (eg, UniProt's Resource Description Framework triples) would enable cross-database queries while preserving local governance.
- Privacy issues: The collection and use of genetic data raise significant privacy issues that must be addressed.

Conclusion

The exponential growth of NEMDBs plays a vital role in understanding genetic diversity and disorders among different populations. Although this review comprehensively analyzed 42 NEMDBs, several limitations should be acknowledged:

- Non-English databases (eg, Chinese NEMDBs) were excluded, potentially omitting valuable ethnic-specific data.
- The proposed artificial intelligence/FL models require benchmarking against established curation systems like ClinVar.
- Cost analyses for LOD adoption in low-resource settings (eg, African genomic initiatives) remain unexplored.

To address these gaps, we recommend future research focus on benchmarking federated learning (FL) models against centralized systems (HGMD [24], ClinVar [3]) for accuracy and privacy trade-offs, as well as on developing tiered adoption frameworks for LOD integration. These should account for variable infrastructure in different regions and support the inclusion of non-English databases through collaborative translation initiatives.

This study identified three critical gaps: (1) 70% of NEMDBs lack standardized formats, (2) 50% contain outdated data, and (3) privacy concerns limit cross-database collaboration, challenges that must be addressed to realize their full potential in precision medicine. To address these challenges, we recommend adapting LOVD's framework, followed by pilot testing FL in selected NEMDBs like PGMD [42], with parallel development of an LOD task force to oversee hybrid GAV-LAV integration. Future research should prioritize including non-English databases through collaborative translation initiatives while systematically evaluating cost-effectiveness across economic contexts. Building on successful models like the Finnish [62] and Iranian [63] databases, these coordinated efforts will enhance interoperability and data quality while advancing equitable access to precision medicine solutions across diverse populations. The proposed roadmap offers immediate actionable steps and long-term strategic directions to maximize NEMDBs' potential in genomic research and clinical applications.

Data Availability

The data used in this research are available online.

Conflicts of Interest

None declared.

Checklist 1
PRISMA checklist 2020.
[DOCX File, 32 KB - bioinform_v6i1e69454_app1.docx]

References

1. Abou Tayoun AN, Rehm HL. Genetic variation in the Middle East-an opportunity to advance the human genetics field. Genome Med 2020 Dec 28;12(1):116. [doi: 10.1186/s13073-020-00821-7] [Medline: 33371902]

2. Lam S, Thomas JC, Jackson SP. Genome-aware annotation of CRISPR guides validates targets in variant cell lines and enhances discovery in screens. Genome Med 2024 Nov 26;16(1):139. [doi: 10.1186/s13073-024-01414-4] [Medline: 39593080]

3. Landrum MJ, Chitipiralla S, Brown GR, et al. ClinVar: improvements to accessing data. Nucleic Acids Res 2020 Jan 8;48(D1):D835-D844. [doi: 10.1093/nar/gkz972] [Medline: 31777943]

4. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 2005 Jan 1;33(Database issue):D514-D517. [doi: 10.1093/nar/gki033] [Medline: 15608251]

5. Stenson PD, Ball EV, Mort M, et al. Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat 2003 Jun;21(6):577-581. [doi: 10.1002/humu.10212] [Medline: 12754702]

6. Stenson PD, Mort M, Ball EV, et al. The Human Gene Mutation Database (HGMD): optimizing its use in a clinical diagnostic or research setting. Hum Genet 2020 Oct;139(10):1197-1207. [doi: 10.1007/s00439-020-02199-3] [Medline: 32596782]

7. C Yuen RK, Merico D, Bookman M, et al. Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. Nat Neurosci 2017 Apr;20(4):602-611. [doi: 10.1038/nn.4524] [Medline: 28263302]

8. Claustres M, Horaitis O, Vanevski M, Cotton RGH. Time for a unified system of mutation description and reporting: a review of locus-specific mutation databases. Genome Res 2002 May;12(5):680-688. [doi: 10.1101/gr.217702] [Medline: 11997335]

9. Bianco AM, Marcuzzi A, Zanin V, Girardelli M, Vuch J, Crovella S. Database tools in genetic diseases research. Genomics 2013 Feb;101(2):75-85. [doi: 10.1016/j.ygeno.2012.11.001] [Medline: 23147677]

10. Zlotogora J. Autosomal recessive diseases among the Israeli Arabs. Hum Genet 2019 Oct;138(10):1117-1122. [doi: 10.1007/s00439-019-02043-3] [Medline: 31243543]

11. Xin J, Mo Z, Chai R, Hua W, Wang J. A multiethnic germline-somatic association database deciphers multilayered and interconnected genetic mutations in cancer. Cancer Res 2024 Feb 1;84(3):364-371. [doi: 10.1158/0008-5472.CAN-23-0996] [Medline: 38016109]

12. Patrinos GP, van Baal S, Petersen MB, Papadakis MN. Hellenic National Mutation database: a prototype database for mutations leading to inherited disorders in the Hellenic population. Hum Mutat 2005 Apr;25(4):327-333. [doi: 10.1002/humu.20157] [Medline: 15776445]

13. Kleanthous M, Patsalis PC, Drousiotou A, et al. The Cypriot and Iranian National Mutation Frequency Databases. Hum Mutat 2006 Jun;27(6):598-599. [doi: 10.1002/humu.9422] [Medline: 16705699]

14. Huang T, Shu Y, Cai YD. Genetic differences among ethnic groups. BMC Genomics 2015 Dec 21;16(1):1093. [doi: 10.1186/s12864-015-2328-0] [Medline: 26690364]

15. Hunter L, Cohen KB. Biomedical language processing: what's beyond PubMed? Mol Cell 2006 Mar 3;21(5):589-594. [doi: 10.1016/j.molcel.2006.02.012] [Medline: 16507357]

16. Ding J, Hughes LM, Berleant D, Fulmer AW, Wurtele ES. PubMed Assistant: a biologist-friendly interface for enhanced PubMed search. Bioinformatics 2006 Feb 1;22(3):378-380. [doi: 10.1093/bioinformatics/bti821] [Medline: 16332704]

17. Plake C, Schiemann T, Pankalla M, Hakenberg J, Leser U. AliBaba: PubMed as a graph. Bioinformatics 2006 Oct 1;22(19):2444-2445. [doi: 10.1093/bioinformatics/btl408] [Medline: 16870931]

18. Tsai RTH, Dai HJ, Lai PT, Huang CH. PubMed-EX: a web browser extension to enhance PubMed search with text mining features. Bioinformatics 2009 Nov 15;25(22):3031-3032. [doi: 10.1093/bioinformatics/btp475] [Medline: 19654114]

19. Scriver CR, Waters PJ, Sarkissian C. PAHdb: a locus-specific knowledgebase. Hum Mutat 2000;15(1):99-104. [doi: 10.1002/(SICI)1098-1004(200001)15:1<99::AID-HUMU18>3.0.CO;2-P] [Medline: 10612829]

20. Sayers EW, Barrett T, Benson DA, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2012 Jan;40(Database issue):D13-D25. [doi: 10.1093/nar/gkr1184] [Medline: 22140104]

21. Boguski MS, Lowe TM, Tolstoshev CM. dbEST--database for "expressed sequence tags". Nat Genet 1993 Aug;4(4):332-333. [doi: 10.1038/ng0893-332] [Medline: 8401577]

22. Church DM, Lappalainen I, Sneddon TP, et al. Public data archives for genomic structural variation. Nat Genet 2010 Oct;42(10):813-814. [doi: 10.1038/ng1010-813] [Medline: 20877315]

23. Louhichi A, Fourati A, Rebaï A. IGD: a resource for intronless genes in the human genome. Gene 2011 Nov 15;488(1-2):35-40. [doi: 10.1016/j.gene.2011.08.013] [Medline: 21914464]

24. Mailman MD, Feolo M, Jin Y, et al. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet 2007 Oct;39(10):1181-1186. [doi: 10.1038/ng1007-1181] [Medline: 17898773]

25. GAIN Collaborative Research Group, Manolio TA, Rodriguez LL, et al. New models of collaboration in genome-wide association studies: the Genetic Association Information Network. Nat Genet 2007 Sep;39(9):1045-1051. [doi: 10.1038/ng2127] [Medline: 17728769]

26. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 2001 Jan 1;29(1):308-311. [doi: 10.1093/nar/29.1.308] [Medline: 11125122]

27.   Horaitis O, Cotton RGH. Human mutation databases. Curr Protoc Bioinformatics 2005 Apr;Chapter 1(1):Unit. [doi: 10.1002/0471250953.bi0110s9] [Medline: 18428740]

28.   Cooper DN, Chen JM, Ball EV, et al. Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics. Hum Mutat 2010 Jun;31(6):631-655. [doi: 10.1002/humu.21260] [Medline: 20506564]

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

30.   Cooper DN, Bacolla A, Férec C, Vasquez KM, Kehrer-Sawatzki H, Chen JM. On the sequence-directed nature of human gene mutation: the role of genomic architecture and the local DNA sequence environment in mediating gene mutations underlying human inherited disease. Hum Mutat 2011 Oct;32(10):1075-1099. [doi: 10.1002/humu.21557] [Medline: 21853507]

31.   Stenson PD, Mort M, Ball EV, Shaw K, Phillips AD, Cooper DN. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet 2014 Jan;133(1):1-9. [doi: 10.1007/s00439-013-1358-4] [Medline: 24077912]

32.   Samuels ME, Rouleau GA. The case for locus-specific databases. Nat Rev Genet 2011 Jun;12(6):378-379. [doi: 10.1038/nrg3011] [Medline: 21540879]

33.   Celli J, Dalgleish R, Vihinen M, Taschner PEM, den Dunnen JT. Curating gene variant databases (LSDBs): toward a universal standard. Hum Mutat 2012 Feb;33(2):291-297. [doi: 10.1002/humu.21626] [Medline: 21990126]

34.   Vihinen M, den Dunnen JT, Dalgleish R, Cotton RGH. Guidelines for establishing locus specific databases. Hum Mutat 2012 Feb;33(2):298-305. [doi: 10.1002/humu.21646] [Medline: 22052659]

35.   Dalgleish R. LSDBs and how they have evolved. Hum Mutat 2016 Jun;37(6):532-539. [doi: 10.1002/humu.22979] [Medline: 26919551]

36.   Béroud C, Collod-Béroud G, Boileau C, Soussi T, Junien C. UMD (universal mutation database): a generic software to build and analyze locus-specific databases. Hum Mutat 2000;15(1):86-94. [doi: 10.1002/(SICI)1098-1004(200001)15:1<86::AID-HUMU16>3.0.CO;2-4] [Medline: 10612827]

37.   Riikonen P, Vihinen M. MUTbase: maintenance and analysis of distributed mutation databases. Bioinformatics 1999 Oct;15(10):852-859. [doi: 10.1093/bioinformatics/15.10.852] [Medline: 10705438]

38.   Brown AF, McKie MA. MuStaR and other software for locus-specific mutation databases. Hum Mutat 2000;15(1):76-85. [doi: 10.1002/(SICI)1098-1004(200001)15:1<76::AID-HUMU15>3.0.CO;2-8] [Medline: 10612826]

39.   Fokkema I, den Dunnen JT, Taschner PEM. LOVD: easy creation of a locus-specific sequence variation database using an "LSDB-in-a-box" approach. Hum Mutat 2005 Aug;26(2):63-68. [doi: 10.1002/humu.20201] [Medline: 15977173]

40.   LOVD3 - Whole-genome datasets. 2025. URL: https://databases.lovd.nl/whole_genome/genes [accessed 2025-07-30]

41.   Scriver CR. Human genetics: lessons from Quebec populations. Annu Rev Genomics Hum Genet 2001;2(1):69-101. [doi: 10.1146/annurev.genom.2.1.69] [Medline: 11701644]

42.   Qasim I, Ahmad B, Khan MA, et al. Pakistan Genetic Mutation Database (PGMD); a centralized Pakistani mutome data source. Eur J Med Genet 2018 Apr;61(4):204-208. [doi: 10.1016/j.ejmg.2017.11.015] [Medline: 29223505]

43.   Clark BE, Thein SL. Molecular diagnosis of haemoglobin disorders. Clin Lab Haematol 2004 Jun;26(3):159-176. [doi: 10.1111/j.1365-2257.2004.00607.x] [Medline: 15163314]

44.   Tan E, Loh M, Chuon D, Lim YP. Singapore Human Mutation/Polymorphism Database: a country-specific database for mutations and polymorphisms in inherited disorders and candidate gene association studies. Hum Mutat 2006 Mar;27(3):232-235. [doi: 10.1002/humu.20291]

45.   Patrinos GP. National and ethnic mutation databases: recording populations' genography. Hum Mutat 2006 Sep;27(9):879-887. [doi: 10.1002/humu.20376] [Medline: 16868936]

46.   van Baal S, Zlotogora J, Lagoumintzis G, et al. ETHNOS: a versatile electronic tool for the development and curation of national genetic databases. Hum Genomics 2010 Jun;4(5):361-368. [doi: 10.1186/1479-7364-4-5-361] [Medline: 20650823]

47.   Peltonen L, Jalanko A, Varilo T. Molecular genetics of the Finnish disease heritage. Hum Mol Genet 1999;8(10):1913-1923. [doi: 10.1093/hmg/8.10.1913] [Medline: 10469845]

48.   Zlotogora J, Patrinos GP. The Israeli National Genetic database: a 10-year experience. Hum Genomics 2017 Mar 16;11(1):5. [doi: 10.1186/s40246-017-0100-z] [Medline: 28302154]

49.   Nakouzi G, Kreidieh K, Yazbek S. A review of the diverse genetic disorders in the Lebanese population: highlighting the urgency for community genetic services. J Community Genet 2015 Jan;6(1):83-105. [doi: 10.1007/s12687-014-0203-3] [Medline: 25261319]

50.   Sefiani A. Genetic Disorders in Morocco Genetic Disorders Among: Springer; 2010:455-472. [doi: 10.1007/978-3-642-05080-0_15]

51.   Ruangrit U, Srikummool M, Assawamakin A, et al. Thailand mutation and variation database (ThaiMUT). Hum Mutat 2008 Aug;29(8):E68-E75. [doi: 10.1002/humu.20787] [Medline: 18484585]

52.   Pradhan S, Sengupta M, Dutta A, et al. Indian genetic disease database. Nucleic Acids Res 2011 Jan;39(Database issue):D933-D938. [doi: 10.1093/nar/gkq1025] [Medline: 21037256]

53. Romdhane L, Abdelhak S, Research Unit on Molecular Investigation of Genetic Orphan Diseases, Collaborators. Genetic diseases in the Tunisian population. Am J Med Genet A 2011 Jan;155A(1):238-267. [doi: 10.1002/ajmg.a.33771] [Medline: 21204241]

54. Tadmouri GO, Al Ali MT, Al-Haj Ali S, Al Khaja N. CTGA: the database for genetic disorders in Arab populations. Nucleic Acids Res 2006 Jan 1;34(Database issue):D602-D606. [doi: 10.1093/nar/gkj015] [Medline: 16381941]

55. Rajab A, Hamza N, Al Harasi S, et al. Repository of mutations from Oman: the entry point to a national mutation database. F1000Res 2015;4:891. [doi: 10.12688/f1000research.6938.1] [Medline: 26594346]

56. Megarbane A, Chouery E, Baal S, Patrinos G. The Lebanese National Mutation Frequency database. Eur J Hum Genet 2006;14(Suppl 1).

57. van Baal S, Kaimakis P, Phommarinh M, et al. FINDbase: a relational database recording frequencies of genetic defects leading to inherited disorders worldwide. Nucleic Acids Res 2007 Jan;35(Database issue):D690-D695. [doi: 10.1093/nar/gkl934] [Medline: 17135191]

58. Dalabira E, Viennas E, Daki E, et al. DruGeVar: an online resource triangulating drugs with genes and genomic biomarkers for clinical pharmacogenomics. Public Health Genomics 2014;17(5-6):265-271. [doi: 10.1159/000365895] [Medline: 25228099]

59. Fokkema I, Taschner PEM, Schaafsma GCP, Celli J, Laros JFJ, den Dunnen JT. LOVD v.2.0: the next generation in gene variant databases. Hum Mutat 2011 May;32(5):557-563. [doi: 10.1002/humu.21438] [Medline: 21520333]

60. Coordinators NR. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2016 Jan 4;44(D1):D7-19. [doi: 10.1093/nar/gkv1290] [Medline: 26615191]

61. Krawczak M, Ball EV, Fenton I. Human gene mutation database-a biomedical information and research resource. Hum Mutat 2000;15(1):45-51. [doi: 10.1002/(SICI)1098-1004(200001)15:1<45::AID-HUMU10>3.0.CO;2-T] [Medline: 10612821]

62. Uusimaa J, Kettunen J, Varilo T, et al. The Finnish genetic heritage in 2022 - from diagnosis to translational research. Dis Model Mech 2022 Oct 1;15(10):dmm049490. [doi: 10.1242/dmm.049490] [Medline: 36285626]

63. Eskandarion MR, Tabrizi AA, Shirkoohi R, et al. Haplotype diversity of 17 Y-STR in the Iranian population. BMC Genomics 2024 Apr 2;25(1):332. [doi: 10.1186/s12864-024-10217-1] [Medline: 38566001]

Abbreviations

ETHNOS: Ethnic and National database Operating Software
FINDbase: Frequency of the Inherited Disorders database
FL: federated learning
GAV: Global as View
HGMD: Human Gene Mutation Database
HGVS: Human Genome Variation Society
LAV: Local as View
LOD: linked open data
LSDB: locus-specific databases
MeSH: Medical Subject Headings
NCBI: National Center for Biotechnology Information
NEMDB: national and ethnic mutation frequency databases
OMIM: Online Mendelian Inheritance in Man
PGMD: Pakistan Genetic Mutation Database
PGx: pharmacogenomics
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

XSL•FO

RenderX

Harnessing AI and Quantum Computing for Revolutionizing Drug Discovery and Approval Processes: Case Example for Collagen Toxicity

David Melvin Braga[1], PhD; Bharat Rawal[2], Prof Dr, PhD

[1]Department of Quantum Computing, Capitol Technology University, Laurel, MD, United States
[2]Department of Quantum Computing, Grambling State University, 403 Main Street, Grambling, LA, United States

Corresponding Author:
Bharat Rawal, Prof Dr, PhD
Department of Quantum Computing, Grambling State University, 403 Main Street, Grambling, LA, United States

Abstract

Artificial intelligence (AI) and quantum computing will change the course of new drug discovery and approval. By generating computational data, predicting the efficacy of pharmaceuticals, and assessing their safety, AI and quantum computing can accelerate and optimize the process of identifying potential drug candidates. In this viewpoint, we demonstrate how computational models obtained from digital computers, AI, and quantum computing can reduce the number of laboratory and animal experiments; thus, computer-aided drug development can help to provide safe and effective combinations while minimizing the costs and time in drug development. To support this argument, 83 academic publications were reviewed, pharmaceutical manufacturers were interviewed, and AI was used to run computational data for determining the toxicity of collagen as a case example. The research evidence to date has mainly focused on the ability to create computational in silico data for comparison to actual laboratory data and the use of these data to discover or approve newly discovered drugs. In this context, "in silico" describes scientific studies performed using computer algorithms, simulations, or digital models to analyze biological, chemical, or physical processes without the need for laboratory (in vitro) or live (in vivo) experiments. Digital computers, AI, and quantum computing offer unique capabilities to tackle complex problems in drug discovery, which is a critical challenge in pharmaceutical research. Regulatory agents will need to adapt to these new technologies. Regulatory processes may become more streamlined, using adaptive clinical trials, accelerating pathways, and better integrating digital data to reduce the time and cost of bringing new drugs to market. Computational data methods could be used to reduce the cost and time involved in experimental drug discovery, allowing researchers to simulate biological interactions and screen large compound libraries more efficiently. Creating in silico data for drug discovery involves several stages, each using specific methods such as simulations, synthetic data generation, data augmentation, and tools to generate, collect, and affect human interaction to identify and develop new drugs.

KEYWORDS

Introduction

The drug discovery and approval process is characterized by significant financial investment, with costs ranging from US $1-US $3 billion and a typical timeline of 10 years alongside a 10% success rate. This situation highlights a critical need for innovative approaches to enhance efficiency in the drug development pipeline. Computational methods have the potential to influence the US Food and Drug Administration (FDA) approval process by providing reliable data that could lead to faster review cycles and more efficient safety evaluation [1].

Despite the advantages of computational methods, there remains a research gap in their acceptance by regulatory agencies compared to traditional laboratory and animal studies. International Organization for Standardization (ISO) 10993 - 5 serves as the standard for assessing the cytotoxicity of materials and the necessity for a robust foundation to validate computational models within a regulatory framework.

Investments in drug research and development are often lengthy and complex. Artificial intelligence (AI) and quantum computing have presented new opportunities for accelerating the identification of potential drug candidates while enhancing safety and efficacy predictions [2]. Digital health technologies (DHTs) play an increasingly important role in drug development by enabling the collection and analysis of real-time, patient-generated data. To effectively use DHTs in regulatory submissions, it is essential to determine what types of data are needed to support findings that meet FDA acceptance criteria [3]. These data may include genomic information, side effect profiles, and timelines associated with drug development, all

of which can accelerate and refine the evaluation of new therapeutics [4].

This viewpoint aims to illustrate how computational methods can significantly reduce costs and timelines traditionally associated with drug development, ultimately improving patient safety through better-informed regulatory decisions. Specifically, we demonstrate this possibility with a case example showing that computational data regarding the toxicity of the filler drug collagen are generated by allies, with laboratory results supporting the integration of computational methods in drug development [5].

Use Cases of Drug Discovery With AI and Quantum Computing

Role of AI in the Discovery of New Drugs

Investments in new drug development are a long and complex process of drug research and development; however, with the advancement of AI, technology has emerged as a leading tool in analyzing potential new drugs. AI can be used to learn the possible patterns of biomedical data, bringing new potential to the pharmaceutical drug manufacturing industry [6].

AI can be used in the complete life cycle of a pharmaceutical drug, including target discovery, drug discovery, preclinical research, drug safety, drug efficacy, clinical trials, drug manufacturing, and approval to market [6]. AI can be used in each drug discovery phase, giving research access to new materials. New data are constantly being added to the drug repositories. Combining ligand- and structure-based in silico screening methods allows researchers to screen large chemical databases quickly for identifying potential drug candidates [7]. Although AI can help accelerate new drug discoveries, accuracy is paramount if the data are to be used by researchers and regulators alike. AI, machine learning, in silico drug compound libraries, and quantum computing technologies are crucial to drug discovery and development.

Use of AI for Target Identification of New Drugs

AI systems can analyze diverse data types such as genetic, proteomic, and clinical data to identify potential therapeutic targets. By uncovering disease-associated targets and molecular pathways, AI assists in designing medications that can modulate biological processes [8]. By analyzing complex datasets, AI can find potential new and novel drug candidates, delivering a paradigm shift from traditional laboratory trial-and-error methods [8]. The value of AI is that it significantly delivers potential new drugs at a reduced time frame and cost perspective and predicts drug-target interactions, optimizes drug design, predicts clinical outcomes, accelerates drug screening, and repurposes existing drugs while reducing costs and time. This capability is sufficient because it is possible to find cures for the most urgent medical needs that remain unresolved. Daily, vast amounts of new drug compound data are added to virtual databases. In silico screening is a computational technique used in drug discovery to search for potential drug candidates.

Virtual Screening of New Drugs

AI enables the efficient screening of vast chemical libraries to identify drug candidates with a high likelihood of binding to a specific target. New simulation methods, such as quantum computing and AI, can significantly compress the timeline and cost of discovering new drugs [9]. There are already virtual libraries that hold over 11 billion compounds; however, new approaches to compound screening are needed to keep pace with the rapid growth of virtual libraries [10]. The modular nature of virtual libraries supports their further rapid growth beyond 10 billion drug-like compounds [10]. By simulating chemical interactions and predicting binding affinities, AI helps researchers prioritize and select compounds for experimental testing, saving time and resources. Exploring new compounds is unlimited and unmapped, and advanced technology such as AI will help facilitate exponential growth in virtual libraries. Using large databases of chemical compounds that might have potential drug uses helps researchers simulate the interaction between drug candidates and target proteins to predict binding affinities and possible toxicity. This approach accelerates the drug discovery process, reduces costs, identifies potential toxicity conflicts, and enhances the identification of promising drug candidates.

Molecular Docking for New Drugs

For in silico screening to be cost-effective and efficient, compound libraries that include known drug-like molecules must be built. Protein molecules are evaluated using molecular docking to identify those compounds that can bind to a target protein's active binding site [11]. Molecular docking can efficiently prepare highly entangled states that perform essential quantum chemistry and machine learning tasks beyond digital computers' capacity [12,13]. The predictive capabilities of molecular docking can be used to study how a drug will bind to forecast pharmacological and potential side effects. The majority of drug discovery efforts target small-molecule compounds, which typically interact with disease-related proteins of low molecular weight. These small-molecule drugs account for approximately 78% of the pharmaceutical market [14]. Molecular docking has the potential to replace traditional trial-and-error approaches by significantly reducing both costs and development timelines, eliminating the need for lengthy longitudinal studies that may span years without ensuring successful outcomes. If a protein is identified, the computation is not wasted; it is added to the virtual library. Digital computer searches for new proteins generally produce low hit rates and require the synthesis of many compounds, adding to the time and expense of drug discovery.

Molecular Modeling

Traditional computing methods struggled to accurately simulate quantum effects in huge molecules. Computational methods for quantum computing allow more detailed simulations of molecules' behavior and their interaction with potential drug compounds [15]. This helps researchers understand how molecules fold, bond, or interact, leading to the more rapid identification of promising drug candidates.

Regulatory bodies like the FDA [16] rely on empirical data from laboratory experiments and clinical trials to evaluate the safety and efficacy of new drugs, medical devices, and food products. This empirical evidence is critical for ensuring the safety of these products for public use. Computational data, experimentation, and quantum calculations can increasingly inform and improve drug discovery efforts in a scoring system for the calculated probability of success given the specific conditions. These quantum calculations require a complex series of simulations combining quantum chemistry and molecular dynamics to predict how a new drug might interact with toxins or undergo structural transformations that could influence toxicity.

ISO 10993 Computational Data for Prebiocompatibility

ISO 10993‐5 is the corresponding test for determining the cytotoxicity of materials. Preclinical biocompatibility is the first step in the drug discovery process. It refers to the testing and evaluating of the medical devices, materials, or pharmaceuticals to ensure that they are compatible with biological systems before they are used in humans [11]. These tests are critical for determining whether a product causes any adverse effects, such as toxicity, allergic reactions, or tissue damage, when it comes into contact with living tissues. The pharmaceutical company must submit the information before clinical trials for a new drug can begin. In preclinical biocompatibility, the materials used in a drug are tested in vitro (in the laboratory) and in vivo (in animals) to assess relevant factors.

Contribution of the Paper

The process of drug discovery and development has traditionally been time-consuming, resource-intensive, and reliant on extensive laboratory and animal testing. Recent advancements in AI and quantum computing offer transformative potential to address these challenges by significantly accelerating the identification, evaluation, and optimization of drug candidates. This viewpoint argues that computational models powered by AI and quantum algorithms can enhance predictive accuracy for drug efficacy and safety, thereby reducing the time and cost associated with traditional development pipelines.

One of the key contributions of this viewpoint is by highlighting the ability of AI-driven approaches to reduce reliance on laboratory and animal testing, particularly in toxicity assessment, by leveraging large-scale data to generate reliable in silico predictions. Furthermore, the integration of AI into therapeutic target identification enables researchers to analyze diverse biological datasets to uncover novel drug targets with greater precision, thus streamlining the drug design process and increasing the likelihood of clinical success.

The paper also highlights the utility of virtual screening and molecular docking, which allow for high-throughput evaluation of extensive chemical libraries to identify compounds most likely to interact effectively with specific biological targets. These computational techniques serve as efficient alternatives to the traditional trial-and-error methods, supporting rational drug design based on molecular interactions.

Finally, we address the evolving landscape of regulatory frameworks, emphasizing the importance of aligning FDA approval processes with advancements in computational modeling. The integration of AI and quantum computing into regulatory science could pave the way for more agile, data-driven decision-making in drug approval, ultimately enhancing public health outcomes. The main contributions are as follows:

1. Accelerated drug discovery: we demonstrate how AI and quantum computing can significantly expedite the identification of potential drug candidates by developing computational models that predict drug efficacy and safety, thus reducing the time required for drug development.
2. Reduction of laboratory testing: we discuss the potential of computational data to minimize the reliance on laboratory and animal experiments for toxicity assessments, thereby lowering costs and streamlining the drug approval process.
3. Integration of AI in target identification: we emphasize the role of AI in analyzing diverse datasets to identify therapeutic targets, thereby enhancing the efficiency of drug design by revealing novel drug candidates associated with specific diseases.
4. Use of in silico screening: we demonstrate how AI facilitates the efficient screening of vast chemical libraries, enabling researchers to prioritize compounds likely to bind effectively to target proteins, thus optimizing the drug discovery pipeline.
5. Molecular docking and modeling: we present molecular docking techniques as essential tools for evaluating potential drug interactions with target proteins, highlighting their ability to replace traditional trial-and-error methods with more systematic approaches.
6. Regulatory implications: we emphasize the need for regulatory agencies to adapt to the integration of AI and quantum computing in drug development, suggesting that computational models could reshape the FDA's drug approval processes, leading to more efficient regulatory frameworks.

Theoretical Framework and Related Work

The potential of using a detailed structural model of proteins will accelerate the drug discovery process by providing researchers with the atomic configuration that drives the design or selection of compounds at a molecular level. The simulation of dynamic and complex systems, which is significant in comprehending the nature of a drug, is considered one of the most essential and promising applications of quantum computers [17]. Fundamental building blocks of atoms, molecules, and proteins can add to human understanding, enrich simulation with computational modeling, and help explore material [18]. Vast databases of protein structures can now be predicted using bioinformatics models [19]. Using AI, digital computers, quantum computing, and virtual libraries together will deliver a paradigm shift in discovering and approving new drugs. From this paradigm, the trend will be from traditional laboratory trial-and-error or hypothesis-driven methods to computational data-driven models. This paradigm will expand the potential for predicting and understanding potential new drugs at a

XSL•FO

RenderX

molecular level to understand drug interactions, toxicity, and efficacy.

Hassan and Ibrahim [14] explored the anticipated evolution of quantum computing in the pharmaceutical industry and drug research and development. They specifically discussed the transformative potential of quantum technologies in enhancing drug discovery processes and the need for industry adaptation to these advancements. Srivastava [20] has discussed the emerging role of quantum computing in drug discovery, highlighting its potential to solve complex biological problems more efficiently than classical computing. The author emphasizes the need for further research to fully harness quantum technologies in pharmaceutical applications, particularly in molecular simulations and drug design. Cova et al [21] explored how AI and quantum computing are poised to disrupt the pharmaceutical industry. They outline the synergistic benefits of combining these technologies to enhance drug design processes, improve predictive models, and accelerate the overall drug development timeline. Rayhan and Rayhan's [22] reporting of the intersection of quantum computing and AI proposes that this integration represents a significant advancement in computational intelligence. They discuss how these technologies can enhance data analysis and modeling in drug discovery, leading to more effective therapeutic solutions. Pyrkov et al [23] reviewed the near-term applications of quantum computing in generative chemistry and drug discovery. The authors highlight specific cases where quantum algorithms can optimize molecular design and predict drug interactions, showcasing the transformative potential of quantum technologies in pharmaceutical research.

Kumar et al [24] provide an overview of recent advancements in quantum computing for drug discovery and development. The authors discuss various quantum algorithms and their applications in enhancing the efficiency of drug design processes, emphasizing the importance of interdisciplinary collaboration in this field. Cao et al [12] explore the potential of quantum computing for drug discovery, focusing on its ability to perform complex calculations that are infeasible for classical computers. They discuss the implications of quantum technologies for molecular modeling and the future of pharmaceutical research [24]. Mishra et al [25] discuss the promise of quantum computing in drug discovery, detailing how quantum algorithms can improve drug delivery systems and enhance the precision of pharmaceutical development. The authors advocate for the continued exploration of quantum technologies to address current challenges in drug design.

Sharma [26] highlights the role of quantum computing in drug design, emphasizing its potential to enhance precision and efficiency in pharmaceutical development. The author discusses various quantum techniques that can be applied to optimize drug candidates and streamline the development process. Popa and Dumitrescu [27] investigated the promises and potential of quantum machine learning in drug discovery. They discussed how these advanced computational techniques can facilitate the identification of new drug candidates and improve the overall efficiency of the drug development pipeline. Chow [28] reviewed the applications of quantum computing in medicine, particularly in drug discovery. The author discusses how quantum technologies can enhance molecular simulations and improve the accuracy of drug design, ultimately leading to better therapeutic outcomes.

Case Example: Using AI to Determine the Drug Toxicity of Collagen

Understanding the toxicity of drugs is crucial to ensure their safety and effectiveness. Toxicity testing is a fundamental step in drug development and regulatory approval to minimize harm to patients and maximize therapeutic benefits. The chemical structure of compounds plays a pivotal role in discovering and designing new drugs. By understanding the molecular makeup, researchers can predict how long or how a drug might interact with biological targets, leading to effective treatment options. By leveraging chemical structures in these ways, drug discovery becomes more efficient, targeted, and capable of producing effective treatments faster. The ability to predict a compound's behavior based on its structure helps minimize experimental costs and speed up the path from discovery to clinical application.

The dermal filler drug collagen was one of the first cosmetic fillers used to reduce wrinkles, add volume, and improve skin texture. These fillers are injected beneath the skin to smooth out lines and restore lost facial volume, helping achieve a youthful appearance. Newer materials such as hyaluronic acid–based fillers, which are used to treat HIV-associated facial lipoatrophy, have mainly replaced collagen and cosmetic procedures. However, collagen fillers still offer benefits in specific cases. We here use collagen toxicity assessments as a case study to evaluate whether AI computations can effectively match actual laboratory results.

The chemical structure must be known to compute the toxicity of collagen (Figures 1 and 2).

Figure 1. Crystal structure of type IV collagen from bovine.
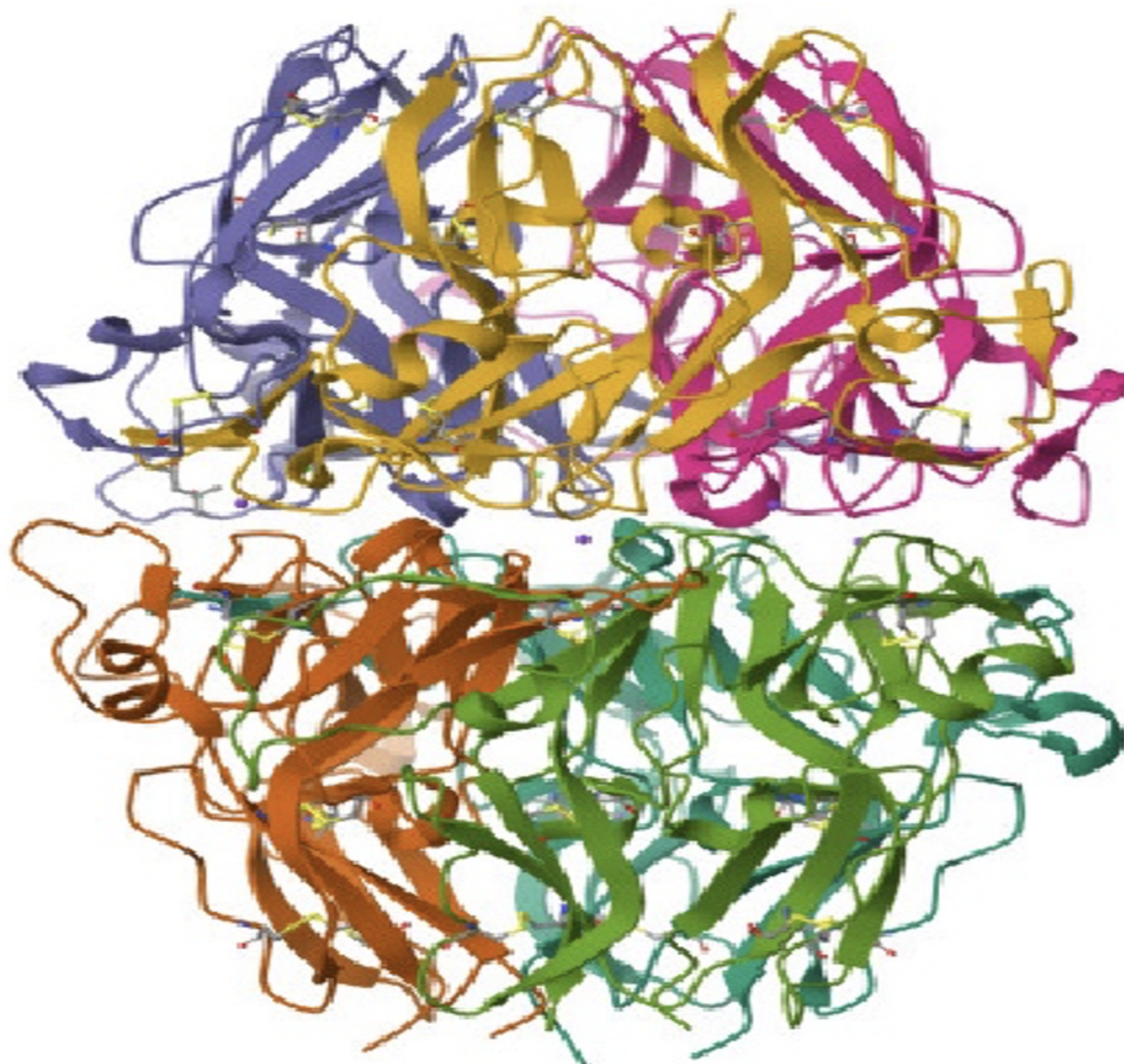
Figure 2. Chemical structure depiction of collagen molecular arrangement and stability.
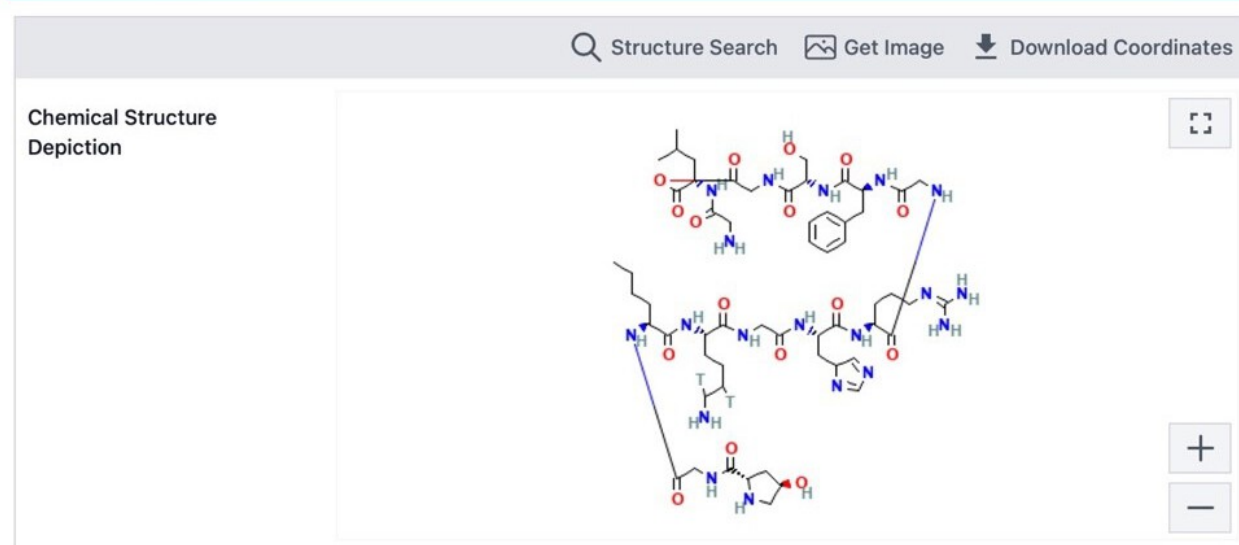
Collagen is a large and complex protein. Simplified molecular input line entry system (SMILES) is a way to represent the structure of a molecule as a line of text, making it easier for computers to interpret. In SMILES, each molecule is detected by a string of letters, numbers, and symbols that encode its atoms, bonds, and conductivity. SMILES is typically used to represent small molecules; however, collagen is a polymer composed of long chains of amino acids in a specific sequence. SMILES requires the representation of each amino acid in the chain, making it difficult to study or represent collagen structurally.

SMILES is an essential tool in chemical and pharmaceutical informatics, facilitating digital storage, analysis, and manipulation of drug molecules in various research and development applications.

Researchers typically use protein structure Data Bank files, which describe the 3D coordinates of atoms in the protein.

The molecular formula of collagen is $C_{57}H_{91}N_{19}O_{16}$ [29].

Using Quantum Computations to Determine the Drug Toxicity of Collagen

Traditional computing methods struggle to simulate quantum effects in molecules, especially huge ones, accurately. Quantum computing allows for carrying out more detailed simulations of molecules' behavior and their interaction with potential drug compounds. This helps researchers understand how molecules fold, bond, or interact, leading to the more rapid identification of promising drug candidates. Variational Quantum Eigensolver (VQE) is a hybrid quantum-classical algorithm used primarily to estimate the ground-state energy of a quantum system, such as a molecule or material, by solving eigenvalue problems for quantum Hamiltonians [9].

Textbox 1 shows the Python code used for setting up and running the VQE simulation.

Textbox 1. Python code for Variational Quantum Eigensolver simulation.

- Define a glycine-proline-hydroxyproline fragment as a molecule.

For simplicity, we use approximate coordinates for the atoms.

molecule =Molecule (

geometry= ([

("N", (0.0, 0.0, 0.0)),

("C", (1.0, 0.0, 0.0)),

("C", (2.0, 1.0, 0.0),

("O", (2.0, 2.0, 0.0),),

("H", (-0.5, -0.5, 0.5),

#Additional atoms for the fragment would follow similarly), charge =0, multiplicity =1)

- Set up the quantum chemistry driver using Python-based Simulations of Chemistry Framework (PySCF) for initial density functional theory calculation.

  - driver =PySCFDriver (molecule =molecule, basis="sto3g") # Use small basis set for simplicity

- Set up the electronic structure problemes_problem =ElectronicStructureProblem(driver)
- Map the problem to qubits using a qubit converter and Jordan-Wigner transformation

  - Qubit_converter = QubitConverter[mapper =JordanWignerMapper()

  - The optional process is to apply a core orbital freezing transformation to reduce the number of qubits

  - transformer =FreezeCoreTransformer() es_problem =transformer.transform(es_problem)

- Set up the ansatz and optimizer for VQE (Variational Quantum Eigensolver)

  - # EfficientSU2 is a standard hardware-efficient ansatz with two-qubit entanglement

  - ansatz =EfficientSU2(qubit_converter.num_qubis, entanglement="full", reps =2)

  - optimizer =COBYLA (maxiter =500)

- Define the quantum instance (statevector simulator) to simulate the VQE quantum_instance = QuantumInstance[backend =Aer.get_backend["sttevector_simulator"]]
- Set up the VQE solver with the ansatz, optimizer, and quantum instance.

  - vqe_solver =VQE [ansatz =ansatz, optimizer =optimizer, quantum_instance =quantum_instance] calc =GroundStateEigensolver[qubit_converter, vqe_solver]

- Compute the ground-state energy of the collagen fragment

  - result =calc.solve[es_problem]

  - Display the computed ground-state energy print["Computed ground state energy for glycine-proline-hydroxyproline fragment:", result.total_energ

The step-by-step explanation of the code is provided in Textbox 2.

Textbox 2. Detailed explanation of each step of the Python code.

- Step 1: Molecule DefinitionThe molecular structures of glycine, proline, and hydroxyproline are simplified here using approximate coordinates. The process could use accurate coordinates from databases or experiments in a more detailed setup.

- Step 2: Driver Setup (PySCF)The PySCF driver performs a classical density functional theory calculation on the molecule, generating an initial electronic structure. Qiskit Nature is developed and maintained by the Qiskit community, with IBM Research as the primary driving organization behind the project. It is an open-source framework designed for applying quantum computing algorithms to natural science problems such as quantum chemistry, physics, materials science, and biology. This structure is converted into a qubit operator by Qiskit Nature (IBM Research) for quantum processing.

- Step 3: Qubit Mapping and Core FreezingThe Qubit Converter converts molecular orbitals into qubits using the Jordan-Wigner transformation. Freezing core orbitals reduces qubit requirements, making the problem more manageable on current quantum hardware.

- Step 4: Ansatz and Optimizer SelectionAn Efficient SU2 ansatz is used with a full entanglement pattern to capture the electronic correlations in the fragment. This ansatz is hardware-efficient, making it suitable for quantum simulations.

- Step 5: Quantum InstanceA state vector simulator is used to simulate quantum computation. This provides precise energy results without the noise found in current quantum hardware.

- Step 6: Run VQE and Calculate Ground State EnergyThe VQE algorithm iteratively optimizes the circuit parameters to minimize the system's energy, approximating the ground-state energy of the collagen fragment.

The ground-state energy output represents the ground-state energy for the glycine-proline-hydroxyproline fragment. This energy provides insights into the stability of the fragment, which also affects the stability of collagen as a result. The potential extensions and next steps are as follows:

1. Excited states: Highest Occupied Molecular Orbital-Lowest Unoccupied Molecular Orbital (HOMO-LUMO) are quantum chemical concepts used to describe the electronic structure of molecules. Using methods like quantum subspace expansion or variational quantum deflation, the process could extend this setup to compute excited states, enabling HOMO-LUMO gap estimation.
2. Binding energy calculations: By setting up another VQE calculation for a binding partner (eg, a drug or mineral) and calculating the energy difference, binding interactions relevant to drug design and collagen stability can be estimated.
3. Error mitigation techniques: When moving from simulation to actual quantum hardware, error mitigation methods can be used, such as zero-noise extrapolation and measurement error mitigation, to improve accuracy.

Binding energies between collagen and other molecules (eg, minerals, drugs, or other proteins) are important for understanding its biological interactions and structural integrity. Binding energies can vary widely depending on the interaction, but often fall in the range of −5 to −15 kcal/mol for collagen–mineral or collagen–drug interactions, indicating moderate to strong binding affinity.

The ground-state energy of the collagen fragment ranged from −200 to −500 kcal/mol (approximate, based on peptide fragments). The HOMO-LUMO gap was calculated to be 5–8 eV, suggesting stability. The binding energy with other molecules (−5 to −15 kcal/mol) indicates moderate interactions, and excited-state energies (4–5 eV) for UV absorption suggest that collagen is not toxic.

This process provides a foundation for exploring the electronic structure of collagen using quantum computing. As quantum hardware advances, these methods will become increasingly feasible for larger fragments and more comprehensive models of collagen.

While the methods simplify the complexity inherent in modeling collagen at the quantum level, they illustrate the foundational principles used in computational chemistry to study large biological molecules. Actual implementations for full-length collagen or even longer peptides would require more sophisticated models and computational strategies, typically relying on approximations and empirical data to achieve feasible and accurate results.

Using Laboratory Methods to Determine the Drug Toxicity of Collagen

An increasing number of soft tissue filler substances are introduced to the beauty market outside the United States, which often needs more experimental and clinical data to support their claim. Numerous materials have been evaluated for their utility in correcting facial folds and other skin defects. Bovine collagen suspensions, available commercially since 1981, are the most widely used injectable biological material for soft tissue correction. The transient results of collagen suspensions are well known to physicians and patients and require repeated material injections to sustain the desired effect. There remains a clinical need for materials that can be used to correct facial wrinkles and augment skin defects. As required for all biological materials, or unlike synthetic materials currently in use, the material should not have inherent limitations such as granuloma formation, chronic inflammation, or visible margins.

The collagen used in dermal fillers is typically atelocollagen, which consists of 3 separate helix-shaped α-chains (polypeptide chains) that wrap around each other and form a 3-stranded helix. Amino acid analysis shows that this is collagen type 1. Each polypeptide chain contains about 1000 cross-linked amino acids. The collagen molecule consists of 2 identical polypeptides,

α-1(1), and a third polypeptide chain that has a different amino acid sequence, a-2(1). The individual polypeptide chains can be separated by sodium dodecyl sulfate-polyacrylamide gel electrophoresis.

A dermal filler is indicated for correcting contour deficiencies of soft tissue. Wrinkles develop because the thickness of the skin's dermal layer significantly diminishes during aging. As a case example, we consider a dermal filler composed of absolutely round and smooth polymethyl methacrylate (PMMA), a synthetic polymer widely used in medical, industrial, and cosmetic applications. The filler comprises PMMA microspheres, 30–42 microns in size, suspended in a water-based carrier gel containing 3.5% bovine collagen, 96.5% buffered isotonic water for injection, and 0.3% lidocaine [30].

The PMMA microspheres are suspended in a solution of partly denatured 3.5% bovine collagen. Following injection of the filler, the collagen vehicle is absorbed by the body within 1–3 months, during which the nondegradable PMMA microspheres stimulate the body to encapsulate each sphere with the patient's collagen. This results in a long-lasting correction of wrinkles and other soft tissue defects [30]. Bovine collagen is converted to atelocollagen by treatment with pepsin to remove the peptide ends, thus reducing its antigenic potential [31].

From the Artes Laboratory report [30], the toxic metal content in the syringe of the semi-permanent dermal filler product Artecoll was determined as follows (Table 1): lead (Pb)=0.03 µg, chromium (Cr)=0.14 µg, cadmium (Cd)=0.017 µg, and mercury (Hg)<0.006 µg per 0.5 g Artecoll. The concentrations of lead, chromium, cadmium, and mercury were reported to be 0.057 ppm, 0.259 ppm, 0.030 ppm, and 0.010 ppm, respectively. This indicates that not only are the individual concentrations of each heavy metal in Artecoll well below 1 ppm, but the combined total of all heavy metals is also less than 0.4 ppm. As a result, the risk of releasing toxic levels of heavy metals from Artecoll is considered negligible [30]. The current permissible exposure limit for chromium was found to be 1 mg/m^3 TWA. The LD50 (median lethal dose) of chromium trioxide subcutaneously injected into a dog was 330 mg/kg body weight [1]. Approximately 1 g of potassium dichromate is considered a lethal dose preceded by gastrointestinal bleeding and massive fluid loss [5]. The revised Immediately Dangerous to Life or Health (IDLH) level for chromium was set to 250 mg Cr/m^3 air [30].

Table . Component specifications for 3.5% atelocollagen.

Parameter	Specification	Method
Collagen (calculated from hydroxyproline)	3.0 – 4.0%	Spectrophotometry
Hydroxyproline	0.41 – 0.55%	Spectrophotometry
Lidocaine HCI[a]	0.27 – 0.33%	HPLC[b]
Heavy metals	<20 ppm	DAB 10[c]
pH	6.8 – 7.8	DAB 10
Pyrogenicity	<36.25 EU/ml	DAB 10
Sterility	Sterile	DAB 10

[a]HCI: hydrochloric acid; a strong, corrosive acid commonly used in chemical reactions, laboratory testing, and pH control.

[b]HPLC: high-performance liquid chromatography; an analytical technique used to separate, identify, and quantify components in a mixture, widely applied in pharmaceuticals, environmental analysis, and biochemistry.

[c]DAB-10: 10-deacetyl baccatin.

The results of the toxicological laboratory data show no evidence that acute exposure to a high chromium concentration would cause irreversible health effects within 30 minutes (Table 2) [32].

Table . Polymethyl methacrylate heavy metals specifications list the daily requirement of chromium.

Item	Specification
Cd	<0.1 ppm
Hg	<0.1 ppm
Pb	<0.2 ppm

The duration of a collagen toxicity test can vary depending on the type and scope of the study:

1. Acute toxicity tests: These are short-term studies, typically lasting a few days to a couple of weeks [33].
2. Subchronic toxicity tests: These studies usually span around 90 days [33,34].
3. Chronic toxicity tests: These long-term studies can last several months to a year or more [33]

## Using AI to Determine the Drug Toxicity of Collagen

AI algorithms can be used to predict toxicity based on the chemical and biological properties of the compounds. AI uses neural networks to analyze molecular graphs or sequences to detect toxicity-related patterns.

The evaluation of pharmacokinetics and toxicity is crucial for designing new therapeutic candidates with in silico virtual screens, and generative AI outputs a vast number of molecules that must be filtered into a tractable number for synthesis and experimental validation. For this case example, the absorption, distribution, metabolism, excretion, and toxicity (ADMET) AI program was used to determine the toxicity of collagen. ADMET is an effective primary filter that evaluates candidate compounds based on their ADMET properties. ADMET-AI is a simple,

fast, and accurate digital computer web interface for predicting the ADMET properties of molecules using machine learning models.

The virtual calculation of the blood-brain barrier is shown in Figure 3 [30], which effectively protects the brain tissue from circulating pathogens and other potentially toxic substances. This calculation shows the toxicity of collagen to be low. Collagen itself was shown to be safe and nontoxic.

Figure 3. Virtual calculation of the blood-brain barrier. hERG: human Ether-à-go-go–related gene.
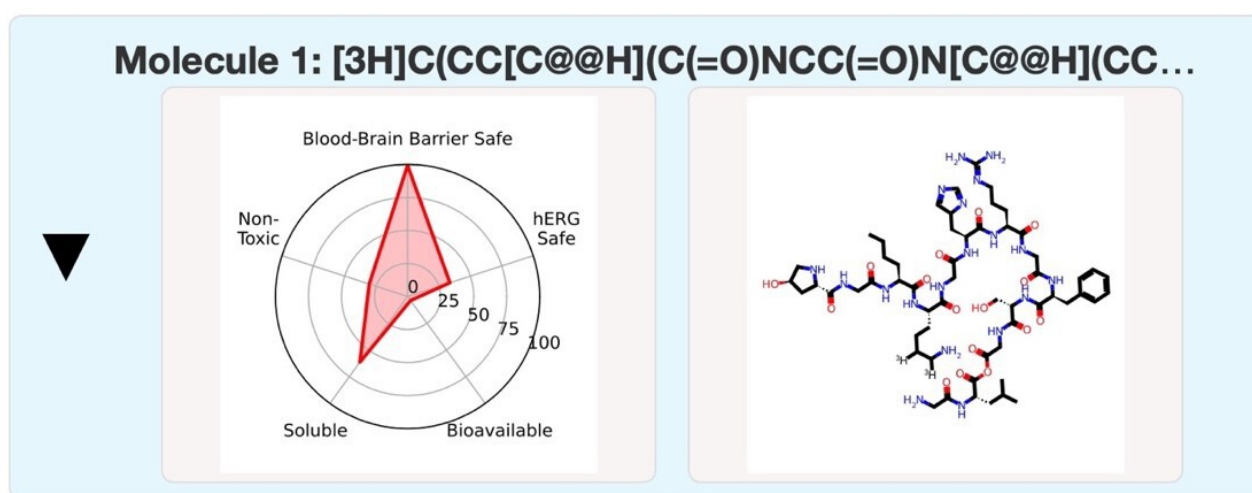
Challenges With the Uses of AI for Drug Discovery

Despite promising advancements, several challenges remain for the integration of AI and quantum computing in drug discovery. The ethical implications of using AI in drug discovery must be addressed. Ensuring transparency in AI algorithms and maintaining accountability in decision-making processes are critical to gaining public trust and regulatory approval, which can be achieved by using an explainable AI approach. Furthermore, the potential for bias in AI models necessitates ongoing scrutiny to ensure equitable access to new therapies.

Data Privacy and Ethics

The use of AI and AI algorithms comes with concern for the privacy and security of user data. Data poising and alterations underlying models put AI users at risk. Implementing federated learning allows for the training of AI models on decentralized data sources without sharing sensitive data. The fully homomorphic encryption technique is used in most federated searching techniques. This is crucial in drug discovery, where patient data and proprietary research information must remain confidential. Federated learning enables collaborative learning across different research institutions or pharmaceutical companies, allowing them to leverage each other's data without compromising privacy. Since raw data are not centralized, the risk of data breaches is minimized, making it a secure choice for handling sensitive information in drug discovery. Fedrated learning can be integrated with the AI and quantum computing techniques discussed in this paper, enhancing the predictive capabilities while maintaining data integrity and privacy.

Validation and Accuracy

Computational models must be rigorously validated against experimental data to ensure that their predictions are reliable. This includes demonstrating that computational methods are accurate and can reliably substitute for laboratory-generated data. Multiple stakeholders (eg, academia, industry, and regulatory bodies) would need to validate and reproduce computational data for different types of products.

AI models rely on large, high-quality datasets, whereas pharmaceutical data are often limited, biased, or proprietary, affecting the model's performance. In addition, AI-generated predictions can lack transparency, making it difficult to understand how a model arrived at a particular conclusion, which is critical in drug development. Although AI predictions can be highly accurate, inconsistencies may still lead to failures in identifying effective drugs or result in overlooking promising candidates. To ensure reliability, AI-driven drug discovery must meet stringent FDA regulatory standards and address ethical concerns, including potential bias in drug development and risks to patient safety. AI models often struggle with the complexity of biological systems, such as multitarget interactions, immune response, and genetic variations. Despite these challenges, AI will continue to improve and is expected to play a significant role in the future of drug discovery. The findings of the present case study were intended to demonstrate that the computational assessment of drug toxicity closely aligns with actual laboratory data. This approach not only replicates laboratory results but does so at a significantly reduced cost.

Strengths of the Proposed Approach

Computational models can lower the costs of bringing new drugs to market by reducing the need for extensive animal studies or large human trials. Quantum computing, AI, and machine learning have improved with respect to accuracy and generalizability, and there is growing potential for their application in areas traditionally requiring laboratory data (eg, toxicology and pharmacodynamics). Advances in quantum computing, molecular dynamics, and systems biology would help computational models closely mimic biological systems and make predictions more reliable.

AI and quantum computing facilitate the drug discovery process from the following aspects:

1. Data analysis and pattern recognition: AI algorithms can analyze vast datasets, including genetic Protonix and clinical data to identify potential therapeutic targets and predict drug interactions. This capability allows researchers to uncover disease-associated targets and molecular pathways more efficiently than traditional methods, which often rely on trial and error [35-37].
2. Molecular simulation: Quantum computing enables more accurate simulations of molecular interactions than classical computers [38]. This allows researchers to explore a broader range of potential drug candidates and significantly predict their efficacy and safety, speeding up the drug discovery process [37].
3. Integration of computational models: The combination of AI and quantum computing allows for the development of sophisticated computational models to simulate complex biological systems. This integration can lead to better-informed decisions in drug development and regulatory processes, ultimately enhancing patient safety [35].
4. Reduction of laboratory testing: By using computational data, the need for extensive laboratory and animal testing can be decreased. This not only reduces cost but also shortens the time required to bring new drugs to market [37].
5. Quantifying development costs: The costs are quantified by evaluating the total expenses incurred during the drug development process, including research and development, clinical trials, and regulatory approvals. Traditional methods can take up to 15 years and cost around US $1 billion, whereas quantum computing can potentially reduce this timeline and cost significantly [37].

Researchers may also review case studies where quantum computing has been implemented in drug discovery to assess the financial and temporal savings achieved compared to conventional methods [37].

The burgeoning field of computational data, propelled by AI and quantum computing advancements, stands to revolutionize new drug discovery and approval processes. Computational methods can significantly accelerate the identification of potential drug candidates, predict their efficacy, and assess safety, thereby reducing the traditional time and cost burdens associated with pharmaceutical development. By integrating AI and quantum computing with extensive chemical databases, researchers can efficiently simulate biological interactions, streamline virtual screening, and predict drug toxicity—ultimately enhancing the likelihood of successful drug development. Furthermore, the implications for the FDA regulatory framework are examined, highlighting how computational data can inform and expedite the approval process, leading to faster review cycles and improved postmarket surveillance. This situation calls for a paradigm shift from traditional laboratory methods to data-driven approaches, emphasizing the need for rigorous validation and collaboration among stakeholders to establish robust regulatory standards for computational models in drug discovery.

AI is far cheaper per compound than laboratory-based testing, especially for initial screenings. For example, screening 1000 compounds via AI might cost US $10,000–US $50,000, depending on the computational setup [20]. The same screening using in vitro methods could cost US $1–US $10 million or US $50–US $500 million using in vivo methods once augmented reality AI models are deemed significant. Once developed and validated, these models significantly reduce long-term expenses, making them more cost-effective than laboratory methods for large-scale or preliminary screenings.

As demonstrated with our case study, AI is often used as a first-pass filter to predict drug toxicity, reducing the number of compounds that need to be tested in the laboratory. By prioritizing only those promising candidates for laboratory testing, researchers can combine the speed and cost-effectiveness of AI with the rigor and accuracy of laboratory results, achieving a balance of cost and reliability.

Summary and Future Prospects

AI and quantum computing offer unique capabilities to tackle complex problems in drug discovery, which is a critical challenge in pharmaceutical research. Regulatory agents will need to adapt to these new technologies. Regulatory processes may become more streamlined, using adaptive clinical trials, accelerating pathways, and better integrating digital data to reduce the time and cost of bringing new drugs to market. Computational data methods could reduce the cost and time involved in experimental drug discovery, allowing researchers to simulate biological interactions and screen large compound libraries more efficiently. Creating virtual data for drug discovery involves several stages, each using specific methods such as simulations, synthetic data generation, data augmentation, and tools to generate, collect, and affect human interaction to identify and develop new drugs. Here, we have emphasized that knowing the molecular structure of a drug is a critical factor in determining its toxicity and for other aspects of the drug discovery and approval process. Using computational data in drug discovery has transformed the pharmaceutical and biotechnology industries by accelerating research, reducing costs and timeliness, and improving the likelihood of success. Overall, the integration of AI and quantum computing represents a transformative shift in drug discovery, offering the potential for faster, more efficient, and more effective therapeutic development. As these technologies continue to evolve, they

XSL·FO

RenderX

will likely play a pivotal role in shaping the future of pharmaceuticals. Nevertheless, several research questions remain to be explored to realize this shift, including:

(1) Can AI reliably predict drug toxicity compared to traditional laboratory results? Hypothesis: The incorporation of quantum computing into molecular modeling improves the predictive capabilities of AI, leading to more accurate toxicity assessments.

(2) Does the integration of quantum computing enhance the accuracy of molecular modeling and drug discovery? Hypothesis: The incorporation of quantum computing into molecular modeling improves the predictive capabilities of AI, leading to more accurate toxicity assessments.

(3) How do AI-driven toxicity predictions compare to laboratory outcomes in terms of cost and time efficiency? Hypothesis: Using AI and quantum computing for toxicity prediction significantly reduces the need for laboratory experiments, thereby decreasing both costs and development time in the drug discovery process.

The convergence of AI and quantum computing holds great potential for revolutionizing drug discovery and approval processes. Continued research is needed to refine quantum algorithms and integrate them with AI systems effectively.

Acknowledgments

Conflicts of Interest

None declared.

References

1. ElZarrad MK, Lee AY, Purcell R, et al. Advancing an agile regulatory ecosystem to respond to the rapid development of innovative technologies. Clin Transl Sci 2022 Jun;15(6):1332-1339. [doi: 10.1111/cts.13267] [Medline: 35319833]
2. Gelernter D. Mirror Worlds: Or: The Day Software Puts the Universe in a Shoebox How It Will Happen and What It Will Mean: Oxford University Press; 1993.
3. Izmailova ES, Demanuele C, McCarthy M. Digital health technology derived measures: biomarkers or clinical outcome assessments? Clin Transl Sci 2023 Jul;16(7):1113-1120. [doi: 10.1111/cts.13529] [Medline: 37118983]
4. Asiri Y. Computing drug-drug similarity from patient-centric data. Bioengineering (Basel) 2023 Feb 1;10(2):182. [doi: 10.3390/bioengineering10020182] [Medline: 36829676]
5. Tao F, Qi Q. Make more digital twins. Nature New Biol 2019 Sep 26;573(7775):490-491. [doi: 10.1038/d41586-019-02849-1]
6. Vora LK, Gholap AD, Jetha K, et al. Artificial intelligence in pharmaceutical technology and drug delivery design. Pharmaceutics 2023 Jul 10;15(7):1916. [doi: 10.3390/pharmaceutics15071916] [Medline: 37514102]
7. Ang D, Rakovski C, Atamian HS. De novo drug design using transformer-based machine translation and reinforcement learning of an adaptive Monte Carlo tree search. Pharmaceuticals (Basel) 2024 Jan 27;17(2):161. [doi: 10.3390/ph17020161] [Medline: 38399376]
8. Han R, Yoon H, Kim G, et al. Revolutionizing medicinal chemistry: the application of artificial intelligence (AI) in early drug discovery. Pharmaceuticals (Basel) 2023 Sep 6;16(9):1259. [doi: 10.3390/ph16091259] [Medline: 37765069]
9. Lee L. Biopharma thought leaders: how AI is accelerating and transforming drug discovery. Biopharma Dealmakers. 2023 Jun. URL: www.nature.com/biopharmdeal [accessed 2025-06-25]
10. Sadybekov AA, Sadybekov AV, Liu Y, et al. Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. Nature New Biol 2022 Jan;601(7893):452-459. [doi: 10.1038/s41586-021-04220-9] [Medline: 34912117]
11. Wang Y, Krstić PS. Multistate transition dynamics by strong time-dependent perturbation in NISQ era. J Phys Commun 2023 Jul 1;7(7):075004. [doi: 10.1088/2399-6528/ace67a]
12. Cao Y, Romero J, Aspuru-Guzik A. Potential of quantum computing for drug discovery. IBM J Res & Dev 2018;62(6):6. [doi: 10.1147/JRD.2018.2888987]
13. Vedral V. The Everything-Is-a-Quantum-Wave Interpretation of Quantum Physics. Quantum Rep 2023;5(2):475-480. [doi: 10.3390/quantum5020031]
14. Hassan A, Ibrahim A. The Coming Quantum Computing Evolution in the Pharmaceutical Industry and Drug R&D. Research Berg Review of Science and Technology 2023;3(11):1-16.
15. U.S. Food & Drug Administration. Artificial Intelligence in Drug Manufacturing. 2023. URL: https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://www.fda.gov/media/165743/download&vedqu=2ahUKEwjjhe-6-P-IAxW_GDQIHSpNAKsQFnoECBcQAQ&usg=AOvVaw0qw5m2HuZ6Xl7znNB4wd0S [accessed 2025-06-06]
16. Pognan F, Beilmann M, Boonen HCM, et al. The evolving role of investigative toxicology in the pharmaceutical industry. Nat Rev Drug Discov 2023 Apr;22(4):317-335. [doi: 10.1038/s41573-022-00633-x] [Medline: 36781957]
17. Huang B, von Lilienfeld OA. Quantum machine learning using atom-in-molecule-based fragments selected on the fly. Nat Chem 2020 Oct;12(10):945-951. [doi: 10.1038/s41557-020-0527-z] [Medline: 32929248]

18. Qiu X, Li H, Ver Steeg G, et al. Advances in AI for protein structure prediction: implications for cancer drug discovery and development. Biomolecules 2024 Mar 12;14(3):339. [doi: 10.3390/biom14030339] [Medline: 38540759]

19. PubChem compound summary for CID 6913668, collagen i, alpha chain (98-110). National Center for Biotechnology Information. URL: https://pubchem.ncbi.nlm.nih.gov/compound/Collagen-I_-alpha-chain-_98-110 [accessed 2025-06-24]

20. Srivastava R. Quantum computing in drug discovery. Infor Syst Smart City 2023;3(1):294. [doi: 10.59400/issc.v3i1.294]

21. Cova T, Vitorino C, Ferreira M, et al. Artificial intelligence and quantum computing as the next pharma disruptors. In: Artificial Intelligence in Drug Design: Springer; 2022, Vol. 2390:321-347. [doi: 10.1007/978-1-0716-1787-8_14]

22. Rayhan A, Rayhan S. Quantum computing and AI: a quantum leap in intelligence. In: AI Odyssey: Unraveling the Past, Mastering the Present, and Charting the Future of Artificial Intelligence: NotunKhabar; 2023, Vol. 424.

23. Pyrkov A, Aliper A, Bezrukov D, et al. Quantum computing for near-term applications in generative chemistry and drug discovery. Drug Discov Today 2023 Aug;28(8):103675. [doi: 10.1016/j.drudis.2023.103675] [Medline: 37331692]

24. Kumar G, Yadav S, Mukherjee A, et al. Recent advances in quantum computing for drug discovery and development. IEEE Access 2024;12:64491-64509. [doi: 10.1109/ACCESS.2024.3376408]

25. Mishra R, Mishra PS, Mazumder R, et al. Quantum computing and its promise in drug discovery. In: Drug Delivery Systems Using Quantum Computing 2024:57-92. [doi: 10.1002/9781394159338]

26. Sharma P. Quantum computing in drug design: enhancing precision and efficiency in pharmaceutical development. Sage Science Review of Applied Machine Learning 2024:1-9.

27. Popa R, Dumitrescu E. Drug discovery in the 21st century: exploring the promises and potential of quantum machine learning. J Contemp Healthc Analytics 2023;7(12):1-13 [FREE Full text]

28. Chow JCL. Quantum computing in medicine. Med Sci (Basel) 2024 Nov 17;12(4):67. [doi: 10.3390/medsci12040067] [Medline: 39584917]

29. Protein Data Bank. Crystal structure of Type IV collagen NC1 domain from bovine lens capsule. 2024. URL: https://www.rcsb.org/structure/1T60 [accessed 2025-06-06]

30. Artes Medical USA, Inc. Report of laboratory result of collagen toxicity. 2002 URL: https://www.accessdata.fda.gov/cdrh_docs/pdf2/P020012C.pdf [accessed 2025-07-20]

31. OEChem 2.3.0 [release 2021.10.14]. Open Eye Scientific. URL: https://docs.eyesopen.com/toolkits/python/oechemtk/releasenotes/version2_3_0.html [accessed 2025-07-11]

32. Bettanti A, Beccari AR, Biccarino M. Exploring the future of biopharmaceutical drug discovery: can advanced AI platforms overcome current challenges. Discov Artif Intell 2004;4(102). [doi: 10.1007/s44163-024-00188-3]

33. Marone PA, Lau FC, Gupta RC, et al. Safety and toxicological evaluation of undenatured type II collagen. Toxicol Mech Methods 2010 May;20(4):175-189. [doi: 10.3109/15376511003646440] [Medline: 20170336]

34. Protein hydrolyzates, animal. ECHA. URL: https://echa.europa.eu/registration-dossier/-/registered-dossier/14739/7/6/2 [accessed 2025-06-06]

35. How artificial intelligence is revolutionizing drug discovery. Petrieflom. URL: https://petrieflom.law.harvard.edu/2023/03/20/how-artificial-intelligence-is-revolutionizing-drug-discovery/ [accessed 2025-06-06]

36. Quantum computing for drug discovery: accelerating research. Quantum Zeitgeist. URL: https://quantumzeitgeist.com/quantum-computing-for-drug-discovery-accelerating-research/#google_vignette [accessed 2025-06-06]

37. Mak KK, Pichika MR. Artificial intelligence in drug development: present status and future prospects. Drug Discov Today 2019 Mar;24(3):773-780. [doi: 10.1016/j.drudis.2018.11.014] [Medline: 30472429]

38. Patel V, Shah M. Artificial intelligence and machine learning in drug discovery and development. Intelligent Medicine 2022 Aug;2(3):134-140. [doi: 10.1016/j.imed.2021.10.001]

Abbreviations

ADMET: absorption, distribution, metabolism, excretion, and toxicity
AI: artificial intelligence
DHT: digital health technology
FDA: US Food and Drug Administration
HOMO-LUMO: Highest Occupied Molecular Orbital-Lowest Unoccupied Molecular Orbital
IDLH: Immediately Dangerous to Life or Health
ISO: International Organization for Standardization
LD50: median lethal dose
PMMA: polymethyl methacrylate
SMILES: simplified molecular input line entry system
VQE: Variational Quantum Eigensolver

XSL·FO

RenderX

Correction: Structural and Functional Impacts of SARS-CoV-2 Spike Protein Mutations: Insights From Predictive Modeling and Analytics

Edem K Netsey[1*], MPhil; Samuel M Naandam[2], PhD; Joseph Asante Jnr[3*], PhD; Kuukua E Abraham[4], MS; Aayire C Yadem[5], PhD; Gabriel Owusu[6], PhD; Jeffrey G Shaffer[7], PhD; Sudesh K Srivastav[7], PhD; Seydou Doumbia[8], PhD; Ellis Owusu-Dabo[9], PhD; Chris E Morkle[10], MEd; Desmond Yemeh[11], PharmD; Stephen Manortey[12], PhD; Ernest Yankson[2], PhD; Mamadou Sangare[13], PhD; Samuel Kakraba[7,11*], PhD

[1]Department of Mathematics and Information Communication Technology, Dambai College of Education, Dambai, Ghana

[10]Department of Mathematics, Suhum Senior High School, Suhum, Ghana

[11]Tulane Center for Aging, School of Medicine, Tulane University, New Orleans, LA, United States

[12]Department of Community Health, Ensign Global University, Kpong, Ghana

[13]Laboratory of Malaria and Vector Research (LMVR), National Institute of Allergy and Infectious Diseases (NIAID), Rockville, MD, United States

[2]Department of Mathematics, School of Physical Sciences, University of Cape Coast, Cape Coast, Ghana

[3]Department of Geriatrics, School of Medicine, University of Arkansas for Medical Sciences, Little Rock, AR, United States

[4]Department of Mathematics, Memphis Shelby County Schools, Memphis, TN, United States

[5]Research and Development, CytoAstra LLC, Little Rock, AR, United States

[6]Office of Research, Celia Scott Weatherhead School of Public Health and Tropical Medicine at Tulane University, New Orleans, LA, United States

[7]Department of Biostatistics and Data Science, Celia Scott Weatherhead School of Public Health and Tropical Medicine at Tulane University, 1440 Canal Street, New Orleans, LA, United States

[8]Department of Public Health, Faculty of Medicine, Malaria Research and Training Center, University of Sciences, Techniques and Technologies of Bamako, Bamako, Mali

[9]School of Public Health, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

[*]these authors contributed equally

Corresponding Author:
Samuel Kakraba, PhD
Department of Biostatistics and Data Science, Celia Scott Weatherhead School of Public Health and Tropical Medicine at Tulane University, 1440 Canal Street, New Orleans, LA, United States

Related Article:

Correction of: https://bioinform.jmir.org/2025/1/e73637

Abstract

In "Structural and Functional Impacts of SARS-CoV-2 Spike Protein Mutations: Insights From Predictive Modeling and Analytics" [1], the authors noted four errors.

Author JAJ has now been marked as an equal contributor, as indicated by the asterisk added to their name.

Affiliation 1 has been changed from the following:

Department of Mathematics and Information Communication Technology, School of Physical Sciences, Dambai College of Education, Dambai, Ghana

The affiliation now reads:

Department of Mathematics and Information Communication Technology, Dambai College of Education, Dambai, Ghana

Affiliation 11 has been changed from the following:

Interdisciplinary Aging Studies, Tulane Center for Aging, School of Medicine, Tulane University, New Orleans, LA, United States

The affiliation now reads:

Tulane Center for Aging, School of Medicine, Tulane University, New Orleans, LA, United States

Affiliation 13 has been changed from the following:

Department of National Institute of Allergy and Infection Diseases (NIAID), 14 Laboratory of Malaria

and Vector Research (LMVR), National Institute of Allergy and Infectious Diseases (NIAID), Rockville, MD, United States

The affiliation now reads:

Laboratory of Malaria and Vector Research (LMVR), National Institute of Allergy and Infectious Diseases (NIAID), Rockville, MD, United States

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

Reference

1. Netsey EK, Naandam SM, Asante Jnr J, et al. Structural and functional impacts of SARS-CoV-2 spike protein mutations: insights from predictive modeling and analytics. JMIR Bioinform Biotech 2025;6:e73637. [doi: 10.2196/73637] [Medline: 12685290]

XSL•FO
RenderX

Immunogenicity of Adalimumab in Bacterial Molecular Mimicry: In Silico Analysis

Diana Isabel Pachón-Suárez[1,2], MD; Germán Mejía-Salgado[2], MSc, MD; Oscar Correa[3], MD; Andrés Sánchez[4], MSc; Marlon Munera[4], MSc, PhD; Alejandra de-la-Torre[2], MD, PhD

[1]Doctorate in Clinical Research, Escuela de Medicina y Ciencias de la Salud, Universidad del Rosario, Bogota, Colombia

[2]Neuroscience (NEUROS) Research Group, Neurovitae Research Center, Institute of Translational Medicine (IMT), Universidad del Rosario, Cra 24 # 63C- 69, Bogota, Colombia

[3]Pulmonology and Immunology in Pediatrics Research Group, Department of Pediatrics, School of Medicine, Universidad Nacional de Colombia, Bogota, Colombia

[4]Faculty of Health, Medical Research Group (GINUMED), Corporación Universitaria Rafael Nuñez, Cartagena, Colombia

Corresponding Author:
Alejandra de-la-Torre, MD, PhD
Neuroscience (NEUROS) Research Group, Neurovitae Research Center, Institute of Translational Medicine (IMT), Universidad del Rosario, Cra 24 # 63C- 69, Bogota, Colombia

Abstract

Background: Adalimumab, a monoclonal antibody targeting tumor necrosis factor α, treats autoimmune diseases but induces antidrug antibodies in 30% to 60% of patients, reducing its efficacy.

Objective: This study aims to investigate molecular mimicry as a mechanism behind this immunogenicity, where bacterial immunoglobulin domains structurally resemble adalimumab's light chain, triggering immune responses.

Methods: Using PSI-BLASTp (National Center for Biotechnology Information) and PRALINE (Center for Integrative Bioinformatics), there are 40 bacterial antigens homologous to adalimumab, with 8 clinically relevant strains.

Results: Structural analysis revealed 94% amino acid identity between the immunoglobulin domain of *Escherichia coli* strain B1 and adalimumab's light chain, and 89.67% similarity with *Corynebacterium pyruviciproducens*. Root mean square deviation values confirmed strong structural homology. Additionally, 5 cross-reactive B-cell epitopes were predicted, suggesting overlapping surfaces that may promote immune cross-reactivity and antidrug antibody development.

Conclusions: This study represents a first step toward identifying a potential microbial factor driving antiadalimumab antibody formation. The predicted cross-reactive regions provide specific candidates for further in vitro validation to confirm molecular mimicry and refine epitope mapping. Understanding these mechanisms may ultimately inform the design of less immunogenic biologics and guide clinical strategies to predict and prevent antidrug antibody formation.

Introduction

Adalimumab is a fully human monoclonal antibody targeting tumor necrosis factor-α, a protein involved in inflammation in various chronic autoimmune conditions [1]. The Food and Drug Administration has approved adalimumab to treat several diseases: rheumatoid arthritis, ankylosing spondylitis, Crohn disease, ulcerative colitis, hidradenitis suppurativa, juvenile idiopathic arthritis, plaque psoriasis, psoriatic arthritis, and uveitis [1]. Despite the humanization of adalimumab, the amino acid sequences of both the heavy and light variable chains near the epitope binding regions within the complementarity-determining regions tend to elicit a robust immune response [1-3]. Immune complexes formed by

adalimumab and antigens can reach 4000 kDa, and despite being humanized, they may still be recognized as foreign, triggering antidrug antibodies (ADAs) that reduce its effectiveness [1,2].

Approximately 30% to 60% of patients on treatment with adalimumab eventually experience a reduction in the effectiveness of the treatment [4]. This waning efficacy is believed to be due, in part, to immunogenicity, which refers to the body's generation of antibodies targeting the biological drug [4]. One potential mechanism underlying this immune response is immune cross-reactivity, where the immune system interacts with structurally similar antigens from different sources. Among these, molecular mimicry—where bacterial proteins share structural homology with therapeutic antibodies like adalimumab—has emerged as a plausible explanation for this

phenomenon [5,6]. Four requirements must be met for an infection to be implicated in the molecular mimicry-based development of an autoimmune response: there must be epidemiological data about exposure to the environmental agent and the development of autoimmunity, structural homology between human antigens (or medication) and pathogens, autoantibodies or autoreactive lymphocytes against both human and pathogen antigens, and in vivo evidence in animal models [2,7].

Immunoglobulin domains, widely recognized for their role in the structure and function of key immunological proteins, are highly conserved units across evolution [8]. Interestingly, these domains are not confined to the immune systems of higher organisms but are also present in bacterial proteins, including those of *Escherichia coli* and other enterobacteria [8,9]. Immunoglobulin-like domains are frequently found in these microorganisms in cell surface proteins and fimbrial organelles, where they play essential roles in host cell adhesion and invasion by pathogenic strains. They serve as structural components of pilus and nonpilus fimbrial systems and are members of the intimin or invasin family of outer membrane adhesins [8]. This dual functionality underscores their significance as a possible evolutionary mechanism through which pathogens leverage these conserved structures to evade or modulate the host immune response, facilitating infection and colonization [8].

The immune cross-reactivity highlights the need for a comprehensive understanding of the immune responses induced by microorganisms in drugs. It has important implications for the development and use of pharmacological therapies. Currently, there is no evidence of immune cross-reactivity between adalimumab and microbial antigens. However, cross-reactive immune responses have been reported with other medications and vaccines, leading to thrombocytopenia and autoimmune diseases. Furthermore, patients with underlying autoimmune conditions who develop infections are more likely to produce anti-drug antibodies [5,10-15].

Molecular mimicry has traditionally been investigated in the context of autoimmune diseases and vaccine responses, where microbial antigens resemble host proteins. However, its potential role in the immunogenicity of therapeutic antibodies remains largely unexplored. In this study, we apply this concept to examine whether infections could trigger cross-reactive immune responses toward biological drugs. Specifically, we investigate possible cross-immunogenicity between the immunoglobulin domain of clinically relevant Gram-positive and Gram-negative bacteria and adalimumab using in silico approaches. To our knowledge, this is the first study to propose a mechanistic link between bacterial antigens and adalimumab immunogenicity supported by sequence and structural evidence. By addressing this unexplored aspect of biologic drug immunogenicity, our work provides a novel conceptual framework that may guide future experimental validation and inform strategies to improve therapeutic antibody safety.

Methods

Study Design

A workflow image of the method is shown in Figure S1 in Multimedia Appendix 1.

Sequences Analysis

Adalimumab's amino acid sequence, identified by its DrugBank Accession Number (DB00051), was obtained from DrugBank [16,17]. The complete amino acid sequences of both the heavy (α) and light (β) chains corresponding to the Fab region were used for the analyses. These sequences include the variable and constant domains, ensuring full structural representation of the monoclonal antibody during alignment and comparison. Adalimumab sequence served as the input for a PSI-BLASTp (version 2.16.0; National Center for Biotechnology Information) search targeting bacterial homologs, using the identifier Bacteria (taxid:2). The length of matched subsequences in PSI-BLASTp is influenced by statistical significance thresholds and the iterative nature of the algorithm, which together help ensure that only meaningful alignments are included in the analysis [18]. Default settings were applied for the general search parameters. For subsequent analyses, amino acid sequences from bacteria of clinical significance to humans were selected [19].

Antigens demonstrating a similarity of ≥30% were considered for further investigation. The amino acid sequences from the chosen microorganisms were aligned with the adalimumab light and heavy chains to ascertain identity levels and pinpoint conserved regions. The PRALINE tool (version 2; Center for Integrative Bioinformatics) [20] facilitated the alignments by identifying regions of similarity, which may indicate functional, structural, or evolutionary relationships among the proteins being compared. The alignment parameters were configured to use BLOSUM62 as the exchange matrix using default parameters unless otherwise specified. Specifically, we used BLOSUM62 as the substitution matrix with a gap opening penalty of 11 and a gap extension penalty of 1 (standard for PSI-BLAST alignments). For the PSI-BLASTp search, 3 iterations were performed with an E-value threshold of 0.001 to enhance sensitivity and identify distant homologs [21]. E-value represents the number of random matches you would expect to find with a score equal to or better than the one observed [22]. Similarly, antigens with a similarity of ≥30% were advanced for further analysis. Also, a high-resolution Protein Data Bank (PDB) file (ID: 3wd5) was sourced from the PDB [23,24], enabling structural analysis.

Modeling Based on Homology

3D models of selected antigens, for which no reports exist in the PDB, were constructed based on homology using the SWISS-MODEL server (ProMod3) [25,26]. These initial models underwent further refinement with UCSF Chimera (version 1.1.3) [27].

Antigens with experimentally resolved 3D structures were sourced directly from the PDB. The visualization of all models was achieved using PyMOL (version 3.0; Schrödinger, Inc) [25]. Structural homology assessments were conducted using

the Ramachandran charts, the Quantitative Model Energy Analysis index, the RMSD metric, and Global Model Quality Estimate (GMQE) values were assessed for all models. RMSD focuses on precise structural alignment, while GMQE provides a broader model quality evaluation. Using both metrics allows researchers to make informed decisions about the reliability and applicability of their homology models in biological research [26,28].

B-Cell Epitope Prediction

The prediction of B-cell epitopes was carried out using the ElliPro server (version 3.0; IEDB Analysis Resource), using the default parameters that are a minimum score protrusion index (PI) threshold of 0.5 and a maximum distance of 6 Å between residue centers for defining discontinuous epitopes. ElliPro identified linear and discontinuous epitopes based on the protein's 3D structure (PDB ID: 3WD5) using the PI of residues [29,30]. B-cell epitope prediction methods, such as those using the Ellipro server, generally achieve accuracy rates ranging from 65% to over 70%. Still, ongoing validation and refinement are necessary due to variability in sensitivity and specificity across different prediction tools [27]. Furthermore, epitopes previously identified for adalimumab were retrieved from the Immune Epitope Database (IEDB). This step was essential to explore the potential molecular mimicry between antigens from bacteria and those associated with adalimumab. Epitopes conserved between adalimumab and its bacterial homologs were visualized on the 3D model of the monoclonal antibody using PyMOL version 3.0.

Major Histocompatibility Complex Class II–Dependent T-Cell Epitope Prediction

T-cell epitope prediction was performed using the NetMHCIIpan version 4.1 web server (IEDB Analysis Resource) [31]. This platform uses an ensemble of deep neural networks trained on large quantitative binding datasets (IC values) derived from multiple human leukocyte antigen (HLA) class II molecules, enabling panallelic prediction of peptide-HLA interactions.

The FASTA sequences of the VH and VL chains were analyzed separately. The alleles HLA-DRB1*01:01, HLA-DQA1*01:01/DQB1*05:01, HLA-DPA1*01:03/DPB1*02:01, and HLA-DQA1*05:01/DQB1*02:01 were selected due to their frequency and immunogenetic relevance in diverse human populations commonly used in therapeutic immunogenicity assessments. Epitope scanning was performed with a 15-amino-acid window (15-mer) sliding by one residue to ensure maximal coverage of overlapping peptides. The software generated an affinity score (nM) and a percentile rank for each predicted peptide. Following the server's guidelines, peptides with rank ≤2.0% were classified as strong binders, while those with 2.0%<rank≤10% were considered weak binders.

Filtering, Ranking, and Epitope Selection

All prediction outputs were exported in CSV format. The data were subsequently filtered to retain only peptides with an affinity score ≥0.2 and low percentile rank, indicative of stable peptide–major histocompatibility complex (MHC) II complexes. Scores and ranks were compared across both antibody chains

to identify high-affinity regions and potential immunogenic hotspots.

Population Coverage Analysis

To assess the global relevance and potential immunological reach of the predicted MHC class II–restricted T-cell epitopes, a population coverage (PC) analysis was conducted using the population coverage tool available at the IEDB [32].

This analysis estimates the fraction of individuals within defined human populations that are likely to present one or more of the predicted epitopes, based on the distribution frequency of HLA alleles. The tool integrates the predicted epitope–HLA binding data obtained from NetMHCIIpan 4.1 with HLA genotypic frequencies derived from the Allele Frequency Net Database, which compiles large-scale datasets from diverse ethnic and geographical groups worldwide.

All epitopes predicted as strong or weak binders (rank ≤10%) across the selected HLA-DR, HLA-DP, and HLA-DQ alleles were included as input. The analysis was performed for multiple population sets, including global coverage, South American, and European cohorts, representing regions with significant therapeutic use of adalimumab and diverse HLA genetic backgrounds.

The algorithm computes several parameters: projected PC (%), representing the cumulative percentage of individuals expected to respond to at least one of the selected epitopes. Average number of epitope–HLA combinations recognized per individual, reflecting immune response redundancy. PC90 value (PC 90%), indicating the minimal number of epitope–HLA combinations required to cover 90% of the target population.

The results were exported in CSV format and visualized as bar and cumulative distribution plots to illustrate interregional variability in potential T-cell responsiveness. This step provides an estimate of the breadth and universality of the predicted epitope set, allowing prioritization of epitopes with the highest immunological representativeness across human populations.

Conservation Analysis of Immunoglobulin Domain in Bacteria

The conservation of amino acid residues from the immunoglobulin domain across various bacterial species in relation to adalimumab was analyzed using the Rate4Site algorithm on the ConSurf server (version 1.00) [33]. This algorithm calculates position-specific evolutionary rates using an empirical Bayesian approach. The rates are normalized and categorized into 9 grades, with highly conserved residues assigned a score of 9 and highly variable residues receiving a score of 1. The thresholds for these categories are based on the normalized evolutionary rates calculated by Rate4Site. Residues with scores of 1‑3 are considered highly variable, reflecting higher evolutionary rates and frequent mutations or substitutions across species. Conversely, residues with scores of 7 to 9 are classified as highly conserved, indicating minimal variability and strong evolutionary pressure to maintain their structure and function across species. These conservation rates were then visualized using the structural model of adalimumab obtained from the PDB using the Chimera tool [27].

Allergenicity Prediction Using AllerTOP

The potential allergenicity of the predicted T-cell epitopes derived from the light chain of the adalimumab antibody was evaluated using the AllerTOP v.2.0 server [34,35]. This bioinformatics tool applies an alignment-independent approach based on auto- and cross-covariance transformation of protein sequences into uniform-length vectors, followed by machine learning classification using a k-nearest neighbor algorithm trained on a curated dataset of known allergens and nonallergens.

All epitopes predicted by NetMHCIIpan 4.1 as MHC class II binders (rank ≤10%) were used as input sequences in FASTA format. Each peptide was analyzed individually to determine its probability of being classified as "Probable Allergen" or "Probable Non-Allergen," according to the physicochemical descriptors of amino acid residues (hydrophobicity, size, flexibility, and secondary structure propensity).

The results were automatically compared to the training dataset of AllerTOP, which includes more than 2400 allergenic and 2400 nonallergenic proteins, allowing for an indirect homology-free prediction of allergenic potential. The outcomes were exported and tabulated, recording for each peptide (1) the most similar protein identified, (2) the source organism, and (3) the allergenicity classification.

Peptides classified as "Probable Allergen" were further analyzed for potential molecular mimicry with known allergens of plant, fungal, or arthropod origin to assess possible cross-reactivity risks. The combined use of NetMHCIIpan 4.1 and AllerTOP v.2.0 allowed an integrated evaluation of both T-cell immunogenicity and allergenicity potential for the light chain–derived epitopes of adalimumab.

Ethical Considerations

This study was conducted entirely using in silico methods and publicly available, deidentified data. No human participants, personal information, or identifiable records were accessed; therefore, issues of privacy and confidentiality do not apply. No compensation was provided, as no participants, either human or animal subjects, were involved. Consequently, institutional ethical approval and informed consent were not required for this research.

Results

PSI-BLASTp

To explore potential molecular similarities between the monoclonal antibody and bacterial proteins, a PSI-BLASTp search was performed. This analysis was conducted to identify possible cross-reactive epitopes that could contribute to off-target interactions or immunological cross-reactivity, which are relevant for understanding antibody specificity and safety. The search revealed 40 significant matches between the monoclonal sequence and bacterial antigens. From these, 8 sequences corresponding to bacteria of clinical relevance were selected for further analysis (Table 1). Comparative alignment between the adalimumab light chain and the identified bacterial homologs demonstrated an average amino acid identity of 64%. This identity represents amino acid residues that are identical and located in the same position when the sequences are aligned. Higher sequence similarity increases the likelihood of cross-reactivity, as it suggests that the compared molecules may share structurally conserved epitopes capable of being recognized by the same antibodies. The most conserved region was located between residues 74 and 150 (Multimedia Appendix 2), suggesting a potential structural or functional similarity in this segment. The analysis revealed that adalimumab light chain shares sequence homology with a bacterial protein containing an immunoglobulin domain, suggesting possible evolutionary or conformational parallels. Similar results were obtained for the heavy chain, although some alignments involved hypothetical bacterial proteins. Consequently, subsequent analyses focused on the heavy chain and the homologs that could be fully annotated (Table S1 in Multimedia Appendix 3).

Table . PSI-BLASTp results. This table presents the root mean square deviation (RMSD) values, which measure the average structural deviation between aligned molecules.

Bacteria	Antigen	Similarity (%)	Cod genbank	RMSD[a]
Escherichia coli	IDCP[b]	94	HEC3531043·1	0.2
Corynebacterium pyru-viciproducens	IDCP	89.67	WP_280195946·1	0.35
Klebsiella pneumoniae	IDCP	82.86	WP_317090695·1	0.3
Vibrio vulnificus	IDCP	75.00	MCA0777086·1	0.5
Pseudomonas sp.	IDCP	67.80	MCX2891938·1	0.6
Helicobacter pylori	IDCP	62.84	WP_304481324.1	0.5
Salmonella enterica	IDCP	55.56	MBS2599521.1	0.9
Staphylococcus aureus	IDCP	43.84	WP_282719268.1	0.8

[a]RMSD values below 1 Å indicate an exceptionally high degree of structural similarity, often reflecting near-identical alignments. This level of similarity is particularly relevant in studies of molecular mimicry, as it suggests that the structures may share conserved functional or antigenic regions, increasing the likelihood of cross-reactivity. Such low RMSD values underscore the robustness of the alignments and the potential biological significance of the identified matches.

[b]IDCP: immunoglobulin domain–containing protein.

XSL•FO
RenderX

To further characterize the similarities detected through PSI-BLASTp, pairwise alignments were conducted between the adalimumab light chain and the bacterial homologs. This analysis was designed to quantify the degree of sequence conservation and to identify bacterial species exhibiting the closest resemblance to the therapeutic antibody. The results showed identity values ranging from moderate to high

(43.84%‐94%) between adalimumab and the bacterial antigens. The highest sequence conservation was observed with *E. coli* and *Corynebacterium pyruviciproducens* (Table 1 and Figure 1). These findings suggest that certain bacterial proteins share notable similarity with adalimumab, which may be relevant for understanding potential cross-reactive interactions.

Figure 1. Binary alignment of adalimumab light chain with its closest homologs. The binary alignments demonstrated the highest degree of conservation with adalimumab, *Escherichia coli,* and *Corynebacterium pyruviciproducens*, specifically in residues 18‐231 of *E. coli* and residues 28‐241 of *C. pyruviciproducens*. These findings highlight potential regions of molecular mimicry between adalimumab and bacterial proteins.
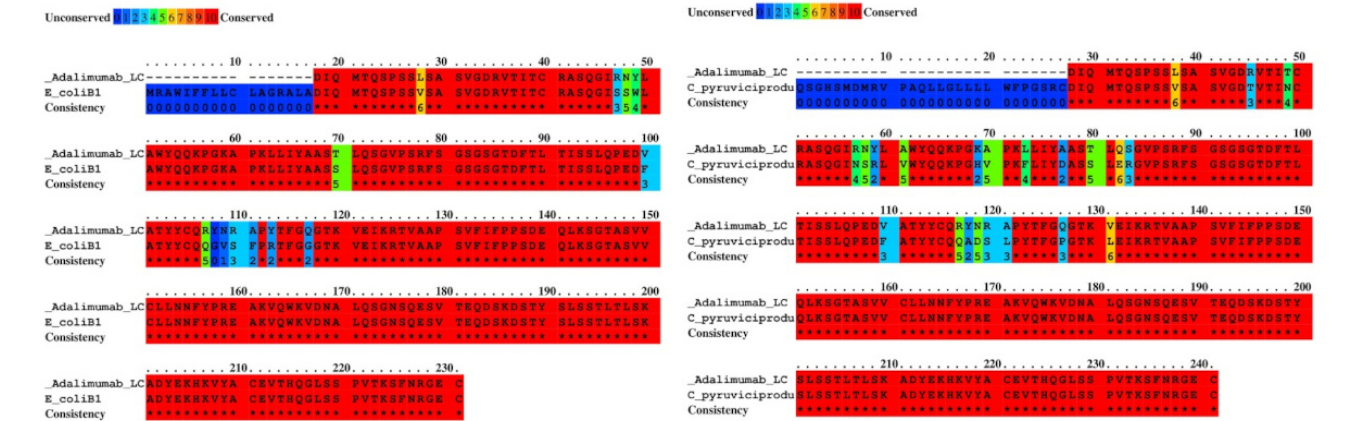
Modeling and Structural Analysis

To evaluate whether the sequence similarities observed translated into comparable 3D conformations, structural models were generated for the bacterial antigens listed in Table 1. This analysis aimed to determine the extent to which these bacterial proteins might adopt folds resembling those of adalimumab, thereby providing structural evidence of potential mimicry. The resulting models showed that the bacterial antigens consistently adopted the characteristic immunoglobulin-like fold (Figure 2). In the cases of *C. pyruviciproducens*, *E. coli*, *Staphylococcus aureus*, and *Vibrio vulnificus*, the predicted structures were organized as dimers. The GMQE scores indicated reliable model quality, with the lowest value (0.70) corresponding to *Salmonella enterica*. This parameter combines information from the sequence alignment and the quality of the structural template to estimate the expected accuracy of the final model; GMQE

values range from 0 to 1, and scores above 0.6 are generally considered indicative of high-confidence structural predictions [26]. Furthermore, RMSD analyses demonstrated a high degree of structural similarity between adalimumab and the bacterial models (Figure 3). RMSD values represent the average atomic distance between 2 superimposed structures, where values below 1 Å suggest a nearly identical spatial organization [36]. In particular, the remarkably low RMSD observed for *E. coli* (0.2 Å) provides compelling evidence of molecular mimicry, as such minimal deviation indicates that both molecules share an almost indistinguishable folding pattern. This structural conservation not only supports the sequence-based similarities but also implies that the bacterial proteins could expose epitopes in a conformation highly compatible with antibody recognition, thereby favoring potential cross-reactivity or recognition from adalimumab antibodies to bacterial proteins and adalimumab.

Figure 2. The 3D models. Structures modeled for bacterial antigens homologous to adalimumab light chain adopted a typical fold of an immunoglobulin domain.

Corynebacterium pyruviciproducens

Escherichia coli

Klebsiella pneumoniae

Pseudomonas species

Helicobacter pylori

Salmonella enterica

Staphylococcus aureus

Vibrio vulnificus
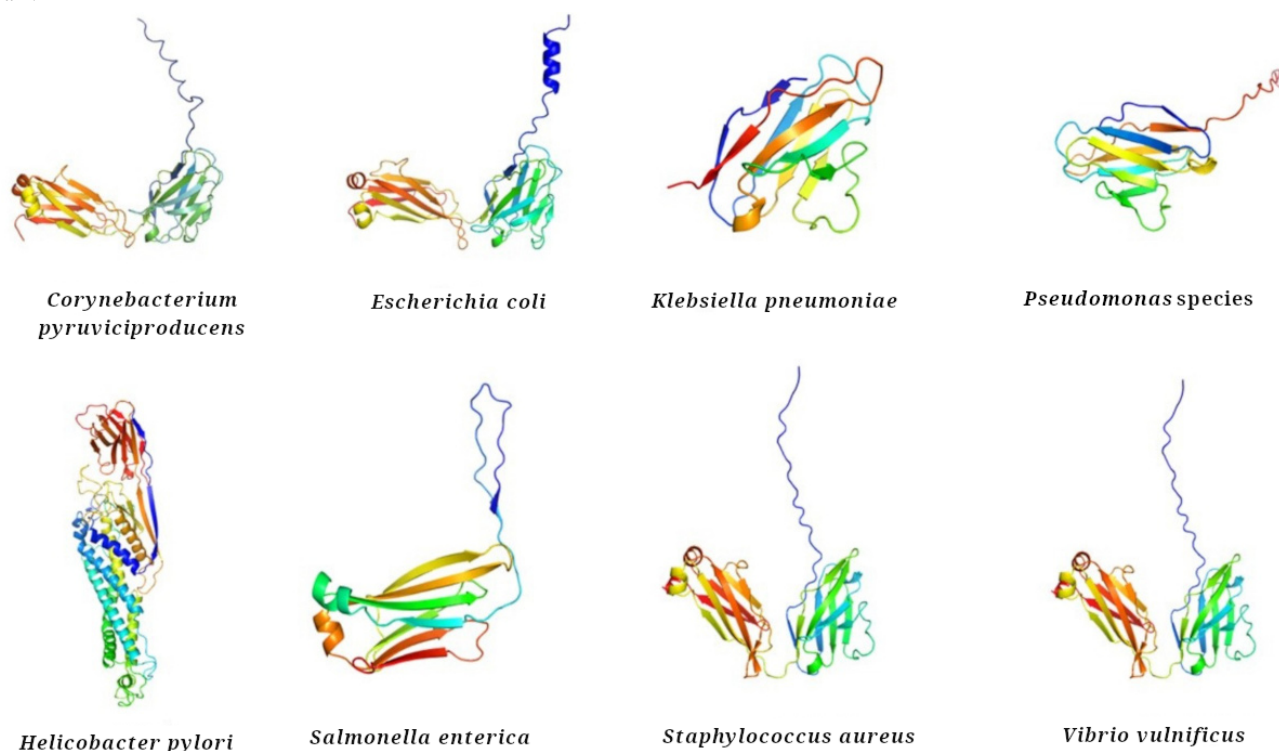
Figure 3. Structural homology. The 3D structure of adalimumab light chain (colored in brown) was superimposed onto each of the modeled structures for bacterial homologs. Analysis revealed a match among the compared structures, indicating a high degree of structural homology.

Corynebacterium pyruviciproducens

Escherichia coli

Klebsiella pneumoniae

Pseudomonas species

Helicobacter pylori

Salmonella enterica

Staphylococcus aureus

Vibrio vulnificus
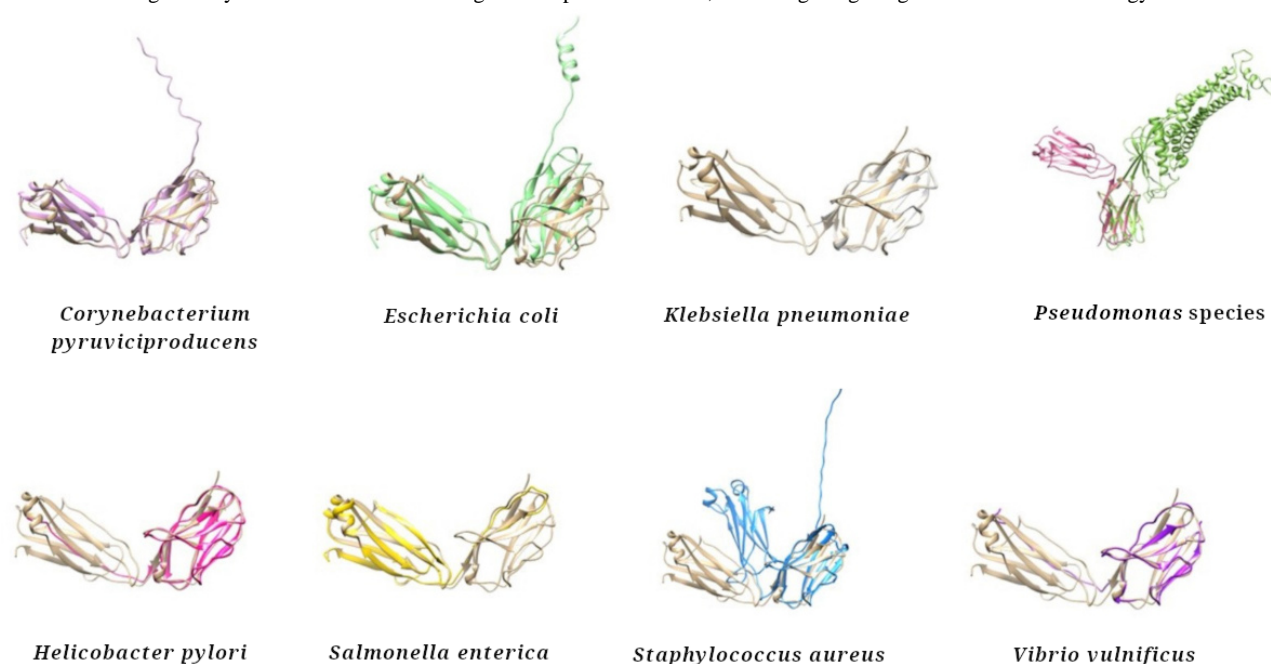
B-Cell Epitope Prediction

To explore whether the sequence similarities between adalimumab and bacterial homologs could translate into shared antigenic regions, an in silico prediction of B-cell epitopes was performed. This computational approach aimed to identify potential linear epitopes within adalimumab that might overlap with conserved bacterial sequences and therefore represent possible cross-reactive sites. The analysis predicted 5 B-cell epitopes in adalimumab showing varying degrees of similarity

to bacterial homologs (Table 2). The predicted epitopes differed in length, with epitopes 1 and 5 containing the largest number of residues, while epitope 2 comprised only 4 residues shared among the bacterial antigens examined. Structural projection of the predicted epitopes revealed that potential cross-reactive regions are distributed across different areas of the antibody surface (Figure 4A). Additionally, surface modeling indicated that these epitopes collectively occupy a substantial portion of the molecular surface (Figure 4B). Although the predicted epitopes ranged from 4 to 16 amino acids, even short conserved

sequences can be relevant for antigen recognition, as complementarity-determining regions within antibodies—often only 6‐20 residues long—are primarily responsible for specific antigen binding [37].

Table . Epitopes predicted on adalimumab to be conserved among bacterial homologs[a].

Epitope	Sequences	Start	End	Residues	Score
1	TLSKADYEKHKV	220	232	12	0.802
2	SSLQ	116	119	4	0.5
3	SGSGTD	105	110	6	0.651
4	SVGDR	51	55	5	0.653
5	HQGLSSPVTKSFN-RGE	240	255	16	0.676

[a]The scores assigned to each predicted epitope range from 0 to 1, with values closer to 1 indicating a stronger prediction. A score of 0.802 for epitope 1, for example, suggests a high level of confidence in its conservation and potential functional relevance across bacterial homologs. The biological relevance of these results suggests that, among the entire antigen sequence, the region corresponding to the predicted epitope is the most likely to be immunogenic and therefore to be recognized by antibodies.

Figure 4. Cross-reactive B-cell epitope prediction. According to Ellipro and multiple alignment tools, 5 linear epitopes are shared between adalimumab light chain and its homologs, which could be implicated in cross-reactivity. (A) A cartoon model illustrates the location of epitopes on the 3D structure of adalimumab. (B) A surface model displays the area occupied by the predicted epitopes.
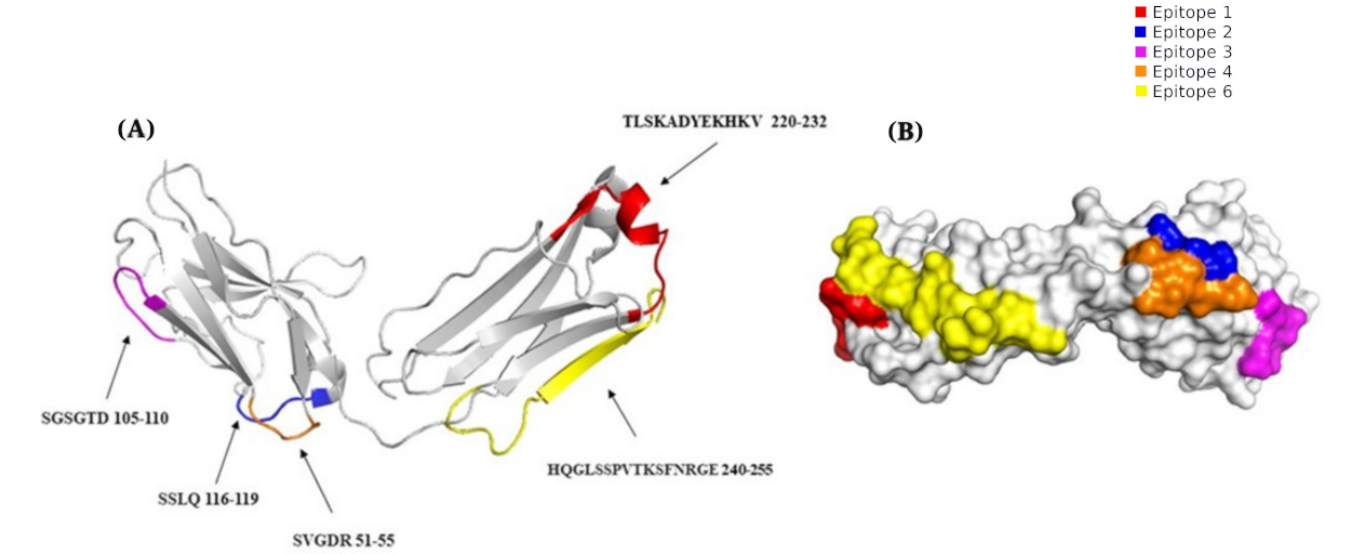
MHC Class II–Dependent T Epitope Prediction

T-cell epitope predictions using the IEDB MHC II binding tool identified multiple adalimumab-derived peptides with strong affinity (percentile ≤2%) for HLA-DRB1 alleles previously associated with anti-adalimumab antibodies. Both heavy- and light-chain regions contained potential CD4[+] T-cell epitopes presented by these risk alleles, supporting a T-cell–dependent mechanism of immunogenicity (Tables S2 and S3 in Multimedia Appendix 3). Screening of the clinically relevant bacterial proteomes revealed peptides with comparable high-affinity binding to the same HLA molecules and partial sequence similarity to adalimumab epitopes (not shown). This overlap suggests that microbial antigens may share HLA-restricted motifs with adalimumab, potentially enabling cross-reactive T-cell responses that contribute to antidrug antibody formation.

PC Analysis

PC analysis performed using the IEDB Population Coverage tool showed that the predicted class II epitopes from the adalimumab light chain could be presented by approximately 15.2% of the European population (Multimedia Appendix 4 and Table S4 in Multimedia Appendix 3). The mean number of epitope–HLA combinations recognized per individual was 3.19 (SD), with a PC90 value of 2.48. These results indicate that a limited yet notable fraction of the population carries HLA alleles capable of presenting these epitopes, supporting the existence of potential interindividual variability in T-cell responsiveness to adalimumab.

Conservation Analysis of Immunoglobulin Domain in Bacteria

To evaluate whether the structural similarities between adalimumab and bacterial homologs reflect conserved evolutionary features, a conservation analysis of the immunoglobulin domain was performed using the ConSurf server. This analysis aimed to identify amino acid residues that are evolutionarily conserved across bacterial antigens and the antibody, which could indicate the maintenance of structural or functional motifs important for protein stability or interaction.

The results, summarized in Figure 5, highlight conserved residues mapped onto the amino acid sequence and the corresponding 3D structures. The cartoon representations illustrate that several regions of the immunoglobulin domain remain highly conserved among the analyzed bacterial species. The conservation score gradient, represented by ConSurf's color scale from cyan (variable=grade 1) to purple (highly conserved=grade 9), emphasizes that key residues within the core of the domain exhibit strong conservation, suggesting evolutionary pressure to maintain structural integrity in these regions.

Figure 5. The conservation analysis of individual amino acids in adalimumab was conducted using the ConSurf server. (A) Display of the amino acid sequence, highlighting evolutionarily conserved residues of immunoglobulin domains of the bacterial antigens in adalimumab. (B) and (C) are cartoon models that illustrate the conserved regions on the 3D structure of adalimumab, demonstrating that its amino acid sequence and structure are preserved across different species. The degrees of conservation were mapped onto the sequence and structure, employing the ConSurf color-coding scheme, where shades range from cyan (representing variable, grade 1) to purple (indicating highly conserved, grade 9) positions.
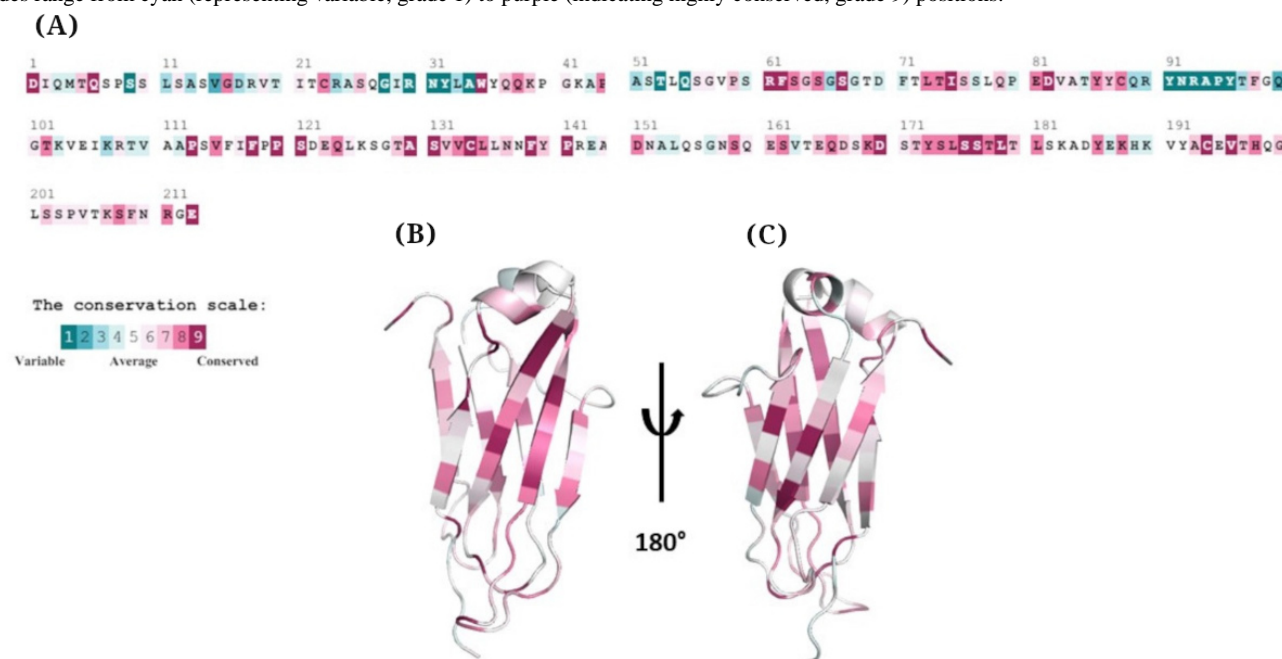
Allergenicity Prediction Using AllerTOP

Allergenicity analysis performed using the AllerTOP server revealed that several of the predicted adalimumab-derived peptides share sequence similarity with known or probable allergens. Notably, matches were identified with allergenic proteins from *Sarcoptes scabiei* (Sarc s 1), *Artemisia vulgaris* (Art v 3), and *Malassezia sympodialis*, as well as with plant-derived allergens such as hydroxyproline-rich glycoproteins from *Oryza sativa* and neoxanthin synthase from *Solanum tuberosum*. Some peptides also showed similarity to human proteins, including ATP synthase F1 assembly factor 2 and B-cell lymphoma 6 protein, suggesting potential immunological cross-reactivity. Overall, the presence of sequences with predicted allergenic properties indicates that these epitopes may elicit immune recognition and could contribute to the immunogenic potential of adalimumab.

Discussion

Principal Findings

This study identified several bacterial antigens that share significant sequence and structural similarities with adalimumab, particularly within immunoglobulin-like domains. Using in silico analyses, we found that these bacterial proteins exhibit conserved folds and low RMSD values relative to adalimumab, supporting the hypothesis of molecular mimicry. Such a resemblance may provide a mechanistic explanation for the development of ADAs and the loss of therapeutic response observed in some patients treated with adalimumab. To our knowledge, this is the first study to propose a mechanism of cross-reactivity between adalimumab and microbial antigens supported by in silico structural and sequence evidence. Moreover, it is the first report describing specific bacterial epitopes with potential clinical relevance in the context of adalimumab immunogenicity.

Up to 60% of patients who receive adalimumab eventually show a decline in the effectiveness of the treatment [4]. Therapeutic failure is thought to be caused by ADAs, but the reasons for its formation are still unknown. Some past infections may have contributed to the development of ADAs [15]. In this setting, molecular mimicry could explain why prior infections produce these antibodies. Molecular mimicry may cause antibodies to be generated during past infections to inadvertently neutralize adalimumab by cross-reacting with its epitopes, diminishing its therapeutic effect. So, we propose using bioinformatics to investigate this phenomenon. In support of this hypothesis, our in silico T-cell epitope predictions identified adalimumab peptides with strong HLA-DRB1 binding—alleles previously linked to anti-adalimumab antibodies—and revealed bacterial peptides capable of binding the same HLA molecules, suggesting a potential cross-reactive, T-cell–mediated mechanism underlying ADA formation.

To establish a link between certain microbes and molecular mimicry, we must follow a 4-tiered evidence approach [2,7]. Previously, an epidemiological connection was established [15]. Now, using in silico methods, we have identified potential antigens with significant identity, indicating possible mimicry. Although the analyzed fragments are relatively short, they were selected from clinically relevant bacteria due to their potential involvement in molecular mimicry processes. Information regarding their immunodominant regions is currently unavailable, as this is the first study addressing these specific bacterial-antibody similarities. This finding suggests that clinically relevant bacteria, including *E. coli*, *C. pyruviciproducens*, *Klebsiella pneumoniae*, *V. vulnificus*, *Pseudomonas aeruginosa*, *Helicobacter pylori*, *S. enterica*, and *S. aureus*, may harbor molecular motifs capable of inducing cross-reactivity. These microorganisms are of major medical concern as they encompass commensals with pathogenic potential, such as *E. coli* [38]; opportunistic pathogens linked to severe infections like *C. pyruviciproducens* [39,40]; and multidrug-resistant strains such as *K. pneumoniae* [41,42]. Others, including *V. vulnificus* and *P. aeruginosa*, are associated with life-threatening conditions such as necrotizing soft-tissue infections and ventilator-associated pneumonia, respectively [43,44]. Moreover, *H. pylori*, *S. enterica*, and *S. aureus* contribute to chronic and systemic diseases ranging from gastritis and typhoid fever to osteomyelitis and sepsis [45-47]. In addition, population coverage analysis using the IEDB tool showed that the predicted class II epitopes from the adalimumab light chain could be presented by approximately 15% of the European population, suggesting that only a subset of individuals possess HLA alleles capable of recognizing these epitopes and may therefore be more prone to T-cell–mediated immunogenic responses.

The observed sequence similarity within immunoglobulin-like domains highlights a plausible mechanism through which antibodies targeting microbial antigens could recognize adalimumab epitopes, potentially impairing its therapeutic efficacy. This warrants further experimental validation through in vitro inhibition assays and in vivo studies to confirm the immunological and clinical significance of these mimicry events. These predictions also provide candidate regions for in vitro validation, enabling the refinement of epitope mapping through assays such as enzyme-linked immunosorbent assay or peptide-binding tests. Additionally, the complete amino acid sequences of the bacterial antigens are available for future recombinant cloning and seroprevalence studies in adalimumab-treated patients.

In addition, allergenicity prediction using the AllerTOP server indicated that several adalimumab-derived peptides share sequence similarity with known or probable allergens from diverse sources, including *S. scabiei*, *A. vulgaris*, and *O. sativa*. This overlap suggests that certain epitopes may possess inherent immunostimulatory properties or cross-reactivity with environmental allergens, further supporting their potential contribution to the immunogenic profile of adalimumab.

Our results have cast a new light on the intricate role of molecular mimicry in drug development, merging concepts from diverse therapeutic domains. Historically, molecular structures resembling those of microbial antigens have been exploited to elicit immunological benefits, enhancing the body's defenses against diseases and improving vaccine efficacy, and even cancer response [48-51]. However, our findings underscore a more intricate reality where these structural analogies, notably those present in adalimumab, bear the potential to induce ADA development. This duality reflects broader pharmaceutical experiences, where beneficial immunogenic mimicry can, in some contexts, inadvertently lead to immune cross-reactions with conditions like thrombocytopenia and infectious and autoimmune diseases [5,10-14]. The resulting immunological disarray, wherein the body cannot differentiate between therapeutic agents and microbial antigens, might trigger an unwarranted immune response against the host's own tissues, manifesting in a range of adverse clinical outcomes such as autoimmune disease relapses, serum sickness, hypersensitivity reactions, or symptoms of autoimmune disease. Thus, our study bridges the existing knowledge gap by revealing how the same molecular resemblances that have been leveraged for therapeutic gain may also carry risks that must be carefully navigated in the continuum of drug research and patient care.

Unraveling the connections of shared molecular similarities between drugs and various microorganisms—pathogens and commensals alike—can illuminate potential reasons behind the unintended effects some medications might have on the immune system's response to treatment. Specifically, within the therapeutic framework, molecular mimicry may introduce complexities in treatment modalities due to its potential to elicit aberrant immune responses. Such phenomena can attenuate the efficacy of the medication or precipitate adverse immunological reactions through ADA development. We also consider it may be beneficial to adjust the drug's formulation and design to lessen its immunogenic potential while maintaining therapeutic effectiveness, a task that is complex yet crucial [52].

In the realm of precision medicine, the significance of molecular mimicry between monoclonal antibodies and microorganisms is profound. This study highlights the presence of antigens with immunoglobulin domains in microbes such as *E. coli*, *H. pylori*, and *S. aureus*, common colonizers [53-55], as well as in those linked to severe diseases, including *K. pneumoniae*, *V. vulnificus*, *S. enterica*, *C. pyruviciproducens*, and *Pseudomonas* species [40,41,43,56,57]. These immunoglobulin domains, which are involved in diverse binding and molecular recognition processes, have been identified across a spectrum of functional groups, including molecular transport, morphoregulation, and cell adhesion to virus receptors, shape recognition, and toxin neutralization [8,9]. The remarkable functional versatility of the immunoglobulin superfamily extends to cell phenotype markers and regulators of gene transcription, among others. Thus, the structural and functional parallels found in this study underscore the need for careful consideration of molecular mimicry in drug development, particularly in the design of monoclonal antibodies, due to their potential to elicit unintended immune responses or interfere with microbial commensals critical to human health.

It is important to acknowledge the limitations of our study. The actual structures may differ from the models we propose, as in silico modeling and epitope prediction analyses are not

definitive. Nevertheless, bioinformatic approaches offer significant advantages by efficiently guiding research efforts [27]. They play a crucial role in the initial evaluation of hypotheses, helping determine whether further in vitro studies are warranted. Additionally, bacteria may retain remnants of vectors through horizontal and vertical transmission, whether occurring naturally or during biotechnological processes. However, since not all the bacteria analyzed are associated with biotechnology, any bacterial contamination, if present, is more likely to result from natural processes and evolutionary mechanisms. Moreover, given that all the bacterial antigens studied contain an immunoglobulin-like domain—a feature widely reported across various organisms and particularly in enterobacteria [8]—we consider it unlikely that these findings are due to contamination.

While in silico predictions provide valuable preliminary insights, we acknowledge that they cannot fully confirm the occurrence of molecular mimicry in a biological context. Nonetheless, the results provide a rational basis for the design of forthcoming serological studies, guiding the selection of candidate antigens and regions with potential cross-reactivity. This work represents a first step toward understanding the mechanisms underlying cross-reactive immune responses between microbial antigens and therapeutic antibodies. The high degree of structural similarity observed suggests a biologically plausible mechanism of cross-reactivity, which we plan to further investigate through in vitro validation assays. In addition, we acknowledge that immunogenicity is a multifactorial process influenced not only by molecular mimicry but also by protein-specific properties, manufacturing conditions, and patient-related factors. Nonetheless, molecular mimicry remains one of the least explored contributors to therapeutic antibody immunogenicity, and our findings provide the first evidence-supported hypothesis proposing this mechanism in the context of adalimumab.

As we look toward future pharmaceutical innovation, the insights gained from this inquiry advocate for an integrated approach. This approach should encompass a thorough investigation of the interplay between drugs, the human microbiome, and pathogenic microorganisms. By doing so, we can strive to harness the positive aspects of molecular mimicry while mitigating its risks, thereby advancing the field of medicine with a more informed and cautious perspective.

Conclusions

In conclusion, examining adalimumab's structural similarities with key microorganisms such as bacteria offers a nuanced perspective on molecular mimicry's dual role in medicine. While its utility in enhancing therapeutic benefits is established, we urge a critical reevaluation based on our findings that raise the possibility of adverse immune reactions due to ADAs. Our results also point to the need to advance in the confirmation through in vitro and in vivo tests of this cross-reactivity, because this would make it a necessary and judicious approach to drug design, incorporating an integrated analysis of drug-pathogen-microbiome interactions to safeguard therapeutic efficacy and patient health.

Data Availability

Additional data is available from the corresponding author on request.

Authors' Contributions

DIP-S, AD-L-T, and MM conceptualized the study. DIP-S and MM curated the data. Formal analysis was performed by DIP-S, MM, GM-S, AS, and AD-L-T. Investigation was carried out by DIP-S, MM, AS, and AD-L-T. Methodology development was conducted by DIP-S, MM, GM-S, and AD-L-T. Validation was performed by AS and OC. Visualization was prepared by DIP-S. The original draft was written by DIP-S, while DIP-S, MM, GM-S, AS, AD-L-T, and OC contributed to reviewing and editing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Computational workflow.
[PNG File, 299 KB - bioinform_v6i1e83872_app1.png]

Multimedia Appendix 2
Multiple alignment between adalimumab light chain and homologs. The multiple alignments revealed a 64% identity among the compared amino acid sequences. The highest conservation is observed between residues 71 and 159.
[PNG File, 855 KB - bioinform_v6i1e83872_app2.png]

Multimedia Appendix 3
PSI-BLASTp results and predicted class II T-cell epitopes for adalimumab chains, including HLA-DRB1*11:01–restricted peptides and projected population coverage.
[DOCX File, 26 KB - bioinform_v6i1e83872_app3.docx]

Multimedia Appendix 4
Population coverage analysis for class II epitopes predicted from the adalimumab light chain using the Immune Epitope Database (IEDB) population coverage tool. The bar and cumulative distribution plots represent the percentage of individuals in the European population predicted to recognize at least 1 human leukocyte antigen (HLA)–epitope combination, indicating an estimated coverage of approximately 15%.
[PNG File, 257 KB - bioinform_v6i1e83872_app4.png]

References

1. Ellis CR, Azmat CE. Adalimumab: StatPearls Publishing; 2024. URL: http://www.ncbi.nlm.nih.gov/books/NBK557889/ [accessed 2024-02-26]
2. Thaper D, Prabha V. Molecular mimicry: an explanation for autoimmune diseases and infertility. Scand J Immunol 2018 Aug;88(2):e12697. [doi: 10.1111/sji.12697] [Medline: 29935034]
3. Strand V, Gonçalves J, Hickling TP, Jones HE, Marshall L, Isaacs JD. Immunogenicity of biosimilars for rheumatic diseases, plaque psoriasis, and inflammatory bowel disease: a review from clinical trials and regulatory documents. BioDrugs 2020 Feb;34(1):27-37. [doi: 10.1007/s40259-019-00394-x] [Medline: 31721107]
4. Cludts I, Spinelli FR, Morello F, Hockley J, Valesini G, Wadhwa M. Anti-therapeutic antibodies and their clinical impact in patients treated with the TNF antagonist adalimumab. Cytokine 2017 Aug;96:16-23. [doi: 10.1016/j.cyto.2017.02.015] [Medline: 28279855]
5. Segal Y, Shoenfeld Y. Vaccine-induced autoimmunity: the role of molecular mimicry and immune crossreaction. Cell Mol Immunol 2018 Jun;15(6):586-594. [doi: 10.1038/cmi.2017.151] [Medline: 29503439]
6. Kanduc D. Peptide cross-reactivity: the original sin of vaccines. Front Biosci (Schol Ed) 2012 Jun;4(4):1393-1401. [doi: 10.2741/s341] [Medline: 22652881]
7. Cusick MF, Libbey JE, Fujinami RS. Molecular mimicry as a mechanism of autoimmune disease. Clin Rev Allergy Immunol 2012 Feb;42(1):102-111. [doi: 10.1007/s12016-011-8294-7] [Medline: 22095454]
8. Bodelón G, Palomino C, Fernández L. Immunoglobulin domains in Escherichia coli and other enterobacteria: from pathogenesis to applications in antibody technologies. FEMS Microbiol Rev 2013 Mar;37(2):204-250. [doi: 10.1111/j.1574-6976.2012.00347.x] [Medline: 22724448]
9. Hüttener M, Hergueta J, Bernabeu M, et al. Roles of proteins containing immunoglobulin-like domains in the conjugation of bacterial plasmids. mSphere 2022 Feb;7(1):e0097821. [doi: 10.1128/msphere.00978-21] [Medline: 34986320]
10. Dotan A, Shoenfeld Y. Perspectives on vaccine induced thrombotic thrombocytopenia. J Autoimmun 2021 Jul;121:102663. [doi: 10.1016/j.jaut.2021.102663] [Medline: 34020254]
11. Bozkurt B, Kamat I, Hotez PJ. Myocarditis with COVID-19 mRNA vaccines. Circulation 2021 Aug;144(6):471-484. [doi: 10.1161/CIRCULATIONAHA.121.056135] [Medline: 34281357]
12. Garg RK, Paliwal VK. Spectrum of neurological complications following COVID-19 vaccination. Neurol Sci 2022 Jan;43(1):3-40. [doi: 10.1007/s10072-021-05662-9] [Medline: 34719776]
13. Bragazzi NL, Bridgewood C, Sharif K, et al. HPV vaccines and lupus: current approaches towards preventing adverse immune cross-reactivity. Expert Rev Vaccines 2019 Jan;18(1):31-42. [doi: 10.1080/14760584.2019.1557519] [Medline: 30526148]
14. Hausmann OV, Seitz M, Villiger PM, Pichler WJ. The complex clinical picture of side effects to biologicals. Med Clin North Am 2010 Jul;94(4):791-804. [doi: 10.1016/j.mcna.2010.03.001] [Medline: 20609863]
15. Hässler S, Bachelet D, Duhaze J, et al. Clinicogenomic factors of biotherapy immunogenicity in autoimmune disease: a prospective multicohort study of the ABIRISK consortium. PLoS Med 2020 Oct;17(10):e1003348. [doi: 10.1371/journal.pmed.1003348] [Medline: 33125391]
16. Knox C, Law V, Jewison T, et al. DrugBank 3.0: a comprehensive resource for "OMICS" research on drugs. Nucleic Acids Res 2011 Jan;39(Database issue):D1035-D1041. [doi: 10.1093/nar/gkq1126] [Medline: 21059682]
17. Guzenko D, Burley SK, Duarte JM. Real time structural search of the Protein Data Bank. PLoS Comput Biol 2020 Jul;16(7):e1007970. [doi: 10.1371/journal.pcbi.1007970] [Medline: 32639954]
18. Hu G, Kurgan L. Sequence similarity searching. Curr Protoc Protein Sci 2019 Feb;95(1):e71. [doi: 10.1002/cpps.71] [Medline: 30102464]
19. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL. NCBI BLAST: a better web interface. Nucleic Acids Res 2008 Jul 1;36:W5-W9. [doi: 10.1093/nar/gkn201] [Medline: 18440982]
20. PRALINE multiple sequence alignment. Center for Integrative Bioinformatics. URL: https://www.ibi.vu.nl/programs/pralinewww [accessed 2025-04-25]

21.   Protein BLAST: search protein databases using a protein query. National Center for Biotechnology Information. URL:
        https://blast.ncbi.nlm.nih.gov/Blast.
        cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&BLAST_SPEC=&LINK_LOC=blasttab&LAST_PAGE=blastx
        [accessed 2024-12-18]

22.   BLAST statistics: the expect value. National Center for Biotechnology Information. URL: https://www.nlm.nih.gov/ncbi/
        workshops/2023-08_BLAST_evol/e_value.html [accessed 2024-12-12]

23.   Research Collaboratory for Structural Bioinformatics Protein Data Bank. URL: https://www.rcsb.org [accessed 2025-12-02]

24.   Wheeler DL, Barrett T, Benson DA, et al. Database resources of the National Center for Biotechnology Information. Nucleic
        Acids Res 2006 Jan;34:D173-D180. [doi: 10.1093/nar/gkj158] [Medline: 16381840]

25.   Mooers BHM. Simplifying and enhancing the use of PyMOL with horizontal scripts. Protein Sci 2016 Oct;25(10):1873-1882.
        [doi: 10.1002/pro.2996] [Medline: 27488983]

26.   Waterhouse A, Bertoni M, Bienert S, et al. SWISS-MODEL: homology modelling of protein structures and complexes.
        Nucleic Acids Res 2018 Jul 2;46(W1):W296-W303. [doi: 10.1093/nar/gky427] [Medline: 29788355]

27.   Pettersen EF, Goddard TD, Huang CC, et al. UCSF Chimera: a visualization system for exploratory research and analysis.
        J Comput Chem 2004 Oct;25(13):1605-1612. [doi: 10.1002/jcc.20084] [Medline: 15264254]

28.   Baker D, Sali A. Protein structure prediction and structural genomics. Science 2001 Oct;294(5540):93-96. [doi:
        10.1126/science.1065659] [Medline: 11588250]

29.   Ponomarenko J, Bui HH, Li W, et al. ElliPro: a new structure-based tool for the prediction of antibody epitopes. BMC
        Bioinformatics 2008 Dec 2;9:514. [doi: 10.1186/1471-2105-9-514] [Medline: 19055730]

30.   ElliPro: antibody epitope prediction. Immune Epitope Database. URL: http://tools.iedb.org/ellipro/ [accessed 2024-12-18]

31.   MHC-II binding predictions. IEDB Analysis Resource. URL: https://tools.iedb.org/mhcii/ [accessed 2025-12-02]

32.   Population coverage. IEDB Analysis Resource. URL: https://tools.iedb.org/population/ [accessed 2025-12-02]

33.   Ashkenazy H, Abadi S, Martz E, et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary
        conservation in macromolecules. Nucleic Acids Res 2016 Jul 8;44(W1):W344-W350. [doi: 10.1093/nar/gkw408] [Medline:
        27166375]

34.   AllerTOP. Drug Design and Bioinformatics Lab. URL: https://www.ddg-pharmfac.net/AllerTOP [accessed 2025-12-02]

35.   Dimitrov I, Bangov I, Flower DR, Doytchinova I. AllerTOP v.2—a server for in silico prediction of allergens. J Mol Model
        2014 Jun;20(6). [doi: 10.1007/s00894-014-2278-5] [Medline: 24878803]

36.   Holm L. Using Dali for protein structure comparison. In: Gáspári Z, editor. Structural Bioinformatics: Methods and Protocols:
        Humana; 2020:29-42. [doi: 10.1007/978-1-0716-0270-6_3]

37.   Svilenov HL, Sacherl J, Protzer U, Zacharias M, Buchner J. Mechanistic principles of an ultra-long bovine CDR reveal
        strategies for antibody design. Nat Commun 2021 Nov;12(1):6737. [doi: 10.1038/s41467-021-27103-z] [Medline: 34795299]

38.   Rendón MA, Saldaña Z, Erdem AL, et al. Commensal and pathogenic Escherichia coli use a common pilus adherence
        factor for epithelial cell colonization. Proc Natl Acad Sci U S A 2007 Jun 19;104(25):10637-10642. [doi:
        10.1073/pnas.0704104104] [Medline: 17563352]

39.   Tong J, Liu C, Summanen PH, Xu H, Finegold SM. Corynebacterium pyruviciproducens sp. nov., a pyruvic acid producer.
        Int J Syst Evol Microbiol 2010 May;60(Pt 5):1135-1140. [doi: 10.1099/ijs.0.011783-0] [Medline: 19666798]

40.   Fernandez LV, Fortuny AS, Rodriguez EF. Mastitis por Corynebacterium pyruviciproducens y Corynebacterium amycolatum
        en mujeres inmunocompetentes no lactantes. Rev Argent Microbiol 2021 Jan;53(1):39-42. [doi: 10.1016/j.ram.2020.06.006]
        [Medline: 32718823]

41.   Arato V, Raso MM, Gasperini G, Berlanda Scorza F, Micoli F. Prophylaxis and treatment against Klebsiella pneumoniae:
        current insights on this emerging anti-microbial resistant global threat. Int J Mol Sci 2021 Apr 14;22(8):4042. [doi:
        10.3390/ijms22084042] [Medline: 33919847]

42.   Li D, Huang X, Rao H, et al. Klebsiella pneumoniae bacteremia mortality: a systematic review and meta-analysis. Front
        Cell Infect Microbiol 2023;13:1157010. [doi: 10.3389/fcimb.2023.1157010] [Medline: 37153146]

43.   Yun NR, Kim DM. Vibrio vulnificus infection: a persistent threat to public health. Korean J Intern Med 2018
        Nov;33(6):1070-1078. [doi: 10.3904/kjim.2018.159] [Medline: 29898575]

44.   Reynolds D, Kollef M. The epidemiology and pathogenesis and treatment of Pseudomonas aeruginosa infections: an update.
        Drugs (Abingdon Engl) 2021 Dec;81(18):2117-2131. [doi: 10.1007/s40265-021-01635-6] [Medline: 34743315]

45.   Mladenova I. Clinical relevance of Helicobacter pylori infection. JCM 2021;10(16):3473. [doi: 10.3390/jcm10163473]
        [Medline: 34441769]

46.   Galán-Relaño Á, Valero Díaz A, Huerta Lorenzo B, et al. Salmonella and salmonellosis: an update on public health
        implications and control strategies. Animals (Basel) 2023 Nov 27;13(23):3666. [doi: 10.3390/ani13233666] [Medline:
        38067017]

47.   Touaitia R, Mairi A, Ibrahim NA, Basher NS, Idres T, Touati A. Staphylococcus aureus: a review of the pathogenesis and
        virulence mechanisms. Antibiotics (Basel) 2025 May 6;14(5):470. [doi: 10.3390/antibiotics14050470] [Medline: 40426537]

48.   Bondareva M, Budzinski L, Durek P, et al. Cross-regulation of antibody responses against the SARS-CoV-2 spike protein
        and commensal microbiota via molecular mimicry. Cell Host Microbe 2023 Nov 8;31(11):1866-1881. [doi:
        10.1016/j.chom.2023.10.007] [Medline: 37944493]

49. Li N, Zhang G, Chen Y, et al. Identification of the mimotopes within the major capsid protein L1 of human papillomavirus types 18 and 45, using neutralizing monoclonal antibodies. Int J Biol Macromol 2021 Mar;174:587-595. [doi: 10.1016/j.ijbiomac.2021.01.137] [Medline: 33493568]

50. Horwacik I, Golik P, Grudnik P, et al. Structural basis of GD2 ganglioside and mimetic peptide recognition by 14G2a antibody. Mol Cell Proteomics 2015 Oct;14(10):2577-2590. [doi: 10.1074/mcp.M115.052720] [Medline: 26179345]

51. Gilchuk P, Murin CD, Milligan JC, et al. Analysis of a therapeutic antibody cocktail reveals determinants for cooperative and broad ebolavirus neutralization. Immunity 2020 Feb 18;52(2):388-403. [doi: 10.1016/j.immuni.2020.01.001] [Medline: 32023489]

52. Pham NB, Meng WS. Protein aggregation and immunogenicity of biotherapeutics. Int J Pharm 2020 Jul;585:119523. [doi: 10.1016/j.ijpharm.2020.119523] [Medline: 32531452]

53. Martinson JNV, Walk ST. Escherichia coli residency in the gut of healthy human adults. EcoSal Plus 2020 Sep;9(1):10. [doi: 10.1128/ecosalplus.ESP-0003-2020] [Medline: 32978935]

54. Maixner F, Krause-Kyora B, Turaev D, et al. The 5300-year-old Helicobacter pylori genome of the Iceman. Science 2016 Jan 8;351(6269):162-165. [doi: 10.1126/science.aad2545] [Medline: 26744403]

55. Gherardi G. Staphylococcus aureus infection: pathogenesis and antimicrobial resistance. Int J Mol Sci 2023 May 3;24(9):8182. [doi: 10.3390/ijms24098182] [Medline: 37175886]

56. Gal-Mor O. Persistent infection and long-term carriage of typhoidal and nontyphoidal salmonellae. Clin Microbiol Rev 2019 Jan;32(1):e00088-18. [doi: 10.1128/CMR.00088-18] [Medline: 30487167]

57. Qin S, Xiao W, Zhou C, et al. Pseudomonas aeruginosa: Pathogenesis, virulence factors, antibiotic resistance, interaction with host, technology advances and emerging therapeutics. Sig Transduct Target Ther 2022;7(1):199. [doi: 10.1038/s41392-022-01056-1]

Abbreviations

ADA: antidrug antibody
GMQE: Global Model Quality Estimate
HLA: human leukocyte antigen
IDCP: immunoglobulin domain-containing protein
IEDB: Immune Epitope Database
MHC: major histocompatibility complex
PC: population coverage
PDB: Protein Data Bank
PI: protrusion index
RMSD: root mean square deviation

XSL•FO
RenderX

Optimizing Feature Selection and Machine Learning Algorithms for Early Detection of Prediabetes Risk: Comparative Study

Mahmoud B Almadhoun; MA Burhanuddin

Fakulti Kecerdasan Buatan dan Keselamatan Siber, Universiti Teknikal Malaysia, Melaka, Durian Tunggal, Malaysia

Corresponding Author:
MA Burhanuddin
Fakulti Kecerdasan Buatan dan Keselamatan Siber, Universiti Teknikal Malaysia, Melaka, Durian Tunggal, Malaysia

Abstract

Background: Prediabetes is an intermediate stage between normal glucose metabolism and diabetes and is associated with increased risk of complications like cardiovascular disease and kidney failure.

Objective: It is crucial to recognize individuals with prediabetes early in order to apply timely intervention strategies to decelerate or prohibit diabetes development. This study aims to compare the effectiveness of machine learning (ML) algorithms in predicting prediabetes and identifying its key clinical predictors.

Methods: Multiple ML models are evaluated in this study, including random forest, extreme gradient boosting (XGBoost), support vector machine (SVM), and k-nearest neighbors (KNNs), on a dataset of 4743 individuals. For improved performance and interpretability, key clinical features were selected using LASSO (Least Absolute Shrinkage and Selection Operator) regression and principal component analysis (PCA). To optimize model accuracy and reduce overfitting, we used hyperparameter tuning with RandomizedSearchCV for XGBoost and random forest, and GridSearchCV for SVM and KNN. SHAP (Shapley Additive Explanations) was used to assess model-agnostic feature importance. To resolve data imbalance, SMOTE (Synthetic Minority Oversampling Technique) was applied to ensure reliable classifications.

Results: A cross-validated ROC-AUC (receiver operating characteristic area under the curve) score of 0.9117 highlighted the robustness of random forest in generalizing across datasets among the models tested. XGBoost followed closely, providing balanced accuracy in distinguishing between normal and prediabetic cases. While SVMs and KNNs performed adequately as baseline models, they exhibited limitations in sensitivity. The SHAP analysis indicated that BMI, age, high-density lipoprotein cholesterol, and low-density lipoprotein cholesterol emerged as the key predictors across models. The performance was significantly enhanced through hyperparameter tuning; for example, the ROC-AUC for SVM increased from 0.813 (default) to 0.863 (tuned). PCA kept 12 components while maintaining 95% of the variance in the dataset.

Conclusions: It is demonstrated in this research that optimized ML models, especially random forest and XGBoost, are effective tools for assessing early prediabetes risk. Combining SHAP analysis with LASSO and PCA enhances transparency, supporting their integration in real-time clinical decision support systems. Future directions include validating these models in diverse clinical settings and integrating additional biomarkers to improve prediction accuracy, offering a promising avenue for early intervention and personalized treatment strategies in preventive health care.

Introduction

A prediabetic state is characterized by elevated blood sugar levels, considered as an intermediate stage between normal glucose metabolism and type 2 diabetes [1]. In individuals with a high risk of diabetes, cardiovascular disease, and kidney complications, early diagnosis and intervention in prediabetes is important for delaying or preventing progression to diabetes [2]. In spite of lifestyle interventions, adherence remains one of the biggest challenges, which necessitates early and accurate detection.

While biochemical markers like fasting glucose and glycated hemoglobin are valuable, they may not capture the full spectrum of prediabetes risk factors, resulting in missed diagnoses and delayed interventions. To address this, a wide set of predictors, including clinical and genetic data, needs to be incorporated. This issue can be overcome by machine learning (ML), which can analyze complex relationships between a broad range of biomarkers [3]. By leveraging ML algorithms, this study aims to enhance the accuracy of prediabetes risk assessment and early detection.

A feature selection technique such as LASSO (Least Absolute Shrinkage and Selection Operator) regression and principal component analysis (PCA) further optimizes these models by focusing on the most apropos predictors, as a consequence improving both efficiency and interpretability [4,5]. Additionally, it reduces model complexity and boosts prediction accuracy by eliminating irrelevant or unnecessary data in ML. Models based on the most impactful clinical features, such as BMI, age, low-density lipoprotein cholesterol (LDL-C), and high-density lipoprotein cholesterol (HDL-C), can capture underlying patterns linked with prediabetes [6].

This paper assesses and compares the predictive power of various ML algorithms such as random forest, extreme gradient boosting (XGBoost), support vector machine (SVM), and k-nearest neighbors (KNNs), inclusive of feature selection methods such as LASSO and PCA. We aim to identify the most effective model and feature selection technique for the detection of early prediabetes, ultimately contributing to highly accurate diagnostics and personalized prevention.

In this study, key predictors such as BMI, age, LDL-C, and HDL-C were identified, which may refine diagnostic criteria and help with targeted prevention. The findings emphasize the capability for ML-based tools to improve prediabetes management and foster better patient outcomes through early intervention.

Various ML models have been used in recent studies to enhance detection accuracy and identify key risk factors associated with prediabetes progression. These approaches underscore the potential of ML in developing effective and clinically applicable prediction models for prediabetes risk.

An important direction is using ensemble and decision tree–based models to predict prediabetes. A study by Liu et al [7] evaluated logistic regression, decision trees, random forests, and XGBoost to predict diabetes progression in older patients with prediabetes. XGBoost was the most accurate model (60.66%), but its generalizability was limited by the dataset's narrow demographic scope. In spite of a minor decline in predictive performance over time, XGBoost showed promise as a model for identifying prediabetes risk factors among older adults. Similarly, Abbas et al [8] developed a model of prediabetes risk score for a Middle Eastern cohort based on random forest, gradient boosting, XGBoost, and deep learning. This model effectively screens risk across different groups of individuals by analyzing demographic and physiological factors, including age, blood pressure, BMI, waist size, and gender. Primary care settings could benefit from the study's focus on noninvasive, easily measurable variables.

Additionally, tree-based models, logistic regression, and LASSO have been commonly used to refine prediabetes risk prediction. Hu et al [9] developed a personalized nomogram that predicted 5-year prediabetes risk among Chinese adults. Using stepwise selection, LASSO, and ML models, Hu et al [9] found that the LASSO model provided the best performance with variables such as age, BMI, fasting blood glucose, and serum creatinine. As a result of this approach, LASSO can generate an accurate yet efficient model even with a limited number of predictive features. In another logistic regression–based study, Yu et al

[10] validated a prediabetes assessment model on a large Chinese dataset. Based on C statistics and calibration plots, the model demonstrated good discrimination, but a cohort study might improve its performance.

Efforts have also been made to incorporate nonlaboratory risk factors into predictive models. In a study by Dong et al [11], lifestyle factors such as sleep duration and recreational activity were incorporated into a model using logistic regression and interpretable ML techniques, especially XGBoost. SHAP (Shapley Additive Explanations) was used to determine variable significance, revealing that lifestyle variables are crucial to the model's detection efficiency. By incorporating clinical and lifestyle predictors, XGBoost can identify undiagnosed prediabetes and diabetes, offering a more comprehensive risk assessment.

As a result of these studies, we can observe that ensemble methods (random forest and XGBoost), regression-based approaches (logistic regression and LASSO), and interpretable ML models (eg, SHAP-enhanced XGBoost) all offer unique strengths in predicting prediabetes risk. According to the results, while tree-based models and ensemble models tend to be more accurate, regression techniques such as LASSO help create interpretable, efficient models, especially when resources are limited.

Methods

Dataset

This study used a dataset that is publicly accessible, which includes health records from 4743 individuals who were examined at the Health Management Center of Peking University Shenzhen Hospital from January 2020 to March 2023. The World Health Organization standards were followed when assessing fasting blood glucose levels, random blood glucose levels, oral glucose tolerance tests, and glycated hemoglobins of participants. Prediabetes was diagnosed if fasting blood glucose was between 6.1 and 6.9 mmol/L or if the blood glucose level was between 7.8 and 11.0 mmol/L after oral glucose tolerance test. Based on glucose metabolism status, participants were classified into 2 groups: normal (1593/4743, 33.6%) and prediabetes (3150/4743, 66.4%). The dataset included 22 features, comprising demographic, clinical, and laboratory variables such as age, BMI, HDL-C, and fasting blood glucose levels. The target variable for the study was binary, with participants categorized as either normal or prediabetic. Since this dataset is open to the public and anonymized, numeric values for individual IDs were preserved for traceability in the preprocessing phase, but they do not contain any personally identifiable information.

Variable Assignment and Data Categorization

In this study, the dataset includes both categorical and numerical variables. The categorical variables, such as status, gender, urine glucose, and urine protein, were assigned specific values to facilitate analysis. These values allow for easy differentiation between groups or conditions. On the other hand, continuous or numerical variables, such as age, BMI, and various blood and urine biomarkers, were used as-is without specific value

assignments since they naturally provide a range of measurements. Table 1 shows the assigned values for each of the categorical variables.

Table . Dataset variables and descriptions for prediabetes risk assessment.

Variable name	Meaning of variable	Type of variable	Assignment description
Status	Glucose metabolic status	Categorical variable	1=normal, 2=prediabetes
Age	Age	Numerical variable	Is unassigned
Gender	Gender	Categorical variable	0=female, 1=male
BMI	Body mass index	Numerical variable	Is unassigned
SBP	Systolic blood pressure	Numerical variable	Is unassigned
U-GLU	Urine glucose	Categorical variable	0=negative, 1=positive
PRO	Urine protein	Categorical variable	0=negative, 1=positive
TP	Total protein	Numerical variable	Is unassigned
ALB	Albumin	Numerical variable	Is unassigned
GLB	Globulin	Numerical variable	Is unassigned
T-BIL	Total bilirubin	Numerical variable	Is unassigned
DB	Direct bilirubin	Numerical variable	Is unassigned
IB	Indirect bilirubin	Numerical variable	Is unassigned
ALT	Alanine aminotransferase	Numerical variable	Is unassigned
AST	Aspartate transaminase	Numerical variable	Is unassigned
BUN	Blood urea nitrogen	Numerical variable	Is unassigned
SCr	Serum creatinine	Numerical variable	Is unassigned
UA	Uric acid	Numerical variable	Is unassigned
TC	Total cholesterol	Numerical variable	Is unassigned
TG	Triglycerides	Numerical variable	Is unassigned
HDL-C	High-density lipoprotein cholesterol	Numerical variable	Is unassigned
LDL-C	Low-density lipoprotein cholesterol	Numerical variable	Is unassigned

Data Preprocessing

Overview

For improved model performance, data preprocessing involved handling missing values through mean imputation, balancing the dataset using SMOTE (Synthetic Minority Oversampling Technique), and scaling features with StandardScaler() and MinMaxScaler(). Through these steps, the dataset was optimized for building reliable ML models for prediabetes risk prediction.

Handling Missing Data

Missing values were imputed using the mean of the corresponding feature, guaranteeing consistency and completeness in the dataset.

Balancing the Dataset

The dataset has an imbalanced class distribution, with 33.6% (1593/4743) representing the normal group (status=1) and 66.4% (3150/4743) representing the prediabetes group (status=2). This type of imbalance can influence the performance of classification models, specifically incorrectly predicting the minority class (normal group in this case), so SMOTE was used to oversample the minority class (normal group). This step

ensured that the ML models were not biased toward the larger class, improving predictive performance [12], particularly for prediabetes detection.

Scaling and Normalization

Scaling and normalization are pivotal steps when preparing continuous variables for models such as KNN, SVM, and LASSO, which are sensitive to feature scaling. To address this, the features are standardized using the "StandardScaler()," which tunes them to have a mean of 0 and an SD of 1. This standardization guarantees that all features are on a similar scale and refines model performance. In addition, normalization can be applied using the "MinMaxScaler()," which transforms the data into a range between 0 and 1 [13].

Exploratory Data Analysis

To obtain an understanding of the relationship across several features and to pick out any patterns, trends, or correlations that may guide next steps, the dataset was completely explored before applying predictive models. Heatmaps were used to visualize the relationship between numerical variables as shown in Figure 1. The main goal of this step is to gain a fruitful understanding of the raw data and arrange it for additional analysis [14]. Among the assessed models, SHAP analysis was

XSL•FO

RenderX

performed solely on the XGBoost classifier due to its alignment with the TreeExplainer framework. Models based on trees benefit from SHAP's precise additive feature attributions, which

are computationally efficient and theoretically robust. XGBoost's built-in support for SHAP made it more interpretable than other models (eg, SVM, KNN, and random forest).

Figure 1. Heatmap distribution of the dataset features. ALB: albumin; ALT: alanine aminotransferase; AST: aspartate transaminase; BUN: blood urea nitrogen; DB: direct bilirubin; GLB: globulin; HDL-C: high-density lipoprotein cholesterol; IB: indirect bilirubin; LDL-C: low-density lipoprotein cholesterol; SBP: systolic blood pressure; SCr: serum creatinine; T-BIL: total bilirubin; TC: total cholesterol; TG: triglyceride; TP: total protein; UA: uric acid.
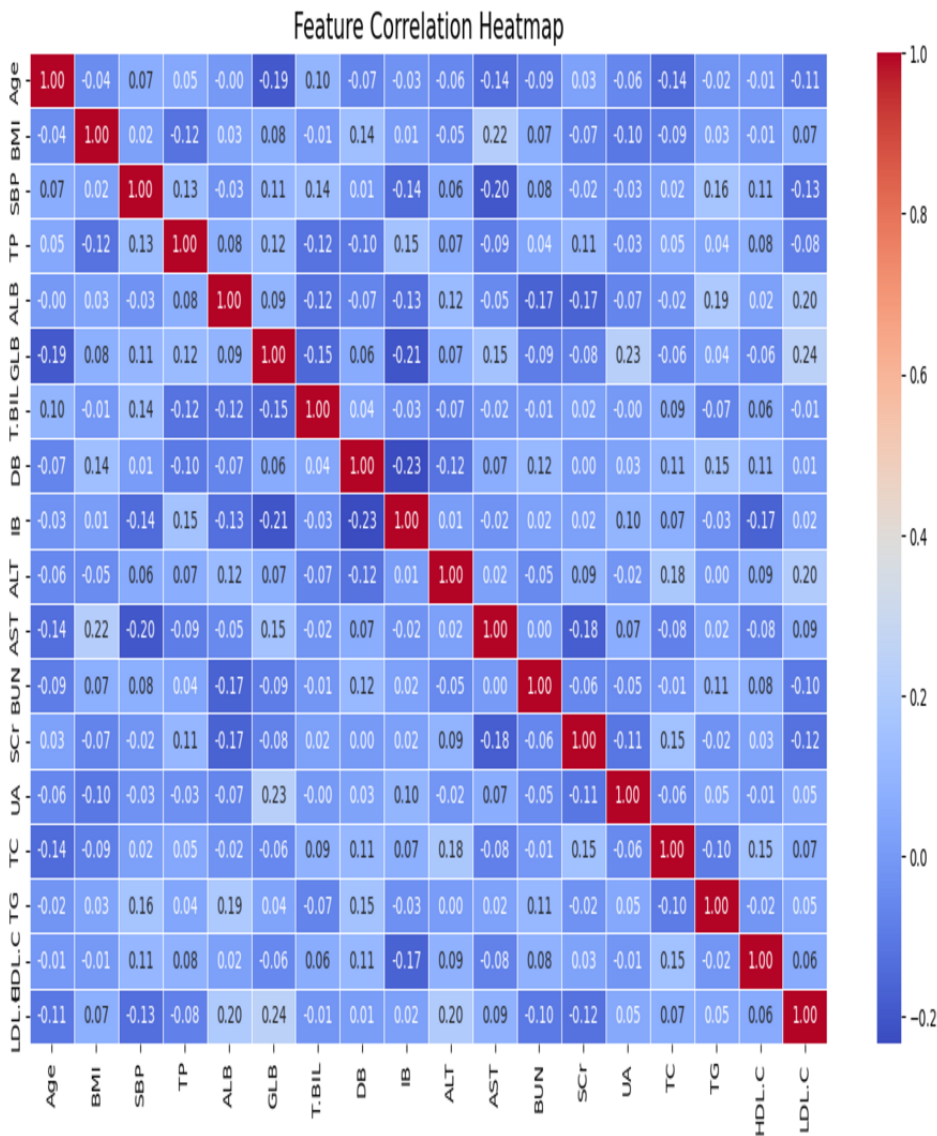
Features Selection

Overview

Two principal features selection techniques were applied after the data exploration phase to choose the most relevant and informative variables. A suitable feature selection not only enhances the performance and interpretability of a model but also reduces computational complexity and the risk of overfitting [15].

LASSO Regression

LASSO regression was used as the first method for feature selection. The LASSO method reduces the number of variables by shrinking the coefficients of less important features to zero, which effectively eliminates them from the model [16]. It is mostly useful for handling multicollinearity as it automatically picks one feature from a set of highly correlated features such as LDL-C and total cholesterol based on the correlation heatmap.

About PCA

The main aim of this technique is to reduce dimensionality in the dataset by transforming the base features into a smaller set of uncorrelated components while keeping most of the variance in the data [17]. In models facing overfitting, such as SVM and XGBoost, PCA reduced multicollinearity and compressed the retaining 95% of the variance in the data. Additionally, PCA reduced the number of variables, simplifying the model and making it more computationally efficient [18].

Before training the predictive models, these features selection techniques were applied. Using only relevant predictors improved model performance and generalizability. By using a structured approach to data exploration and features selection, we lay a strong foundation for building and evaluating ML models for prediabetes risk prediction in the next phase.

Model Development

Overview

In this study, different ML models were used to predict the onset of prediabetes. These algorithms were selected due to their ability to handle high-dimensional data, interpretability, and performance in classification tasks. As well as each model was tuned and evaluated to optimize performance for prediabetes detection.

XGBoost

XGBoost is a powerful gradient-boosting algorithm that constructs an ensemble of decision trees to improve classification accuracy. Each one tree is sequentially trained to emend the errors of the previous trees, which makes it more powerful for tasks with complex relationships between features [19]. XGBoost is known for its performance and speed in handling big datasets, which makes it appropriate for medical prediction tasks like prediabetes diagnosis. In addition, XGBoost applies regularized boosting techniques to overcome the difficulty of the model and correct overfitting; as a result, increasing model accuracy [20].

Random Forest

Random forest is an ensemble learning approach that constructs numerous decision trees during training. Every tree is built using a random subset of features and data samples, and the last prediction is made by averaging the predictions from all trees [21]. Random forest minimizes the risk of overfitting by using a bagging approach and tends to accomplish well on classification issues such as prediabetes detection.

About SVM

SVM is a supervised learning model that separates data points into distinct classes by finding an optimal hyperplane. For the complex relationships between predictors, such as BMI and age, a nonlinear kernel was applied. This method is suited for medical diagnosis since the decision boundary is not linearly separable in high-dimensional spaces [22].

About KNNs

KNN is an uncomplicated, nonparametric classifier that specifies the class label based on the most votes of the KNNs in the feature space [23]. In this study, KNN was used after scaling the features, and the optimal number of neighbors was set through hyperparameter tuning. Despite KNN being computationally intensive for big datasets, its clarity and interpretability make it a beneficial model for prediabetes classification.

Hyperparameter Tuning and Cross-Validation

Overview

Hyperparameter tuning was used for all models to recognize the optimal settings for each algorithm. To achieve that, we used GridSearchCV and RandomizedSearchCV, which systematically explore a range of hyperparameters and choose the set that maximizes model performance.

GridSearchCV

All combinations of hyperparameters are assessed exhaustively through a particular parameter grid. It is a systematic approach to identifying the effective parameter set [17]. With large datasets and complex models, it can be computationally expensive, so this study used GridSearchCV for models with a relatively small hyperparameter search space, which made it feasible to explore all combinations. The KNN algorithm was tuned by tuning the number of neighbors (k) and the distance metric.

RandomizedSearchCV

A randomized search of the hyperparameter space selects hyperparameter settings from the specified ranges [24]. It is more efficient than GridSearchCV when the search space is large because it explores a representative sample of possible combinations instead of testing them all. We used this technique for more complicated models such as random forest and XGBoost when the number of hyperparameters and possible values was too large for a wide search. RandomizedSearchCV assists with identifying optimal hyperparameters by setting a limit on the number of iterations (eg, 40).

Tuning Process for Each Model

XGBoost

The hyperparameters, such as the maximum tree depth, the learning rate, and the subsample ratio, were tuned using RandomizedSearchCV. This approach allowed for a more efficient search through a vast range of parameter values, making it fit for models with big parameter spaces. Random sampling allowed the tuning process to explore a diversity of hyperparameter combinations while preventing overfitting and maximizing classification accuracy.

Random Forest

To optimize hyperparameters such as the number of trees, maximum tree depth, and minimum samples required for a part, RandomizedSearchCV was used. This approach is selected for random forest because of the large search space, as it can easily sample a subset of hyperparameters to explore near-optimal settings.

XSL•FO

RenderX

About SVM

To fine-tune hyperparameters such as the kernel type and penalty parameter C, GridSearchCV was used. Due to the smaller search space for SVM, GridSearchCV is considered the best choice because this approach performs a wide search over the specified parameter values, so it guarantees to find the best possible combinations for the model.

About KNNs

To tune the distance metrics (eg, Euclidean or Manhattan distance) and number of neighbors (*k*), the GridSearchCV method was applied. This approach is useful to pick out the most effective neighborhood size and similarity measures for predicting prediabetes.

This tuning strategy guaranteed that every model was fine-tuned to work optimally for prediabetes prediction.

Cross-Validation Approach

The tuning process for each model included *k*-fold cross-validation to ensure reliable performance estimation and reduce the risk of overfitting. In *k*-fold cross-validation:

- The dataset is divided into *k* equal-sized subsets (folds).
- The model is trained on *k* – 1 folds and tested on the remaining fold. This process is repeated *k* times, with each fold serving as the test set once. The results are averaged to get a final evaluation metric.
- 5-fold cross-validation was used in this study, which balances computational cost and model evaluation reliability.

Through cross-validation, a robust estimate of model performance across various subsets of data is obtained by evaluating how well the model generalizes to unseen data [25]. To choose the best-performing parameter set, this method was used during hyperparameter tuning.

Model Evaluation Metrics

Overview

To evaluate the performance of ML models, various metrics were applied.

Accuracy

This is the measure of the percentage of true predictions made by the model out of all predictions. Nevertheless, accuracy alone can be misleading, particularly when the classes are imbalanced, as in the case of prediabetes diagnosis.

Precision

The proportion of true positive predictions to the total number of positive predictions. High precision indicates that the model produces few false positive errors, which is important in minimizing irrelevant treatments.

Recall (Sensitivity)

The ratio of correct positive predictions to the total actual positives. A higher recall means fewer cases of prediabetes were missed, making it necessary for early prediabetes diagnosis.

F_1-Score

The harmonic means of precision and recall contribute a balance between both metrics. It is mainly valuable when false positives and false negatives have serious consequences.

ROC-AUC Score

The ROC-AUC (receiver operating characteristic area under the curve) assesses the capability of the model to distinguish between both classes (normal and prediabetes). The ROC-AUC score provides an aggregate measure of performance throughout all classification thresholds, where a higher value refers to superior model performance.

Cross-Validated ROC-AUC

In addition to evaluating ROC-AUC on the test set, cross-validated ROC-AUC provides a more reliable estimate of the model's ability to generalize. This metric was calculated using *k*-fold cross-validation, giving a better indication of how the model will perform on unseen data.

By using these evaluation metrics, the comparative performance of the ML models was assessed, with a particular focus on balancing accuracy, precision, recall, and F_1-score to ensure reliable predictions for prediabetes risk assessment.

Results

XGBoost, Random Forest, SVM, and KNN

This section provides a comparative evaluation of the ML models applied in this study—XGBoost, random forest, SVM, and KNN—along with the results of feature selection techniques, such as LASSO regression and PCA. The performance of each model is assessed using multiple evaluation metrics, including accuracy, precision, recall, F_1-score, and ROC-AUC scores, on both the test set and cross-validation. Table 2 shows the performance metrics comparison of the ML models.

Table . Performance metrics comparison of machine learning models.

Model	Accuracy (%)	Precision	Recall	F_1-score	ROC-AUC[a] (test set)	Cross-validated ROC-AUC
XGBoost[b]	74.7	0.8128	0.7889	0.8007	0.7930	0.8600
Random forest	75.9	0.8391	0.7169	0.7732	0.8030	0.9117
SVM[c]	73.9	0.6260	0.6686	0.6466	0.7791	0.8630
KNN[d]	70.8	0.6901	0.6881	0.6890	0.7845	0.8397

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

[b]XGBoost: extreme gradient boosting.

[c]SVM: support vector machine.

[d]KNN: *k*-nearest neighbor.

Model Performance Comparison

Overview

The following subsections present the comparative results of XGBoost, random forest, SVM, and KNN models, each fine-tuned using hyperparameter optimization and evaluated using key performance metrics.

XGBoost

Based on 5-fold cross-validation, the XGBoost model showed a cross-validated ROC-AUC score of 0.86, indicating powerful discrimination between normal and prediabetic cases. In addition, the model achieved a precision of 0.8128, a recall of 0.7889, and an F_1-score of 0.8007 for the prediabetes class. This balanced performance emphasizes the model's strength to effectively minimize both false positives and false negatives, making it an effective method of prediabetes detection.

Random Forest

The random forest model achieved an excellent performance with a cross-validated ROC-AUC score of 0.9117, demonstrating its capability to generalize well across various subsets of the data. The model demonstrated a precision of 0.8391, a recall of 0.7169, and an F_1-score of 0.7732 for the prediabetes class. This indicates that the random forest model not only lowers the likelihood of false positives but also keeps a powerful recall rate, guaranteeing that fewer cases of prediabetes are missed.

About SVM

An SVM model, evaluated through 5-fold cross-validation, achieved a cross-validated ROC-AUC score of 0.8630, indicating its ability to distinguish between normal and prediabetic cases with high accuracy. For the prediabetes class, the model achieved a precision of 0.6260, a recall of 0.6686, and an F_1-score of 0.6466. Despite the SVM model providing a moderate balance between precision and recall, its recall score indicates potential for missing fewer prediabetic cases, making it a feasible choice for early-stage diagnosis.

About KNNs

The KNN model, evaluated using 5-fold cross-validation, demonstrated a cross-validated ROC-AUC score of 0.8397, reflecting its ability to differentiate between normal and prediabetic cases with moderate effectiveness. The model recorded a precision of 0.6901, a recall of 0.6881, and an F_1-score of 0.6890 for the prediabetes class. Although KNN performed slightly lower in terms of accuracy and precision compared to other models, it still provides an interpretable solution for prediabetes.

Performance Enhancement Through Hyperparameter Tuning

To optimize the performance of SVM and KNN, we used GridSearchCV for hyperparameter tuning. For more complex models such as XGBoost and random forest, RandomizedSearchCV was used to efficiently explore broader hyperparameter spaces.

Tables 3 and 4 highlight the improvement in model performance after hyperparameter optimization. All 4 models—XGBoost, random forest, SVM, and KNN—showed notable gains in both ROC-AUC and F_1-score metrics. For instance, XGBoost's ROC-AUC improved from 0.782 to 0.860, and random forest's from 0.807 to 0.9117. These results confirm the effectiveness of using GridSearchCV and RandomizedSearchCV in tailoring model parameters to the dataset, ultimately boosting classification accuracy and robustness. This step is particularly critical for clinical applications, where small improvements in sensitivity or specificity can have substantial impacts on patient outcomes.

The parallel processing option n_jobs = –1 was used to enable parallel processing. Each model required 3-8 minutes to be tuned on a standard multicore computer.

XSL•FO

RenderX

Table . Hyperparameter tuning summary for all models.

Model and hyperparameter		Range or values tested
SVM[a]		
	C	[0.1, 1, 10]
	Kernel	['linear', 'rbf']
	Gamma (rbf)	['scale', 'auto']
KNN[b]		
	n_neighbors	[3, 5, 7, 9, 11]
	Metric	['euclidean', 'manhattan']
XGBoost[c]		
	n_estimators	[50, 100, 200, 300]
	learning_rate	[0.01, 0.05, 0.1, 0.2]
	max_depth	[3, 5, 7, 9]
	Gamma	[0, 0.1, 0.3, 0.5]
	Subsample	[0.6, 0.8, 1.0]
	colsample_bytree	[0.6, 0.8, 1.0]
Random forest		
	n_estimators	[50, 100, 200]
	max_depth	[None, 3, 5]
	min_samples_split	[2, 5]
	min_samples_leaf	[1, 2]
	max_features	['sqrt', 'log2']
	Bootstrap	[True]

[a]SVM: support vector machine.

[b]KNN: *k*-nearest neighbor.

[c]XGBoost: extreme gradient boosting.

Table . Effect of hyperparameter tuning on model performance.

Model and metric		Default	Tuned (GridSearchCV/Randomized-SearchCV)
XGBoost[a]			
	ROC-AUC[b]	0.782	0.860
	F_1-score	0.731	0.801
Random forest			
	ROC-AUC	0.807	0.9117
	F_1-score	0.742	0.773
SVM[c]			
	ROC-AUC	0.813	0.863
	F_1-score	0.591	0.646
KNN[d]			
	ROC-AUC	0.805	0.839
	F_1-score	0.652	0.689

[a]XGBoost: extreme gradient boosting.

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

[c]SVM: support vector machine.

[d]KNN: k-nearest neighbor.

Descriptive Patterns From Exploratory Data Analysis Findings

Overview

Figure 1 shows several important patterns that emerged. The following features are highly correlated.

Strong Positive Correlation

Total cholesterol and LDL-C exhibited a strong positive correlation. As a result, the model may be redundant due to those variables sharing similar information. One of these features could potentially be excluded in the feature selection phase if it has a high correlation. It was found that total protein and albumin exhibit a high correlation, suggesting that combining them may not provide more insight than using either separately.

Weak or No Correlations

Correlations between variables such as age, BMI, and uric acid were weak or negligible. This is a significant finding because these variables may provide unique independent information that makes model-building more effective.

Negative Correlation

A mild negative correlation was found between LDL-C and HDL-C, which is consistent with their known inverse roles in cardiovascular health. Age and HDL-C also exhibited a slight negative correlation, suggesting that lipid profiles might change with aging. Multicollinearity issues happen when highly correlated variables distort the model's ability to differentiate between them due to this exploration in sights. It is crucial to recognize such relationships early in the process so that multicollinearity can be handled, and redundant features can be dropped in the next step, features selection.

A summary plot of SHAP data derived from the XGBoost model is shown in Figure 2. The most significant predictors are age, BMI, HDL-C, and LDL-C. As these variables are well-established risk factors for prediabetes, these findings support clinical intuition. Additionally, SHAP provided valuable visual confirmation that agreed with both the correlation analysis and the LASSO feature selection. Using these exploratory data analysis findings, LASSO regression and PCA were applied for feature selection, ensuring that informative predictors were retained while reducing redundancy and improving interpretability.

Figure 2. SHAP summary plot of XGBoost model. ALB: albumin; ALT: alanine aminotransferase; DB: direct bilirubin; HDL-C: high-density lipoprotein cholesterol; LDL-C: low-density lipoprotein cholesterol; PRO: urine protein; SBP: systolic blood pressure; SCr: serum creatinine; SHAP: Shapley Additive Explanations; TG: triglyceride; U-GLU: urine glucose; UA: uric acid; XGBoost: extreme gradient boosting.
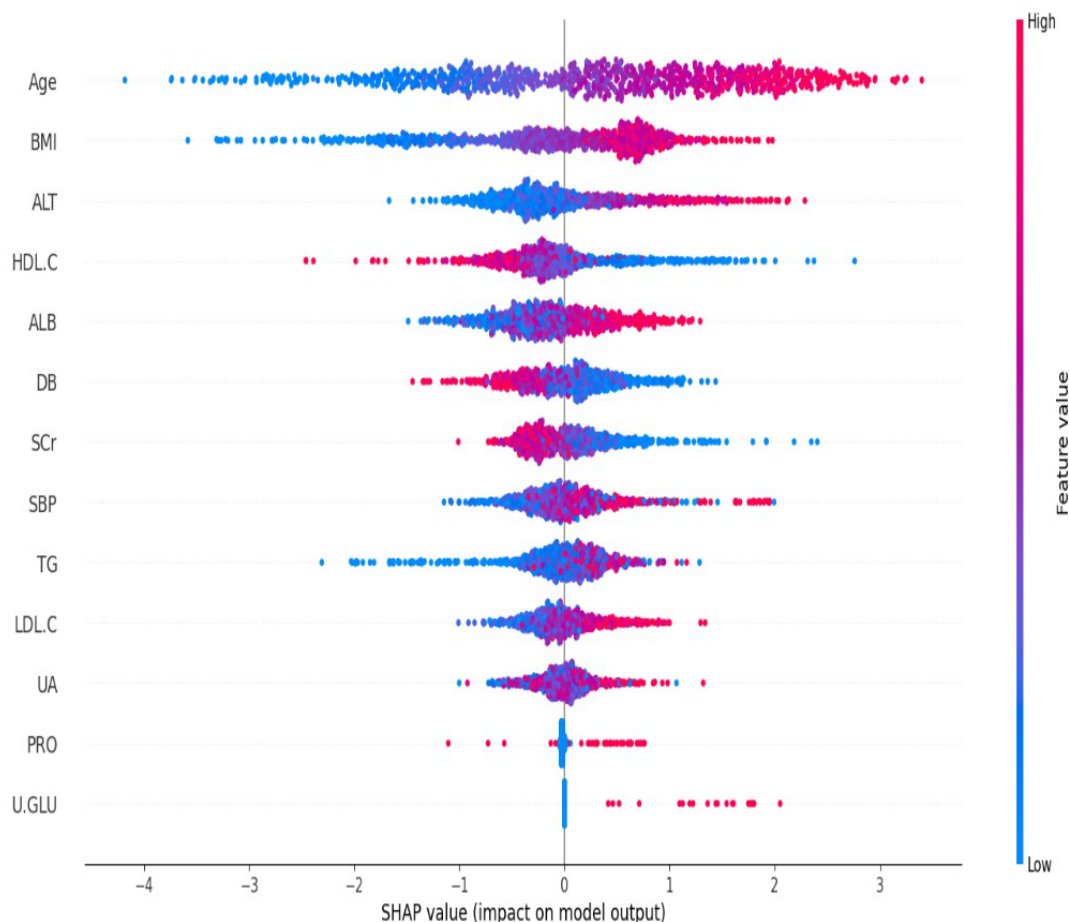
Feature Importance and Selection

Feature selection over LASSO regression guaranteed that every model was trained on the most relevant predictors. During LASSO, features like BMI, age, and HDL-C were consistently identified as significant predictors of prediabetes as shown in Figure 3. These features were retrained in the final model because of their significant predictive power across different iterations. The models differed in which features they emphasized:

- XGBoost identified BMI as the most significant predictor, aligning with established research that links higher BMI with increased prediabetes risk.
- SVM prioritized age as the first predictor, indicating that age may play an additional critical role when nonlinear relationships between variables are considered.
- Random forest and KNN provide insights into other key features such as LDL-C and HDL-C, demonstrating the various aspects of the data that every algorithm emphasizes.

This variance in feature significance underscores the utility of designing diverse models and selection techniques to better understand the predictors of prediabetes risk.

Figure 3. Features importance plots for XGBoost and SVM. ALB: albumin; ALT: alanine aminotransferase; DB: direct bilirubin; HDL-C: high-density lipoprotein cholesterol; LDL-C: low-density lipoprotein cholesterol; PRO: urine protein; SBP: systolic blood pressure; SCr: serum creatinine; SR: ; SVM: support vector machine; TG: triglyceride; UA: uric acid; XGBoost: extreme gradient boosting.
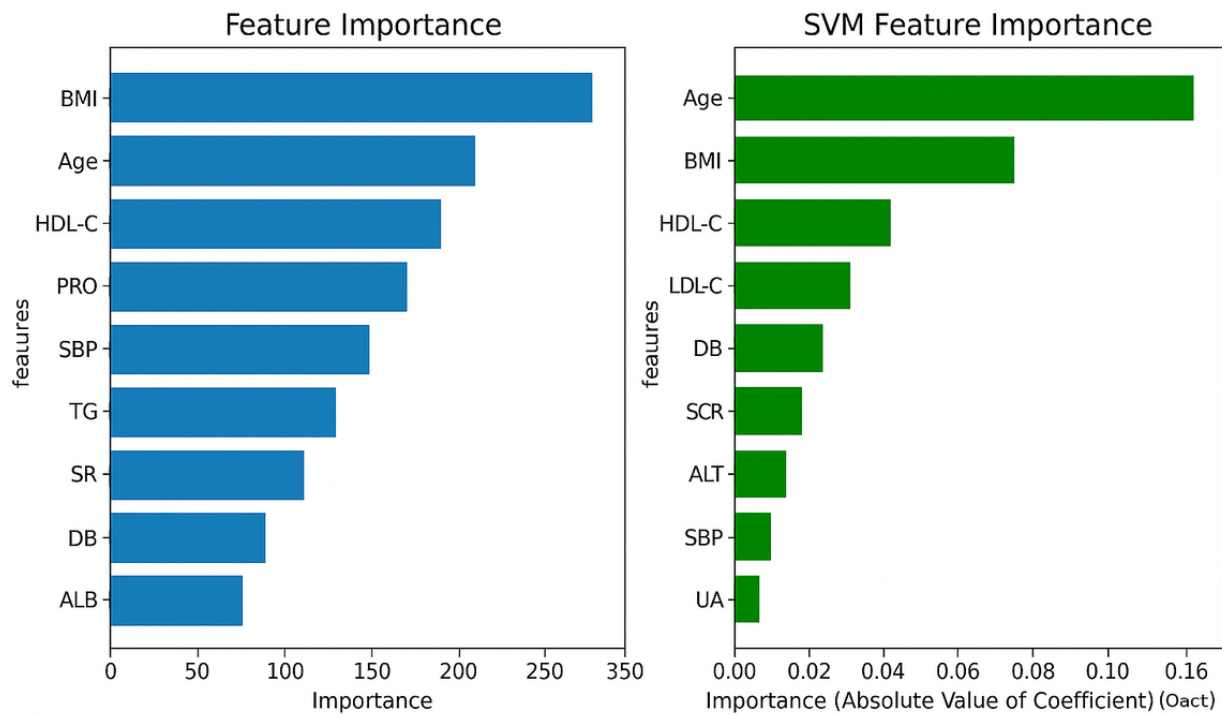
PCA Component Retention

PCA retained 12 principal components, accounting for 95% of the variance in the dataset.

Confusion Matrices

Overview

As shown in Figure 4, the confusion matrix demonstrates that every model's classification performance is detailed in terms of distinguishing normal cases from prediabetic cases. These results reflect the trade-offs each model faces in terms of true positives, false positives, true negatives, and false negatives.

Figure 4. Confusion matrix for XGBoost, SVM, random forest, and KNN models. KNN: *k*-nearest neighbor; SVM: support vector machine; XGBoost: extreme gradient boosting.
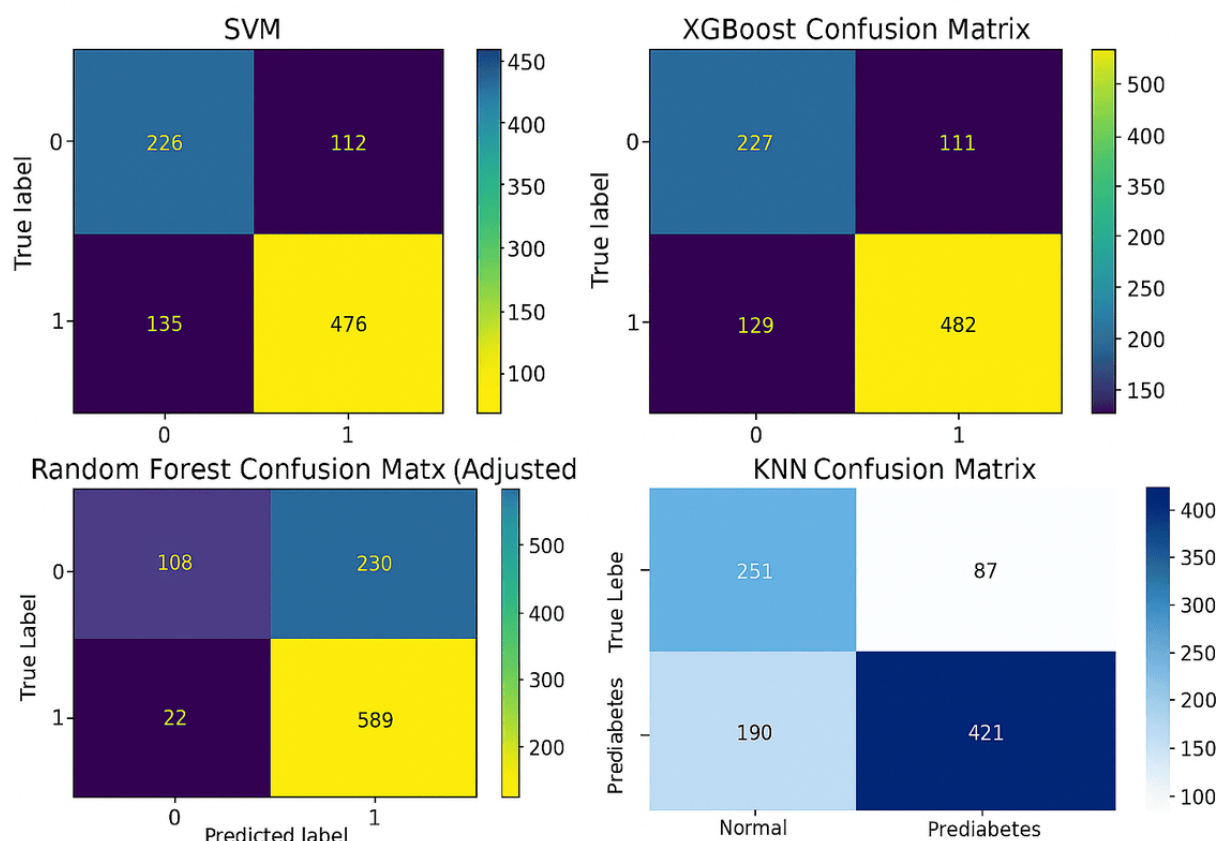
XGBoost

A comparatively balanced classification was accomplished with the XGBoost model, with 482 true positives and 227 true negatives, referring to good sensitivity. However, it recorded 129 false negatives and 111 false positives, proposing some limitations in minimizing misclassification errors, especially false negatives, which are pivotal in clinical settings.

Random Forest

The random forest model (default threshold of 0.5) correctly identified 513 true positives and 208 true negatives, which are better results compared to XGBoost. The model demonstrated a higher sensitivity than other models, as it reduced the number of false negatives to 98. Despite this, 130 false positives were observed, which indicates a slightly higher trade-off in specificity.

A threshold adjustment of 0.2627 substantially improved the random forest's ability to detect prediabetic cases, resulting in 589 true positives and 22 false negatives. A notable rise in false positives (230) and a reduction in true negatives (108) resulted from this adjustment, indicating a move toward maximizing sensitivity over specificity. There may be some advantages to this configuration in scenarios where minimizing missed prediabetic cases is prioritized over averting false positives.

About SVM

For the overall distribution of true positives and true negatives, the SVM model obtained 476 true positives and 226 true negatives, which is like XGBoost's. A total of 135 false

negatives and 112 false positives have been recorded, indicating that while SVM has a strong classification capability, it is more susceptible to false negatives, which limits its effectiveness for early detection cases.

About KNNs

This model performed moderately, generating 421 true positives and 251 true negatives. Even though KNN can effectively detect normal cases, it is less reliable when it comes to identifying prediabetic cases. It showed 190 false negatives and 87 false positives, indicating a higher rate of misclassification.

To summarize, the confusion matrices demonstrate that the random forest model minimizes false negatives better than other models, especially when thresholds are adjusted. Random forest has a significant advantage over XGBoost and SVM when it comes to sensitivity, which makes it particularly suitable for prediabetes detection, where minimizing missed cases is crucial. While KNN is the most effective at identifying normal cases, it lacks the discriminative power necessary to accurately classify prediabetes, illustrating that it may be more fit as a baseline or for smaller datasets.
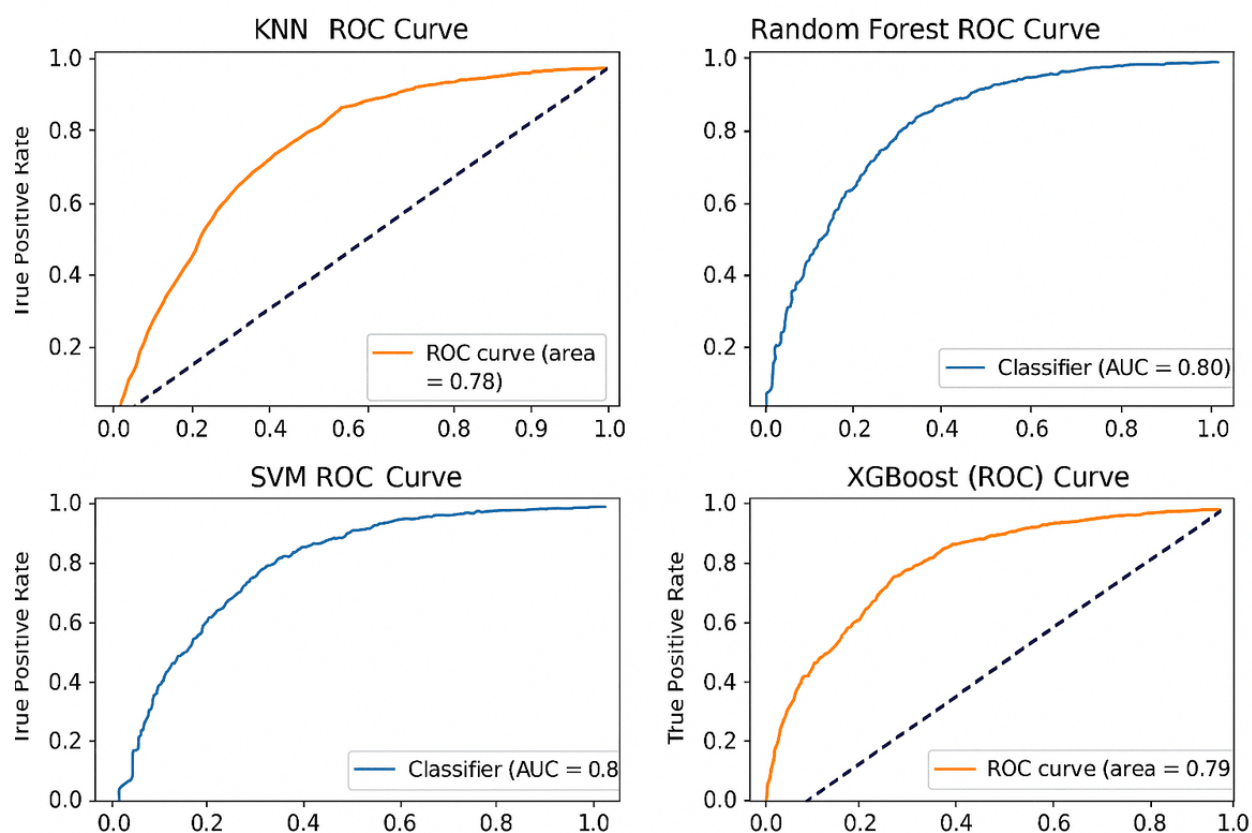
ROC Curves

Overview

Figure 5 shows the ROC (receiver operating characteristic) curves for every model, further clarifies the trade-offs between sensitivity and specificity, and shows the performance of each model in terms of how well it separates between normal and prediabetic cases. The random forest model showed the most

convenient ROC curve, while XGBoost and SVM also displayed powerful curves, suggesting effective categorization performance.

Figure 5. ROC curve comparison across models. KNN: *k*-nearest neighbor; SVM: support vector machine; XGBoost: extreme gradient boosting.

XGBoost

This classifier showed an AUC (area under the curve) of 0.79. The XGBoost ROC curve reflects a relatively good trade-off between the true positive rate (sensitivity) and false positive rate (1 – specificity), indicating that it is an effective classification model, but has some room for improvement in distinguishing classes.

About SVM

The SVM classifier produced a slightly lower AUC of 0.78. However, the SVM struggles slightly more with false positives, as indicated by its ROC curve, which does not consistently approach the top-left corner. Despite this, it performs reasonably well when it comes to classification.

Random Forest

Across the 4 models tested, the random forest model achieved the elevated AUC at 0.80. With a more pronounced upward curve, its ROC curve reflects better differentiation between positive and negative classes, showcasing outstanding classification abilities.

About KNNs

The KNN classifier achieved a score of 0.78, suggesting a fair rank of accuracy in the diagnosis of positive and negative cases. According to the ROC curve for the KNN model, there is a moderate trade-off between the true positive rate (sensitivity) and the false positive rate (1 – specificity). As well, there is

some evidence to suggest that the KNN model has some ability to separate the 2 classes, but its shape suggests that it has room for improvement, as it does not consistently approach the top-left corner, which would indicate an ideal performance.

In a nutshell, all 4 models exhibit durable performance, with AUC values ranging from 0.78 to 0.80. The random forest model manifests as the best-performing classifier, followed closely by XGBoost, SVM, and KNN.

Discussion

Principal Findings

Through systematically integrating model comparison, advanced hyperparameter tuning, and interpretable feature selection techniques, we present a robust, interpretable framework for early prediabetes prediction. By combining SHAP analysis and LASSO regression, this research provides both high performance and transparency, compared to previous studies that focused solely on accuracy.

Comparative Strengths and Limitations of Each Model

Overview

For prediabetes prediction, XGBoost, random forest, SVM, and KNN each show distinct strengths and weaknesses.

Random Forest

In terms of overall discriminative ability, the random forest model accomplished a superior cross-validated ROC-AUC score (0.9117). According to this result, random forest is a robust choice for early detection scenarios as it can generalize to different datasets well. Due to its ability to prioritize recall through threshold adjustments, 22 false negatives were reduced, but false positives increased (230). In view of this trade-off, random forest may be highly powerful when the cost of missing a prediabetic case outweighs the risk of overdiagnosis.

XGBoost

In evaluation, the XGBoost classifier showcased robust performance, as it attained a high precision score of 0.8128 and a balanced recall score. According to these metrics, it seems that XGBoost is particularly adept at minimizing false positives and false negatives, which is highly critical in clinical settings where diagnostic accuracy directly influences patient outcomes. The ROC-AUC score of XGBoost did not surpass that of random forest, despite its ability to balance sensitivity and specificity, making it a viable choice for routine clinical applications.

About SVM

With an AUC of 0.78, the SVM model ranked behind both XGBoost and random forest. Despite their superior performance in high-dimensional spaces and in datasets with clear class separation, SVM models have limited linear separability in the prediabetes dataset, impacting their discriminative power. The model has a good ROC-AUC and F_1-score, with reasonable precision and recall, but when it comes to complex relationships, it lags behind the others. Optimizing feature engineering may upgrade its performance by searching alternative SVM kernels, combining nonlinear interactions, or incorporating alternative kernels.

About KNNs

It performed rationally well in terms of classification performance but ranked lowest in terms of accuracy among the evaluated models, with an accuracy of 70.8% and ROC-AUC of 0.78. Because of its simplicity and reliance on distance metrics, KNN is expected to have lower discriminative power than more complex models such as random forest and XGBoost. This model may be valuable as a baseline model or may be convenient for small datasets with a focus on computational efficiency. The reasonable performance of KNN is a result of its sensitivity to distance metrics and the number of neighbors (k), which may prevent it from catching subtle differences in detecting normal and prediabetic cases. Thus, while KNN may be beneficial in straightforward scenarios, it does not have the same level of precision and recall as more sophisticated models.

Impact of Feature Selection

Feature selection played a crucial role in optimizing the models' performance by focusing on the main relevant predictors. LASSO regression was used to characterize the prime features across models, with BMI, age, LDL-C, and HDL-C consistently emerging as important risk factors for prediabetes. In addition to improving the interpretability of the models, this approach also improved the predictive accuracy by reducing overfitting. The strict feature selection process warranted that the models stayed efficient while maintaining high classification power.

Confusion Matrix and Threshold Analysis

The performance metrics were significantly influenced by adjusting decision thresholds, especially for random forest and XGBoost. A threshold adjustment in random forest minimized the risk of missed diagnoses by reducing false negatives (22 cases). Even so, this came at the expense of a boosted number of false positives (230 cases), suggesting a trade-off between recall and precision. XGBoost, while less sensitive to threshold changes, maintained a balanced approach, limiting both false positives and false negatives effectively. As a result of these outcomes, threshold tuning plays an important role in optimizing model performance for specific clinical applications, such as prioritizing recall in high-risk populations to avoid disease progression.

Clinical Implications

The results suggest that XGBoost and random forest are the most promising models for enhancing prediabetes diagnosis, given their ability to generalize across different datasets and include reliable classification performance. The higher ROC-AUC score achieved over random forest (91.17%) reflects its potential for widespread use in clinical settings, especially where minimizing the risk of missed cases is crucial. The powerful performance of XGBoost among diverse metrics also highlights its practicality for routine screening, where both false positives and false negatives need to be minimized. By adjusting model thresholds, clinicians can customize diagnostic strategies to meet individual patient needs, such as increasing sensitivity for at-risk patients. Even though SVMs and KNNs do not outperform the best models, they still provide useful insights, especially when data dimensionality or simplicity are important factors.

Conclusions

ML models, specifically random forest and XGBoost, have been found to be most sensitive to prediabetes risk assessment, and their performance has powerful discriminative power and high ROC-AUC scores. Combined with feature selection techniques such as LASSO regression, these models offer worthy insights into essential prediabetes predictors, such as BMI, age, and HDL-C. Based on the ROC and AUC analyses, all models—XGBoost, SVM, random forest, and KNN—are viable options for predicting prediabetes. Random forests are robust classifiers because of their ensemble nature, which reduces overfitting and enhances generalizability. SVM and XGBoost also produce competitive results, suggesting their classification abilities can be improved with further parameter tuning. With systematic exploratory data analysis and feature selection, these models can become reliable tools for detecting early prediabetes and offering pathways for optimizing them.

To confirm the generalizability of these models, future research should include validating them in diverse populations, adding biomarkers and genetics to improve prediction accuracy, and integrating these models into clinical decision support systems to assess risk in real time. These models contribute to more

accurate and timely diagnosis of prediabetes, promoting timely intervention and ultimately improving health outcomes.

Conflicts of Interest

None declared.

References

1. Liu Y, Feng W, Lou J, et al. Performance of a prediabetes risk prediction model: a systematic review. Heliyon 2023 May;9(5):e15529. [doi: 10.1016/j.heliyon.2023.e15529]
2. Schwartz JL, Tseng E, Maruthur NM, Rouhizadeh M. Identification of prediabetes discussions in unstructured clinical documentation: validation of a natural language processing algorithm. JMIR Med Inform 2022 Feb 24;10(2):e29803. [doi: 10.2196/29803] [Medline: 35200154]
3. Hathaway QA, Roth SM, Pinti MV, et al. Machine-learning to stratify diabetic patients using novel cardiac biomarkers and integrative genomics. Cardiovasc Diabetol 2019 Jun 11;18(1):78. [doi: 10.1186/s12933-019-0879-0] [Medline: 31185988]
4. De Silva K, Jönsson D, Demmer RT. A combined strategy of feature selection and machine learning to identify predictors of prediabetes. J Am Med Inform Assoc 2020 Mar 1;27(3):396-406. [doi: 10.1093/jamia/ocz204]
5. Dwyer DB, Falkai P, Koutsouleris N. Machine learning approaches for clinical psychology and psychiatry. Annu Rev Clin Psychol 2018 May 7;14:91-118. [doi: 10.1146/annurev-clinpsy-032816-045037] [Medline: 29401044]
6. Talari P, N B, Kaur G, et al. Hybrid feature selection and classification technique for early prediction and severity of diabetes type 2. PLoS ONE 2024;19(1):e0292100. [doi: 10.1371/journal.pone.0292100] [Medline: 38236900]
7. Liu Q, Zhou Q, He Y, Zou J, Guo Y, Yan Y. Predicting the 2-year risk of progression from prediabetes to diabetes using machine learning among Chinese elderly adults. J Pers Med 2022 Jun 27;12(7):7. [doi: 10.3390/jpm12071055] [Medline: 35887552]
8. Abbas M, Mall R, Errafii K, et al. Simple risk score to screen for prediabetes: a cross-sectional study from the Qatar Biobank cohort. J Diabetes Investig 2021 Jun;12(6):988-997. [doi: 10.1111/jdi.13445] [Medline: 33075216]
9. Hu Y, Han Y, Liu Y, et al. A nomogram model for predicting 5-year risk of prediabetes in Chinese adults. Sci Rep 2023;13(1):1-16. [doi: 10.1038/s41598-023-50122-3]
10. Yu LP, Dong F, Li YZ, et al. Development and validation of a risk assessment model for prediabetes in China national diabetes survey. World J Clin Cases 2022 Nov 16;10(32):11789-11803. [doi: 10.12998/wjcc.v10.i32.11789] [Medline: 36405266]
11. Dong W, Tse TYE, Mak LI, et al. Non-laboratory-based risk assessment model for case detection of diabetes mellitus and pre-diabetes in primary care. J Diabetes Investig 2022 Aug;13(8):1374-1386. [doi: 10.1111/jdi.13790] [Medline: 35293149]
12. Liaw LCM, Tan SC, Goh PY, Lim CP. A histogram SMOTE-based sampling algorithm with incremental learning for imbalanced data classification. Inf Sci 2025 Jan;686:121193. [doi: 10.1016/j.ins.2024.121193]
13. Raju VNG, Lakshmi KP, Jain VM, Kalidindi A, Padma V. Study the influence of normalization/transformation process on the accuracy of supervised classification. Presented at: 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT); Aug 20-22, 2020; Tirunelveli, India p. 729-735. [doi: 10.1109/ICSSIT48917.2020.9214160]
14. Da Poian V, Theiling B, Clough L, et al. Exploratory data analysis (EDA) machine learning approaches for ocean world analog mass spectrometry. Front Astron Space Sci 2023;10(May):1-17. [doi: 10.3389/fspas.2023.1134141]
15. Saxena R, Sharma SK, Gupta M, Sampada GC. A novel approach for feature selection and classification of diabetes mellitus: machine learning methods. Comput Intell Neurosci 2022;2022:3820360. [doi: 10.1155/2022/3820360] [Medline: 35463255]
16. Noaro G, Cappon G, Vettoretti M, Sparacino G, Favero SD, Facchinetti A. Machine-learning based model to improve insulin bolus calculation in type 1 diabetes therapy. IEEE Trans Biomed Eng 2021 Jan;68(1):247-255. [doi: 10.1109/TBME.2020.3004031] [Medline: 32746033]
17. Gollapalli M, Alansari A, Alkhorasani H, et al. A novel stacking ensemble for detecting three types of diabetes mellitus using a Saudi Arabian dataset: pre-diabetes, T1DM, and T2DM. Comput Biol Med 2022;147:105757. [doi: 10.1016/j.compbiomed.2022.105757] [Medline: 35777087]
18. Jia W, Sun M, Lian J, Hou S. Feature dimensionality reduction: a review. Complex Intell Syst 2022 Jun;8(3):2663-2693. [doi: 10.1007/s40747-021-00637-x]

19. Hasan MK, Alam MA, Das D, Hossain E, Hasan M. Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access 2020;8:76516-76531. [doi: 10.1109/ACCESS.2020.2989857] [Medline: 34812373]

20. Liu CH, Chang CF, Chen IC, et al. Machine learning prediction of prediabetes in a young male Chinese cohort with 5.8-year follow-up. Diagnostics (Basel) 2024 May 8;14:10. [doi: 10.3390/diagnostics14100979] [Medline: 38786280]

21. Alzyoud M, Alazaidah R, Aljaidi M, et al. Diagnosing diabetes mellitus using machine learning techniques. Int J Data Netw Sci 2024;8:179-188. [doi: 10.5267/j.ijdns.2023.10.006]

22. Olisah CC, Smith L, Smith M. Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. Comput Methods Programs Biomed 2022 Jun;220:106773. [doi: 10.1016/j.cmpb.2022.106773] [Medline: 35429810]

23. Diranisha V, Triayudi A, Komalasari RT. Implementation of k-nearest neighbour (KNN) algorithm and random forest algorithm in identifying diabetes. SAGA J Technol Inform Syst 2024;2(2):234-244. [doi: 10.58905/saga.v2i2.253]

24. Yennimar Y, Rasid A, Kenedy S. Implementation of support vector machine algorithm with hyper-tuning randomized search in stroke prediction. J Sist Inf Ilmu Komput Prima 2023;6(2):61-65. [doi: 10.34012/jurnalsisteminformasidanilmukomputer.v6i2.3479]

25. Yates LA, Aandahl Z, Richards SA, Brook BW. Cross validation for model selection: a review with examples from ecology. Ecol Monogr 2023 Feb;93(1):e1557. [doi: 10.1002/ecm.1557]

Abbreviations

AUC: area under the curve
HDL-C: high-density lipoprotein cholesterol
KNN: *k*-nearest neighbor
LASSO: Least Absolute Shrinkage and Selection Operator
LDL-C: low-density lipoprotein cholesterol
ML: machine learning
PCA: principal component analysis
ROC: receiver operating characteristic
ROC-AUC: receiver operating characteristic area under the curve
SHAP: Shapley Additive Explanations
SMOTE: Synthetic Minority Oversampling Technique
SVM: support vector machine
XGBoost: extreme gradient boosting

XSL•FO
RenderX

In Silico Analysis and Validation of A Disintegrin and Metalloprotease (ADAM) 17 Gene Missense Variants: Structural Bioinformatics Study

Abdelilah Mechnine[1], PhD; Asmae Saih[2], PhD; Lahcen Wakrim[3], PhD; Ahmed Aarab[1], PhD

[1]Biotechnology and Biomolecule Engineering Research Team, Faculty of Sciences and Techniques of Tangier, Abdelmalek Essaâdi University, Ancienne Route de l'Aéroport, Km 10, Ziaten. BP:416, Tangier, Morocco

[2]Laboratory of Biology and Health, URAC 34, Faculty of Sciences Ben M'Sik, University of Hassan II Casablanca, Casablanca, Morocco

[3]Virology Unit, Immunovirology Laboratory, Institut Pasteur du Maroc, Casablanca, Morocco

Corresponding Author:
Abdelilah Mechnine, PhD
Biotechnology and Biomolecule Engineering Research Team, Faculty of Sciences and Techniques of Tangier, Abdelmalek Essaâdi University, Ancienne Route de l'Aéroport, Km 10, Ziaten. BP:416, Tangier, Morocco

Abstract

Background: The protein A disintegrin and metalloprotease (ADAM) domain containing 17, also called tumor necrosis factor alpha–converting enzyme, is mainly responsible for cleaving a specific sequence Pro-Leu-Ala-Gln-Ala-/-Val-Arg-Ser-Ser-Ser in the membrane-bound precursor of tumor necrosis factor alpha. This cleavage process has significant implications for inflammatory and immune responses, and recent research indicates that genetic variants of ADAM17 may influence susceptibility to and severity of SARS-CoV-2 infection.

Objective: The aim of the study is to identify the most deleterious missense variants of ADAM17 that impact protein stability, structure, and function and to assess specific variants potentially involved in SARS-CoV-2 infection.

Methods: A bioinformatics approach was used on 12,042 single-nucleotide polymorphisms using tools including SIFT (Sorting Intolerant From Tolerant), PolyPhen2.0, PROVEAN (Protein Variation Effect Analyzer), PANTHER (Protein Analysis Through Evolutionary Relationships), SNP&GO (Single Nucleotide Polymorphisms and Gene Ontology), PhD-SNP (Predictor of Human Deleterious Single Nucleotide Polymorphisms), Mutation Assessor, SNAP2 (Screening for Non-Acceptable Polymorphisms 2), MUpro, I-Mutant, iStable, InterPro, Sparks-x, PROCHECK (Programs to Check the Stereochemical Quality of Protein Structures), PyMol, Project HOPE (Have (y)Our Protein Explained), ConSurf, and SWISS-MODEL. Missense variants of ADAM17 were collected from the Ensembl database for analysis.

Results: In total, 7 nonsynonymous single-nucleotide polymorphisms (P556L, G550D, V483A, G479E, G349E, T339P, and D232E) were identified as high-risk pathogenic by all prediction tools, and these variants were found to potentially have deleterious effects on the stability, structure, and function of the ADAM17 protein, potentially destroying the entire cleavage process. Additionally, 4 missense variants (Q658H, D657G, D654N, and F652L) in positions related to SARS-CoV-2 infection exhibited high conservation scores and were predicted to be deleterious, suggesting that they play an important role in SARS-CoV-2 infection.

Conclusions: Specific missense variants of ADAM17 are predicted to be highly pathogenic, potentially affecting protein stability and function and contributing to SARS-CoV-2 pathogenesis. These findings provide a basis for understanding their clinical relevance, aiding in early diagnosis, risk assessment, and therapeutic development.

Introduction

The ADAM family, which stands for A disintegrin and metalloprotease, is made up of both single-passage transmembrane proteins and secreted metalloendopeptidases. These enzymes share a distinct domain structure, which includes a prodomain, metalloprotease domain, disintegrin domain, cysteine-rich region, epidermal growth factor–like domain, a transmembrane segment, and a C-terminal cytoplasmic tail [1,2].

However, some human ADAM proteins lack a functional protease domain, meaning that many of ADAMs' roles are centered on protein-protein interactions rather than protease

activity. ADAM proteins belong to the EC 3.4.24.46 enzyme classification and are part of the MEROPS M12B peptidase family. For instance, active ADAM proteases are often referred to as sheddases because they cleave or remove extracellular parts of transmembrane proteins, such as ADAM10, and are able to cleave part of the human epidermal growth factor receptor 2, which then activates the receptor. ADAM genes are present in choanoflagellates, animals, fungi, and certain green algae, while these proteins are not present in most green algae and all land plants because they probably lost it. ADAM proteins have been historically referred to by names like adamalysin or MDC (metalloproteinase type, disintegrin type, cysteine-rich) family [3-6].

ADAM17 is a polypeptide of 824 amino acids, 93,021 Da, and it is located on chromosome 2p25. ADAM17 is hugely expressed in a lot of tissues, such as the brain, kidney, heart, and voluntary muscle, and its expression changes during embryonic development and adult life. ADAM17 is a multidomain protein composed of several conserved domains, starting with an N-terminal signal peptide spanning amino acids (aa 1 ‐ 17), followed by a prodomain (aa 18 ‐ 216), in which there is a cysteine switch-like region $PKVCGY^{186}$ (aa 181 ‐ 188), a metalloenzyme or catalytic domain (aa 217 ‐ 474) with a Zn-binding domain region (aa 405 ‐ 417), a disintegrin cysteine-rich domain (aa 480 ‐ 559), an epidermal growth factor–like region (aa 571 ‐ 602), followed by a cysteine-rich domain (aa 603 ‐ 671), and a transmembrane domain (aa 672 ‐ 694), end by a cytoplasmic tail (aa 695 ‐ 824). Tyr^{702}, Thr^{735}, and Ser^{819} have been shown as cytoplasmic phosphorylation sites, and Ser^{791} has been shown as a cytoplasmic dephosphorylation site. ADAM17 has little or no sequence similarities with other ADAMs, its closest relative is ADAM10; however, their protein sequence homology is a smaller amount than 30% consistent with the National Center for Biotechnology Information Basic Local Alignment Search Tool [7,8].

The purpose of ADAM17 is to treat tumor necrosis factor alpha (TNF-α) both inside the trans-Golgi network's internal membranes and on the cell's surface. The cleavage and release of a soluble ectodomain from membrane-bound proproteins (such as pro-TNF-α) involve this process, which is also known as "excretion" and is recognized to have physiological significance. The first "sheddase" to be discovered, ADAM17, is also thought to be involved in the release of a wide range of membrane-anchored cytokines, cell adhesion molecules, receptors, ligands, and enzymes [9,10].

The 26-kDa type II transmembrane propolypeptide that the TNF-α gene encodes inserts into the cell membrane during maturation, according to the gene's cloning. Pro-TNF-α is physiologically active on the cell surface and can trigger immunological responses by means of juxtacrine intercellular communication. The Ala76-Val77 amide bond of pro-TNF-α, however, is susceptible to proteolytic breakage, which liberates the molecule's soluble 17-kDa extracellular domain (ectodomain). The cytokine known as TNF-α, which is of vital importance in paracrine signaling, is the soluble ectodomain.

ADAM17 catalyzes the proteolytic release of soluble TNF-α [11].

ADAM17 has recently been identified as a key modulator of radiation therapy resistance. Radiation treatment may induce furin-mediated cleavage of the inactive form of ADAM17, converting it into its active form in a dose-dependent manner. This results in increased ADAM17 activity both in vitro and in vivo. In nonsmall cell lung cancer, radiation therapy has also been demonstrated to activate ADAM17, which leads to the excretion of several survival factors, the activation of the growth factor pathway, and the development of radiation resistance [12].

In addition, ADAM17 might be a key player in the Notch signaling pathway when the intracellular Notch domain (from the Notch1 receptor) is released proteolytically following ligand interaction. By controlling the mammary gland's excretion of the epidermal growth factor receptor, amphiregulin ligand, ADAM17 also controls the mitogen-activated protein kinase signaling pathway. Additionally, ADAM17 contributes to the excretion of the cell adhesion protein, L-selectin. To investigate the structural and functional effects of the chosen missense variations of the ADAM17 protein, we used a variety of bioinformatic techniques in the current methodology [13,14].

The primary cellular receptor used by SARS-CoV-2 to infect cells is the enzyme angiotensin-converting enzyme 2 (ACE2). This receptor is recognized by the S protein of SARS-CoV-2, which facilitates the key process of viral entry into a target cell. ADAM17 directly interacts with ACE2, leading to the shedding of ACE2 into the extracellular space, while transmembrane protease, serine 2 (TMPRSS2) not only cleaves ACE2 but also cleaves the SARS-CoV-2 S protein, facilitating membrane fusion and cellular uptake of the virus.

Both ADAM17 and TMPRSS2 act on ACE2, although these proteases can have opposite effects on the loss of ACE2. When the respective proteolytic activities of ADAM17 and TMPRSS2 result in increased shedding of ACE2, this situation may act as a natural barrier to infection. This could be due to the interaction of soluble ACE2 with the virus, preventing it from binding to susceptible tissues [15-21].

Alongside our work on ADAM17 variants and SARS-CoV-2 infection, Cho et al [22] explored in detail the immunogenicity of COVID-19 vaccines in different patients and highlighted immune response variation in terms of COVID-19 host factors. Additionally, Abbas et al [23] have used machine learning to profile RNA 5-methylcytosine modifications, a computational approach that is similarly conceptually related to how we have applied predictive methods in our analysis of ADAM17 variants [23].

This study highlights the potential of bioinformatics-driven variant analysis in exploring high-risk ADAM17 mutations, shedding light on their possible role in SARS-CoV-2 infection and advancing our understanding of ADAM17's impact on immune and inflammatory processes.

Methods

Overview

We collected single-nucleotide polymorphisms (SNPs) of the ADAM17 gene data from the Ensembl database [24]. Only missense variants were extracted from the total SNPs for the first study, and only 7 missense variants were selected and tested for further bioinformatic approaches. For the second study, only variants related to SARS-CoV-2, located between positions 652 and 658, were extracted. In total, 4 missense variants were selected and tested using bioinformatics approaches. The amino acid sequence in FASTA (FAST-All) format was retrieved from the UniProt database [25-27].

Ethical Considerations

This study involved only in silico analyses based on publicly available genomic data retrieved from the Ensembl genome database [24]. The data used are fully anonymized and do not contain any personally identifiable information or involve human or animal subjects. Thus, no ethics approval was required.

Prediction of Deleterious Nonsynonymous SNPs Using SIFT, PolyPhen, PROVEAN, SNAP2, Mutation Assessor, PANTHER, SNP&GO, and PhD-SNP

We used 5 bioinformatic servers for the first study, namely, SIFT (Sorting Intolerant From Tolerant), PolyPhen, PROVEAN (Protein Variation Effect Analyzer), SNAP2 (Screening for Non-Acceptable Polymorphisms 2), and Mutation Assessor. SIFT (version 6.0), a web-based server, was used to predict the impact of a substitution on protein function. A SIFT score >0.05 indicates a tolerated or neutral mutation, while a score <0.05 indicates a deleterious or damaging mutation. PolyPhen-2 (version 2.2) is a web server that predicts the impact of mutations on protein structure and function. It was used to classify mutations into probably damaging, possibly damaging, and neutral. PROVEAN (version 1.1) is a web-based tool that analyzes the functional impact of protein mutations. When the score >2.5, the mutation is considered as neutral and has no effect on the protein. When the score is <2.5, the mutation is considered as deleterious and consequently has a deleterious effect on the protein. SNAP2 is a web-based server that is used to predict the functional effect of a mutation. Based on a neural network method, SNAP2 predicts the changes due to a nonsynonymous single-nucleotide polymorphism (nsSNP) on the secondary structure and compares the solvent accessibility of the native and mutated protein to distinguish them into effect (+100, strongly predicted) or neutral (–100, strongly predicted). Mutation Assessor is a web-based tool that is used to predict the functional effect of a mutation on a protein based on an evolutionary conservation approach. We used 5 bioinformatic servers for the second study, namely, SIFT, PolyPhen, PANTHER (Protein Analysis Through Evolutionary Relationships), SNP&GO (Single Nucleotide Polymorphisms and Gene Ontology), and PhD-SNP (Predictor of Human Deleterious Single Nucleotide Polymorphisms). PANTHER is a web-based tool for predicting nonsynonymous genetic variants that may play a causal role in human disease. PANTHER includes the Position-Specific Evolutionary Preservation tool,

which predicts deleterious or pathogenic variants based on evolutionary conservation across homologous proteins from various organisms. The reference protein sequence of humans as well as sequences from 100 other species is used for these predictions. SNP&GO is a web-based tool using support vector machine (SVM) methods to predict whether a mutation is disease-related based on the protein sequence. The protein sequence is formatted in FASTA, and results are categorized as either neutral or disease-related, with a reliability index greater than 5 indicating a disease-causing mutation. PhD-SNP is also based on SVM and predicts whether a point mutation is a neutral polymorphism or associated with genetic disorders. It uses unique information derived from protein sequence, phylogenetic relationships, and the protein's encoded function to determine whether the variant is disease-associated. This part was inspired by a study by Saih et al [28], who used SIFT, PolyPhen, and PROVEAN consecutively [29-36].

Prediction of Mutation Effect on Stability and Structure of ADAM17 Protein Using I-Mutant, MUpro, and iStable

I-Mutant is a predictor of the effect of a single mutation on protein stability using protein sequences or structures and is an SVM tool based on predicting automatically the stability changes of a protein upon single-point mutations. A delta G>0 indicates a decrease in protein stability, while delta G<0 suggests increased stability. MUpro is a web-based tool used to predict the effect of mutations on the stability (increase or decrease) of a protein. The score >0 means that the mutation results in an increase in the stability of the protein, while a score <0 means that the mutation decreases the stability of the protein. iStable (Integrated predictor for protein stability change upon single mutation) analyzes protein stability using sequence information and predictions from different predictors. In this sequential analysis, 3 predictors are used: I-Mutant2.0, MUpro, and iStable [37-39].

Conservation and Conserved Domain Analysis Using ConSurf and InterPro

ConSurf is a web server that is used for estimating the evolutionary conservation of amino or nucleic acid positions in a protein or DNA or RNA molecule based on the phylogenetic relations between homologous sequences and also for identifying functional regions. A conservation score ranging from 1 to 3 is considered variable, 5 to 6 is intermediate, and 7 to 9 indicates high conservation. InterPro is a web-based server that is used to identify the location of nsSNPs on conserved domains. InterPro recognizes protein motifs and domains, enabling functional characterization of the protein using its database of protein families, domains, and functional site [40-43].

ADAM17 Modeling Using SWISS-MODEL Server and Sparks-X

SWISS-MODEL is a web server dedicated to protein structure homology modeling at different levels of complexity. 3D protein structures provide valuable insights into their molecular function and inform a broad spectrum of applications in life science research. Modeling of protein structures usually requires extensive expertise in structural biology and the use of highly

specialized computer programs for each of the individual steps of the modeling process, and templates selected based on sequence identity and Global Model Quality Estimation score. Sparks-x is a fold-recognition method used to generate 3D protein structures. This tool improves structure prediction through enhanced alignment scoring and the use of SPINE-X, which boosts predictions of secondary structure, backbone torsion angles, and solvent-accessible surfaces [44,45].

Validation of ADAM17 Models Using PROCHECK

PROCHECK (Programs to Check the Stereochemical Quality of Protein Structures) assesses the stereochemical quality of protein structures. It produces PostScript plots analyzing the global and residue-level geometry. PROCHECK-NMR is used to check the quality of structures resolved by nuclear magnetic resonance [46,47].

Prediction of Mutation Effect on Protein Structure Using Project HOPE Server

HOPE (Have (y)Our Protein Explained) server is based on the automatic analysis of mutants, which can provide more clarification of the structural and functional effects on it. HOPE is an application that analyzes mutations automatically and explains the molecular source (origin) of a disease caused by it [48].

Visualization of ADAM17 Native and Mutants Using PyMol

PyMol (version 2.5) is a molecular visualization program used to generate high-quality 3D images of proteins, as well as to edit molecular structures, perform ray tracing, and create molecular animations. PyMol (version 1.2r3pre; Schrödinger, LLC) is written in Python, one of the most popular programming languages, and can be easily extended through Python-based plugins.

All computational analyses were performed using default parameters except where otherwise noted. Tools were accessed from March to November 2024.

Results

Overview

In the first study, a total of 12,042 SNPs were collected from the Ensembl database. PROVEAN, PolyPhen-2, Mutation Assessor, SNAP2, and SIFT programs were used to predict the functional effects of mutations on ADAM17, while MUpro and I-Mutant tools were used to predict the mutation effects on protein stability. Additionally, SWISS-MODEL, ConSurf, and HOPE project were used to evaluate the mutation effects on protein function, structure, and protein-protein interactions. Ten various bioinformatics programs and tools are used to predict the mutation effects during this analysis, as relying on a single program or server is insufficient for accurately assessing mutation impact on proteins.

Among all collected SNPs, only those variants related to SARS-CoV-2 (positions 652 to 658) were extracted. Four missense variants were selected and analyzed using bioinformatics approaches.

Prediction of Deleterious nsSNPs Using SIFT, PolyPhen, PROVEAN, SNAP2, Mutation Assessor, PANTHER, SNP&GO, and PhD-SNP

Of the initial 12,042 SNPs analyzed to predict deleterious effects on the ADAM17 protein, all were first submitted to SIFT; according to SIFT, 88 of these mutations were predicted to be deleterious (index score from 0 to 0.02). These 88 SNPs were subsequently analyzed with PROVEAN, and the results of PROVEAN showed that 60 SNPs were predicted deleterious. Similarly, when analyzed with PolyPhen, 48 of the SNPs were found to be probably damaging, with scores >0.9.

Next, the same 88 SNPs were analyzed using SNAP2, which predicted that 75 SNPs would have functional impacts on the ADAM17 protein with a score more than 1. Finally, Mutation Assessor identified 66 SNPs with a medium functional impact. In summary, from all 12,042 SNPs, only 7 mutations, namely, P556L, G550D, V483A, G479E, G349E, T339P, and D232E, were predicted to have a high functional impact on the ADAM17 protein by all computational tools (Tables 1 and 2).

Table . Prediction of deleterious nonsynonymous single-nucleotide polymorphisms of the ADAM17[a] gene using SIFT[b] and PolyPhen.

Variant ID	Amino acid mutation	SIFT score	SIFT class	PolyPhen score	PolyPhen class
rs1394373815	P 556 L	0	Deleterious	0.997	Probably damaging
rs542316178	G 550 D	0	Deleterious	0.987	Probably damaging
rs777478676	V 483 A	0	Deleterious	0.987	Probably damaging
rs951262662	G 479 E	0	Deleterious	0.996	Probably damaging
rs1192348585	G 349 E	0.01	Deleterious	1	Probably damaging
rs1157021454	T 339 P	0	Deleterious	1	Probably damaging
rs768704961	D 232 E	0	Deleterious	1	Probably damaging

[a]ADAM: A disintegrin and metalloprotease.

[b]SIFT: Sorting Intolerant From Tolerant.

Table . Prediction of deleterious nonsynonymous single-nucleotide polymorphisms of ADAM17[a] gene using PROVEAN[b], Mutation Assessor, and SNAP2[c].

Variant ID	Amino acid mutation	PROVEAN		Mutation Assessor		SNAP2	
		Score	Prediction	Score	Prediction	Score	Prediction
rs1394373815	P 556 L	−8.978	Deleterious	4.095	High	37	Effect
rs542316178	G 550 D	−6.068	Deleterious	4.83	High	76	Effect
rs777478676	V 483 A	−3.496	Deleterious	4.215	High	43	Effect
rs951262662	G 479 E	−7.119	Deleterious	4.78	High	89	Effect
rs1192348585	G 349 E	−7.269	Deleterious	3.83	High	83	Effect
rs1157021454	T 339 P	−5.606	Deleterious	3.735	High	69	Effect
rs768704961	D 232 E	−3.795	Deleterious	3.79	High	84	Effect

[a]ADAM: A disintegrin and metalloprotease.

[b]PROVEAN: Protein Variation Effect Analyzer.

[c]SNAP2: Screening for Non-Acceptable Polymorphisms 2.

For the second study, 4 SARS-CoV-2–related nsSNPs (Q658H, D657G, D654N, and F652L) were submitted to SIFT. According to SIFT, mutations Q658H, D657G, and D654N were predicted to be deleterious (with index scores between 0 and 0.01), and F652L to be tolerated.

PolyPhen classified all 4 mutations as benign, while PANTHER identified D657G and D654N as likely damaging (score>0.57), and Q658H and F652L as possibly damaging (score~0.5).

The 4 nsSNPs were submitted to the SNP&GO program, which indicated that these SNPs would not have effects related to human diseases with scores above 0. The same nsSNPs were also analyzed using PhD-SNP, where the results indicated that the mutation D657G might have a pathogenic impact with a score of 3, while the other 3 mutations (Q658H, D654N, and F652L) were predicted to have neutral impacts (Tables 3 and 4).

Table . Prediction of the deleterious effects of nonsynonymous single-nucleotide polymorphisms related to SARS-CoV-2 using SIFT[a] and PolyPhen.

Variant ID	Amino acid mutation	SIFT score	SIFT class	PolyPhen score	PolyPhen class
rs765452935	Q658H	0.01	Deleterious	0.003	Benign
rs144657795	D657G	0	Deleterious	0.097	Benign
rs758594009	D654N	0	Deleterious	0.063	Benign
rs780262610	F652L	0.81	Tolerated	0.015	Benign

[a]SIFT: Sorting Intolerant From Tolerant.

Table . Prediction of the deleterious effects of nonsynonymous single-nucleotide polymorphisms related to SARS-CoV-2 using PANTHER[a], SNP&GO[b], and PhD-SNP[c].

Variant ID	Amino acid mutation	PANTHER			SNP&GO		PhD-SNP	
		PSEP[d]	Prediction	Pdel	Reliability index	Prediction	Reliability index	Prediction
rs765452935	Q658H	220	Possibly damaging	.50	6	Neutral	3	Neutral
rs144657795	D657G	455	Probably damaging	.57	2	Neutral	3	Disease
rs758594009	D654N	1036	Probably damaging	.85	1	Neutral	5	Neutral
rs780262610	F652L	220	Possibly damaging	.50	8	Neutral	6	Neutral

[a]PANTHER: Protein Analysis Through Evolutionary Relationships.

[b]SNP&GO: Single Nucleotide Polymorphisms and Gene Ontology.

[c]PhD-SNP: Predictor of Human Deleterious Single Nucleotide Polymorphisms.

[d]PSEP: Position-Specific Evolutionary Preservation.

Prediction of Mutation Effects on the Protein Energy and Stability Using I-Mutant, MUpro, and iStable Servers

MUpro results showed that 6 of the 7 selected mutations (G550D, V483A, G479E, G349E, T339P, and D232E) were predicted to decrease the stability of the ADAM17 protein, while the mutation P556L was predicted to increase the stability of the ADAM17 protein. Then, the 7 selected mutations were submitted to I-Mutant. Results of I-Mutant showed that all the 7 mutations (P556L, G550D, V483A, G479E, G349E, T339P, and D232E) were predicted to decrease the stability of ADAM17 protein (Table 5).

Table . Prediction of ADAM17[a] stability using MUpro and I-Mutant tools.

Mutation	MUpro		I-Mutant	
	Delta G	Prediction	Delta G	Prediction
P 556 L	0.20231915	Increase	−0.51	Decrease
G 550 D	−0.35202875	Decrease	−0.76	Decrease
V 483 A	−2.5029256	Decrease	−1.4	Decrease
G 479 E	−0.20347724	Decrease	−0.8	Decrease
G 349 E	−0.096611232	Decrease	−0.48	Decrease
T 339 P	−1.4271878	Decrease	−0.63	Decrease
D 232 E	−1.1774927	Decrease	−0.58	Decrease

[a]ADAM: A disintegrin and metalloprotease.

For the second study, iStable was used to predict the stability of these 4 mutations on the ADAM17 protein. The iStable results indicated that these 4 residues (Q658H, D657G, D654N, and F652L) were predicted to decrease the protein's stability (Table 6).

Table . Stability analysis of ADAM17[a] mutations related to SARS-CoV-2 using iStable.

Mutation	Confidence score	Prediction
Q658H	0.671109	Decrease
D657G	0.846768	Decrease
D654N	0.799807	Decrease
F652L	0.808582	Decrease

[a]ADAM: A disintegrin and metalloprotease.

Prediction of Phylogenetic Conservation Using ConSurf and Study of Conserved Domains Using InterPro

The ConSurf analysis showed that all 7 substitutions are extremely conserved with a conservation score of 9. Six of these mutations (P556L, G550D, V483A, G479E, T339P, and D232E) were predicted to be exposed and functional, while the mutation G349E was predicted to be buried and structural. In the second study, ConSurf results showed 3 substitutions (Q658H, D657G, and D654N) to be highly conserved with a score of 8 and were predicted to be exposed and functional, while substitution (F652L) was very conserved with a score of 7 and predicted to be buried. The full visualization of the ConSurf-based phylogenetic conservation analysis of ADAM17 is provided in Multimedia Appendix 1.

For the second study, the InterPro domains identified include IPR032029, which indicates the ADAM17 proximal membrane domain (580-642), IPR001762, which indicates the disintegrin domain (475-563), IPR034025, which indicates the catalytic domain 17 (223-477) ADAM10/ADAM17, IPR001590 peptidase M12B, and ADAM/reprolysin, and IPR002870, which indicates peptidase M12B propeptide (48-167; Figure 1).

Figure 1. Identification of the ADAM17 protein domain using the InterPro server. ADAM: A disintegrin and metalloprotease.
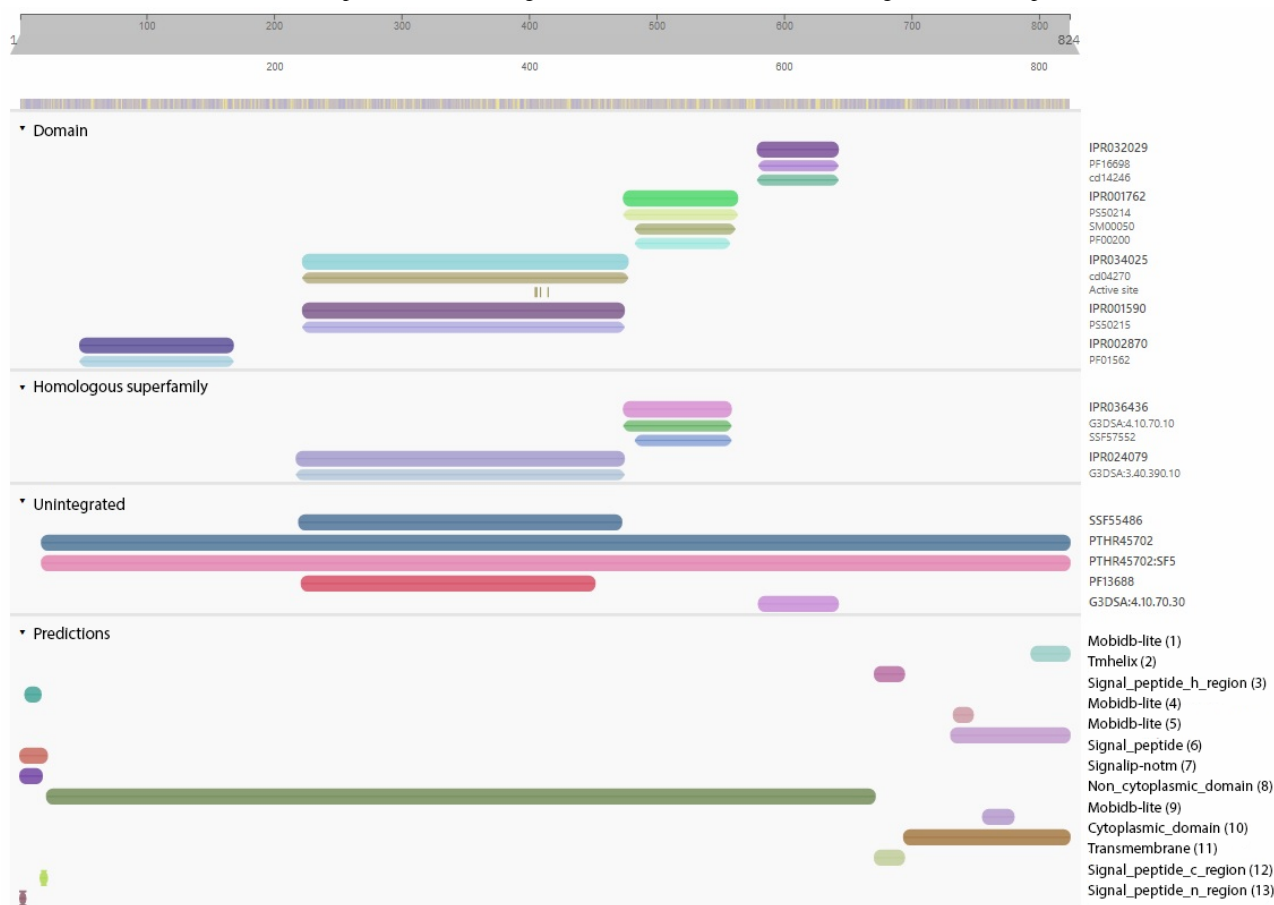
Total number of residues is 824. In the native ADAM17 structure, 604 (82.9%) of amino acids were in the favorable region, 119 (16.3%) in the allowed region, and 6 (0.8%) in the disallowed region. However, in the mutant structure (eg, Q658H), the percentage of the favorable region decreased, and the disallowed region increased, which can be explained by the fact that the mutation impacts the protein and its modeling. All Ramachandran plot results by PROCHECK are provided in Multimedia Appendices 2-6 (Figure 2 and Table 7).

Figure 2. Ramachandran plot of the native model generated by PROCHECK. PROCHECK: Programs to Check the Stereochemical Quality of Protein Structures.
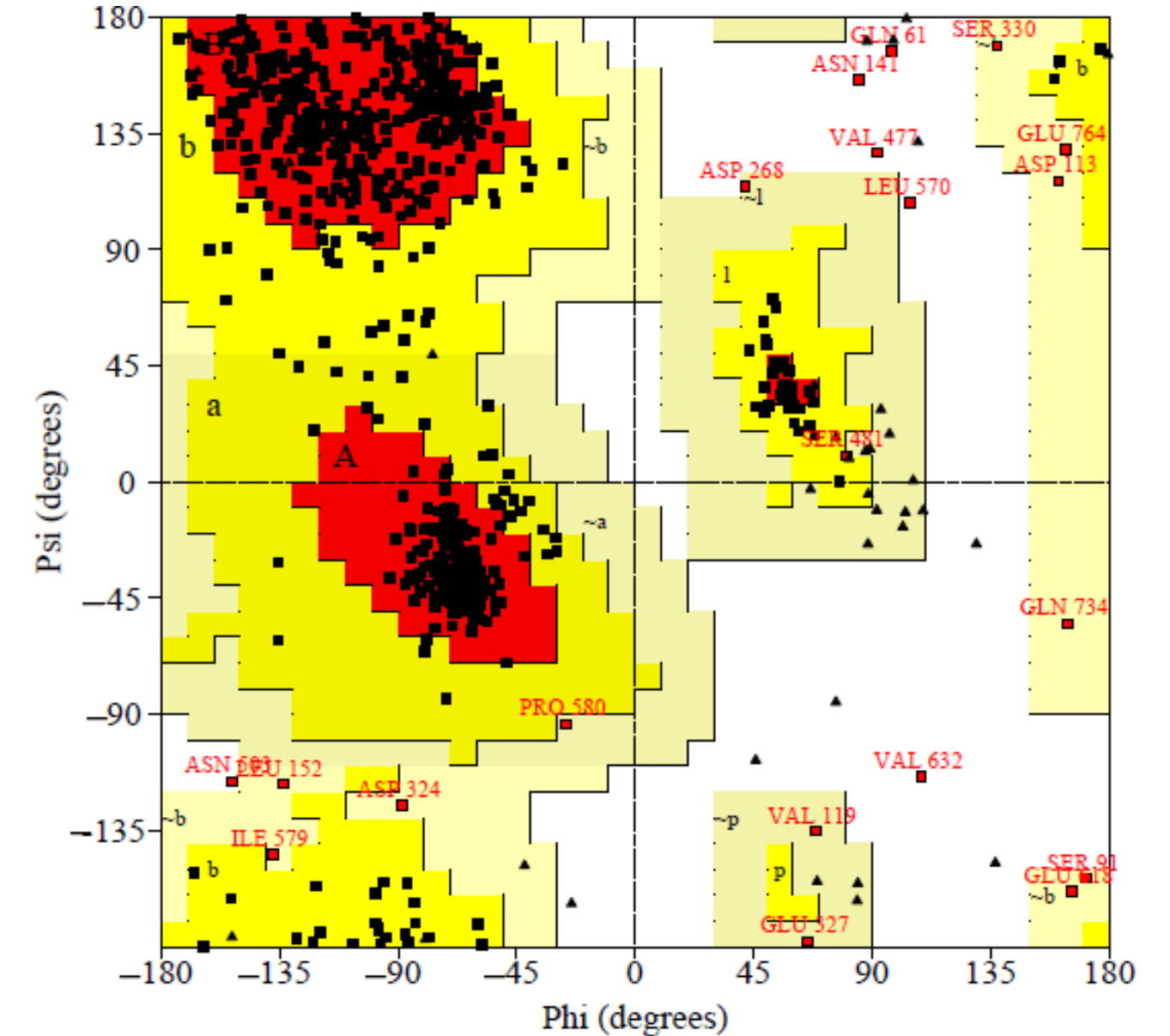
Table . Percentage of different regions for each mutation using PROCHECK[a].

Mutation	Favored region, n (%)	Allowed region, n (%)	Disallowed region, n (%)
Native	604 (82.9)	119 (16.3)	6 (0.8)
Q658H	585 (80)	129 (17.6)	17 (2.3)
D657G	608 (83.3)	112 (15.3)	10 (1.4)
D654N	600 (82.1)	121 (16.6)	10 (1.4)
F652L	609 (83.3)	108 (14.8)	14 (1.9)

[a]PROCHECK: Programs to Check the Stereochemical Quality of Protein Structures.

Modeling of ADAM17 Using SWISS-MODEL and Sparks-X

In this study, we used the SWISS-MODEL server to construct the 3D structure of the native and 7 mutants of the ADAM17 protein. We used 2dw0.1. A (crystal structure of vesicle-associated membrane protein-associated protein 2 from Crotalus Atrox venom [Form 2‐1 crystal]) as a template with a sequence identity equal to 35.21% and resolution of 2.15 A°. In the second study, the Sparks-X server was used to generate the 3D structure of both native and the 4 mutants of ADAM17 molecules.
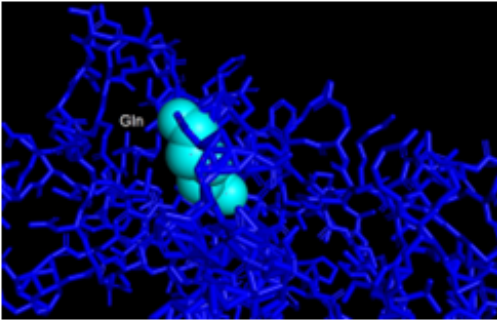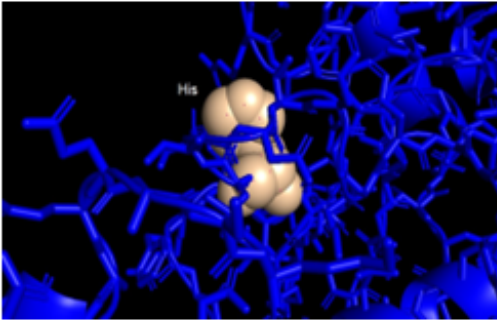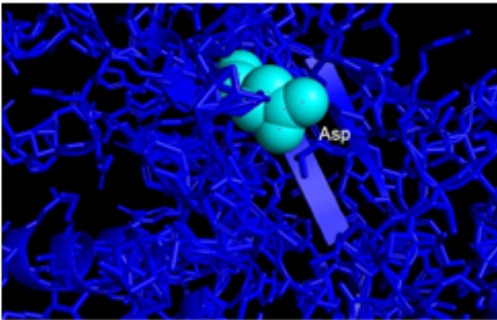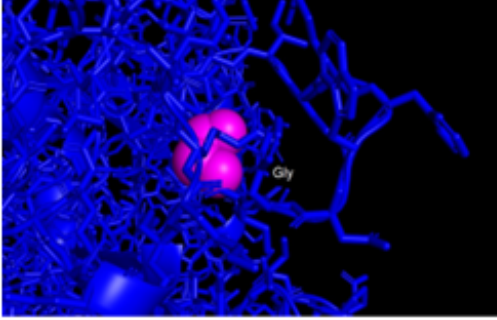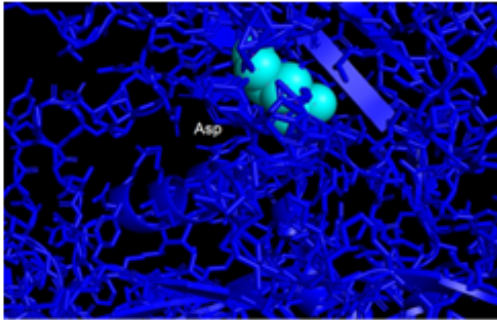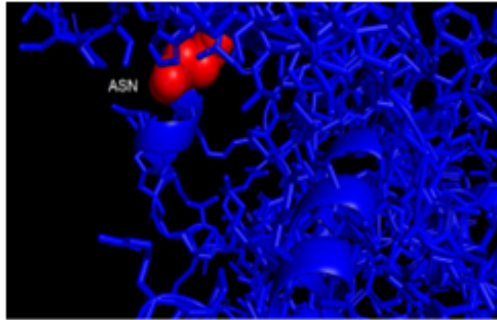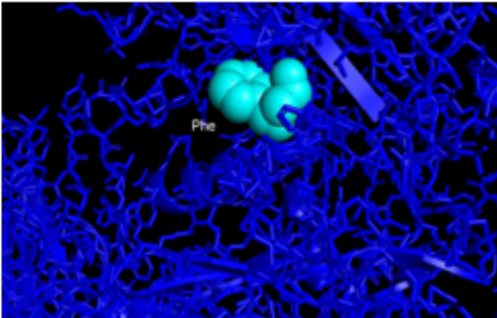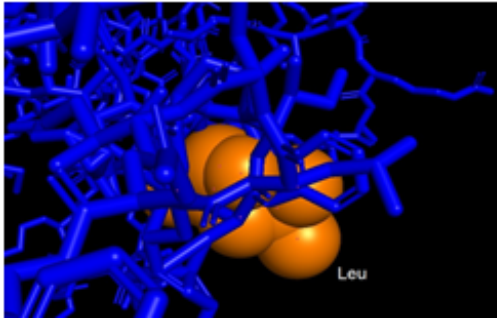
Visualization of ADAM17 Mutations Using the PyMol Program

The 3D structures of the ADAM17 native and mutant models were visualized using PyMol. Structural similarities and differences between the ADAM17 native and its mutants are shown in Figures 3 and 4.

Figure 3. Visualization of ADAM17 native and mutations using the PyMol program. ADAM: A disintegrin and metalloprotease.

Figure 4. Visualization of ADAM17 native and mutations related to SARS-CoV-2 using the PyMol program. ADAM: A disintegrin and metalloprotease.

Prediction of Structural Effects of Mutations in ADAM17 Using the HOPE Server

rs1394373815

The sizes are different between the amino acids of wild-type and mutant. The mutant residue is larger than the wild-type residue, and this might lead to bumps. Prolines are known to have a very rigid structure, sometimes forcing the backbone in a specific conformation. Possibly, the mutation changes a proline with such a function into another residue, thereby disturbing the local structure. The residue is found on the surface of the protein.

rs542316178

The charge of the mutant amino acid differs from that of the wild-type. The mutation introduces a charge, and this can cause repulsion of ligands or other residues with the same charge. The sizes are different between the amino acids of wild-type and mutant. The mutant residue is larger, and this might lead to bumps. The torsion angles for this residue are unusual. Only glycine is flexible enough to make these torsion angles, and mutation into another residue will force the local backbone into an incorrect conformation and will disturb the local structure.

rs777478676

The sizes are different between the amino acids of wild-type and mutant. The mutant residue is smaller, and this might lead to loss of interactions. The mutant residue is situated close to a position that is highly conserved. The mutation introduces an amino acid with different properties, which can disturb this domain and abolish its function.

rs951262662

The mutant amino acid carries a charge that differs from the wild-type counterpart. Because the mutation adds a charge, it can repel ligands or residues with similar charges. The sizes of the amino acids in the mutant and wild-type also differ, with the mutant residue being bulkier, which may result in steric hindrance. The torsion angles for this residue are unusual; glycine is the only amino acid flexible enough to adopt these angles. Mutation to a different residue will force the local backbone into an improper conformation and disturb the surrounding structure.

rs1192348585

The charge of the mutant amino acid contrasts with that of the wild-type. This mutation introduces a charge that may cause repulsive interactions with ligands or other residues carrying the same charge. Size differences between the mutant and wild-type amino acids are notable, as the mutant residue is larger and could cause clashes. The torsion angles for this residue are uncommon; glycine alone has the necessary flexibility to maintain such angles. Mutating to any other residue will impose strain on the local backbone, leading to an incorrect conformation and disruption of the local structural environment.

rs1157021454

The wild-type and mutant residues have different levels of hydrophobicity. At this location, the mutation adds a more hydrophobic residue. This may cause hydrogen bonds to break or disturb the proper. The wild-type residue and the mutant residue share certain properties. This mutation might occur in some rare cases, but it is more likely that the mutation is damaging to the protein.

rs768704961

The sizes are different between the amino acids of wild-type and mutant. The mutant residue is larger, and this might lead to bumps. The mutation is located within a domain and annotated in UniProt as peptidase M12B. Only this residue type was found at this position. Mutation of a 100% conserved residue is usually damaging for the protein.

Discussion

Principal Findings

In this study, different tools were used to identify the most deleterious nsSNPs of the ADAM17 protein, namely, SIFT, PolyPhen, PROVEAN, SNAP2, Mutation Assessor, I-Mutant, MUpro, and ConSurf; these tools were selected according to the following steps: pathogenicity study, stability, and conservation study. Parameters like accuracy, sensitivity, and specificity were chosen to assess their predictive abilities. Without these parameters, it will not be possible to completely evaluate the accuracy of a test.

In this bioinformatic study, we identified 7 nsSNPs (P556L, G550D, V483A, G479E, G349E, T339P, and D232E) from the entire residues of ADAM17. These nsSNPs were predicted by 5 tools: SIFT score of all these mutations ≈0 and classed as deleterious effect on the protein, PolyPhen score≈1 and classed in the probably damaging class, PROVEAN score of all of these mutations is negative (<−3.4) and was predicted deleterious, Mutation Assessor score of all of these mutations is positive (>3.7) and predicted to have a high functional impact on the protein, and SNAP2 score results were positive (>42) and classed to have a functional effect on the protein. In addition, we also evaluated protein stability using I-Mutant and MUpro. I-Mutant predicted that all 7 mutations would decrease the protein stability. MUpro results agreed for most mutations, except that the P556L mutation was predicted to increase the stability. Maximum conservation score by ConSurf means that all mutations were predicted to have functional effects, except the G349E mutation, which was predicted to have a structural effect on ADAM17.

These mutations (P556L, G550D, V483A, G479E, G349E, T339P, and D232E; rs1394373815, rs542316178, rs777478676, rs951262662, rs1192348585, rs1157021454, and rs768704961) are novel for their impact on ADAM17 structure, function, and stability.

The second part of this study focuses on nsSNPs that may be directly related to SARS-CoV-2 due to their positions within the ADAM17 protein. We analyzed 4 nsSNPs of interest (Q658H, D657G, D654N, and F652L), which were found to have the highest conservation scores and were predicted to be deleterious and reducing the stability of ADAM17. We hypothesized that these residues (Q658, D657, D654, and F652) are actively involved in the cleavage of ACE2 by ADAM17, and a mutation at any of these positions could disrupt the entire cleavage process. To support this hypothesis, we used a series of tools to assess the pathogenicity of these mutants.

Comparison to Prior Studies

More than 80 distinct substrates have been discovered to be processed by ADAM17, also referred to as TNFα-converting enzyme, since its discovery. ADAM17, like most other ADAM relations, is understood to process single-spanning membrane proteins like growth factors, cytokines, receptors, chemokines, and regulators of neurological processes and diseases, and ADAM17 processes more than 80 substrates, and lots of them are linked to inflammatory and cancerous diseases. More

recently, molecules important to tumor immunosurveillance have been found to be substrates for ADAM17, and research on the shedding events that this enzyme orchestrates has produced new theories of resistance to common cancer treatments. While ADAM17 features a wide range of substrate profiles, it typically only becomes active in response to triggers that cause disease states, making it a good target for a treatment approach.

The study by Pavlenko et al [49] has demonstrated that there are important ADAM17 residues, namely, R177C, D616N, D657A, and R725H, that play important roles in different cancer types. The R177C mutation affects the prodomain of ADAM17 and causes cecum and central nervous system cancer, the D616N affects the membrane-proximal domain and causes cancer in colon and uterus, the D657A residue affects the membrane-proximal domain and causes colon cancer, and R725H residue affects the cytoplasmic domain and causes colon cancer [49].

Mutations in the ADAM17 gene have been associated with neonatal inflammatory skin and bowel disease, a condition characterized by inflammatory features with neonatal onset, affecting the skin, hair, and gastrointestinal tract. The skin lesions involve perioral and perianal erythema, psoriasiform erythroderma, with flares of erythema, scaling, and widespread pustules. Gastrointestinal symptoms include malabsorptive diarrhea that is exacerbated by intercurrent gastrointestinal infections. The hair is brief or broken; therefore, the eyelashes and eyebrows are wiry and disorganized. The results of this study may be applicable for the analysis of novel missense variants of the ADAM17 gene.

Several studies have demonstrated that residues located between positions 652 and 659 catalyze the shedding of the ACE2 ectodomain by ADAM17 [50,51]. Recent advances in deep learning, such as self-supervised learning, provide promising avenues for enhancing the predictive capabilities of bioinformatics tools like the ones implemented in this work. In addition, application of federated learning with its privacy-preserving analytics approach applied to Internet of Things in smart health care could increase the scope for computational approaches such as that done for the ADAM17 variant [52,53].

Limitations

Our study has several limitations. First, it is based entirely on computational analysis using predictive tools and servers, which may have many varying confidence levels and potential false positive rates that we did not fully address. In addition, the structural analysis using PyMol revealed visual differences between native and mutant proteins, but their functional implications of these structural changes are not thoroughly explored or explained. Second, the 7 deleterious variants and the 4 variants that have a relation with SARS-CoV-2 infection should be confirmed with future laboratory experiments and clinical wet laboratory approaches to figure out the mechanism of these mutations.

Conclusions

In this in silico study of the high-risk missense variants of ADAM17, we identified 7 nsSNPs (P556L, G550D, V483A, G479E, G349E, T339P, and D232E) as the most deleterious mutations in the ADAM17 gene. All 7 mutations were predicted to have damaging effects on the structure, function, and stability of the ADAM17 protein. This study represents the first in silico analysis that evaluates the effect of these missense variants on the function and structure of ADAM17, and these results still require validation with in vitro experiments.

To support this study, in vitro experiments should be conducted to confirm the in silico results. For future research, our results confirm the impact of the 4 named mutations (Q658H, D657G, D654N, and F652L) on the pathology related to SARS-CoV-2, which strongly reinforces the role of ADAM17 in the ectodomain shedding process of ACE2.

Our findings form a basis for understanding the potential implications of ADAM17 variants on disease, which may lead to earlier diagnosis, assessment of risk for progression of related diseases, and may help inform future therapeutic targeting.

Data Availability

All data generated or analyzed during this study are included in this published paper and Multimedia Appendices 1-6.

Authors' Contributions

AM conceptualized and led the study, conducted data analysis, and drafted the manuscript. AS, LW, and AA contributed to data collection, interpretation, and manuscript review.

Conflicts of Interest

None declared.

Multimedia Appendix 1
ConSurf results.

XSL•FO

RenderX

[PDF File, 56 KB - bioinform_v6i1e72133_app1.pdf]

Multimedia Appendix 2
Ramachandran for ADAM17 wild. ADAM: A disintegrin and metalloprotease.
[PDF File, 17 KB - bioinform_v6i1e72133_app2.pdf]

Multimedia Appendix 3
Ramachandran ADAM17 mutation D654N. ADAM: A disintegrin and metalloprotease.
[PDF File, 17 KB - bioinform_v6i1e72133_app3.pdf]

Multimedia Appendix 4
Ramachandran ADAM17 mutation D657G. ADAM: A disintegrin and metalloprotease.
[PDF File, 17 KB - bioinform_v6i1e72133_app4.pdf]

Multimedia Appendix 5
Ramachandran ADAM17 mutation F652L. ADAM: A disintegrin and metalloprotease.
[PDF File, 17 KB - bioinform_v6i1e72133_app5.pdf]

Multimedia Appendix 6
Ramachandran ADAM17 mutation Q658H. ADAM: A disintegrin and metalloprotease.
[PDF File, 17 KB - bioinform_v6i1e72133_app6.pdf]

References

1. Zhong S, Khalil RA. A disintegrin and metalloproteinase (ADAM) and ADAM with thrombospondin motifs (ADAMTS) family in vascular biology and disease. Biochem Pharmacol 2019 Jun;164:188-204. [doi: 10.1016/j.bcp.2019.03.033] [Medline: 30905657]
2. Giebeler N, Zigrino P. A disintegrin and metalloprotease (ADAM): historical overview of their functions. Toxins (Basel) 2016 Apr 23;8(4):122. [doi: 10.3390/toxins8040122] [Medline: 27120619]
3. Edwards DR, Handsley MM, Pennington CJ. The ADAM metalloproteinases. Mol Aspects Med 2008 Oct;29(5):258-289. [doi: 10.1016/j.mam.2008.08.001] [Medline: 18762209]
4. Liu PCC, Liu X, Li Y, et al. Identification of ADAM10 as a major source of HER2 ectodomain sheddase activity in HER2 overexpressing breast cancer cells. Cancer Biol Ther 2006 Jun;5(6):657-664. [doi: 10.4161/cbt.5.6.2708] [Medline: 16627989]
5. Souza JSM, Lisboa ABP, Santos TM, et al. The evolution of ADAM gene family in eukaryotes. Genomics 2020 Sep;112(5):3108-3116. [doi: 10.1016/j.ygeno.2020.05.010] [Medline: 32437852]
6. Blobel CP. Metalloprotease-disintegrins: links to cell adhesion and cleavage of TNF alpha and Notch. Cell 1997 Aug 22;90(4):589-592. [doi: 10.1016/s0092-8674(00)80519-x] [Medline: 9288739]
7. Gooz M. ADAM-17: the enzyme that does it all. Crit Rev Biochem Mol Biol 2010 Apr;45(2):146-169. [doi: 10.3109/10409231003628015] [Medline: 20184396]
8. Xu J, Mukerjee S, Silva-Alves CRA, et al. A disintegrin and metalloprotease 17 in the cardiovascular and central nervous systems. Front Physiol 2016;7:469. [doi: 10.3389/fphys.2016.00469] [Medline: 27803674]
9. Peiretti F, Canault M, Morange P, Alessi MC, Nalbone G. The two sides of ADAM17 in inflammation: implications in atherosclerosis and obesity. Med Sci (Paris) 2009 Jan;25(1):45-50. [doi: 10.1051/medsci/200925145] [Medline: 19154693]
10. Canault M, Leroyer AS, Peiretti F, et al. Microparticles of human atherosclerotic plaques enhance the shedding of the tumor necrosis factor-alpha converting enzyme/ADAM17 substrates, tumor necrosis factor and tumor necrosis factor receptor-1. Am J Pathol 2007 Nov;171(5):1713-1723. [doi: 10.2353/ajpath.2007.070021] [Medline: 17872973]
11. Moss ML, Jin SL, Milla ME, et al. Cloning of a disintegrin metalloproteinase that processes precursor tumour-necrosis factor-alpha. Nature New Biol 1997 Feb 20;385(6618):733-736. [doi: 10.1038/385733a0] [Medline: 9034191]
12. Sharma A, Bender S, Zimmermann M, Riesterer O, Broggini-Tenzer A, Pruschy MN. Secretome signature identifies ADAM17 as novel target for radiosensitization of non-small cell lung cancer. Clin Cancer Res 2016 Sep 1;22(17):4428-4439. [doi: 10.1158/1078-0432.CCR-15-2449] [Medline: 27076628]
13. Sternlicht MD, Sunnarborg SW, Kouros-Mehr H, Yu Y, Lee DC, Werb Z. Mammary ductal morphogenesis requires paracrine activation of stromal EGFR via ADAM17-dependent shedding of epithelial amphiregulin. Development 2005 Sep;132(17):3923-3933. [doi: 10.1242/dev.01966] [Medline: 16079154]
14. Li Y, Brazzell J, Herrera A, Walcheck B. ADAM17 deficiency by mature neutrophils has differential effects on L-selectin shedding. Blood 2006 Oct 1;108(7):2275-2279. [doi: 10.1182/blood-2006-02-005827] [Medline: 16735599]

15. Hoffmann M, Kleine-Weber H, Schroeder S, et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell 2020 Apr 16;181(2):271-280. [doi: 10.1016/j.cell.2020.02.052] [Medline: 32142651]

16. Heurich A, Hofmann-Winkler H, Gierer S, Liepold T, Jahn O, Pöhlmann S. TMPRSS2 and ADAM17 cleave ACE2 differentially and only proteolysis by TMPRSS2 augments entry driven by the severe acute respiratory syndrome coronavirus spike protein. J Virol 2014 Jan;88(2):1293-1307. [doi: 10.1128/JVI.02202-13] [Medline: 24227843]

17. Lambert DW, Yarski M, Warner FJ, et al. Tumor necrosis factor-alpha convertase (ADAM17) mediates regulated ectodomain shedding of the severe-acute respiratory syndrome-coronavirus (SARS-CoV) receptor, angiotensin-converting enzyme-2 (ACE2). J Biol Chem 2005 Aug 26;280(34):30113-30119. [doi: 10.1074/jbc.M505111200] [Medline: 15983030]

18. Monteil V, Kwon H, Prado P, et al. Inhibition of SARS-CoV-2 infections in engineered human tissues using clinical-grade soluble human ACE2. Cell 2020 May 14;181(4):905-913. [doi: 10.1016/j.cell.2020.04.004] [Medline: 32333836]

19. Khan JY, Khondaker MTI, Hoque IT, et al. Toward preparing a knowledge base to explore potential drugs and biomedical entities related to COVID-19: automated computational approach. JMIR Med Inform 2020 Nov 10;8(11):e21648. [doi: 10.2196/21648] [Medline: 33055059]

20. Markovič R, Ternar L, Trstenjak T, Marhl M, Grubelnik V. Cardiovascular comorbidities in COVID-19: comprehensive analysis of key topics. Interact J Med Res 2024 Jul 24;13:e55699. [doi: 10.2196/55699] [Medline: 39046774]

21. Lami F, Elfadul M, Rashak H, et al. Risk factors of COVID-19 critical outcomes in the Eastern Mediterranean Region: multicountry retrospective study. JMIR Public Health Surveill 2022 Mar 15;8(3):e32831. [doi: 10.2196/32831] [Medline: 34736222]

22. Cho K, Park S, Kim EY, et al. Immunogenicity of COVID-19 vaccines in patients with diverse health conditions: a comprehensive systematic review. J Med Virol 2022 Sep;94(9):4144-4155. [doi: 10.1002/jmv.27828] [Medline: 35567325]

23. Abbas Z, Rehman MU, Tayara H, Lee SW, Chong KT. m5C-Seq: machine learning-enhanced profiling of RNA 5-methylcytosine modifications. Comput Biol Med 2024 Nov;182:109087. [doi: 10.1016/j.compbiomed.2024.109087] [Medline: 39232403]

24. ADAM17 gene (ENSG00000151694)—variation table. Ensembl. URL: https://www.ensembl.org/Homo_sapiens/Gene/Variation_Gene/Table?db=core;g=ENSG00000151694;r=2:9488486-9556732 [accessed 2025-08-13]

25. ADAM17—disintegrin and metalloproteinase domain-containing protein 17 (P78536)—FASTA sequence. The UniProt Consortium. URL: https://www.uniprot.org/uniprot/P78536.fasta [accessed 2025-08-18]

26. Howe KL, Achuthan P, Allen J, et al. Ensembl 2021. Nucleic Acids Res 2021 Jan 8;49(D1):D884-D891. [doi: 10.1093/nar/gkaa942] [Medline: 33137190]

27. Bateman A, Martin MJ, Orchard S, et al. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 2021 Jan 8;49(D1):D480-D489. [doi: 10.1093/nar/gkaa1100]

28. Saih A, Baba H, Bouqdayr M, et al. In Silico analysis of high-risk missense variants in human ACE2 gene and susceptibility to SARS-CoV-2 infection. Biomed Res Int 2021;2021:6685840. [doi: 10.1155/2021/6685840] [Medline: 33884270]

29. Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res 2012 Jul;40:W452-W457. [doi: 10.1093/nar/gks539] [Medline: 22689647]

30. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. Nat Methods 2010 Apr;7(4):248-249. [doi: 10.1038/nmeth0410-248] [Medline: 20354512]

31. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics 2015 Aug 15;31(16):2745-2747. [doi: 10.1093/bioinformatics/btv195] [Medline: 25851949]

32. Hecht M, Bromberg Y, Rost B. Better prediction of functional effects for sequence variants. BMC Genomics 2015;16(Suppl 8):S1. [doi: 10.1186/1471-2164-16-S8-S1] [Medline: 26110438]

33. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res 2011 Sep 1;39(17):e118. [doi: 10.1093/nar/gkr407] [Medline: 21727090]

34. Thomas PD, Campbell MJ, Kejariwal A, et al. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 2003 Sep;13(9):2129-2141. [doi: 10.1101/gr.772403] [Medline: 12952881]

35. Capriotti E, Calabrese R, Fariselli P, Martelli PL, Altman RB, Casadio R. WS-SNPs&GO: a web server for predicting the deleterious effect of human protein variants using functional annotation. BMC Genomics 2013;14(Suppl 3):S6. [doi: 10.1186/1471-2164-14-S3-S6] [Medline: 23819482]

36. Capriotti E, Fariselli P. PhD-SNPg: a webserver and lightweight tool for scoring single nucleotide variants. Nucleic Acids Res 2017 Jul 3;45(W1):W247-W252. [doi: 10.1093/nar/gkx369] [Medline: 28482034]

37. Cheng J, Randall A, Baldi P. Prediction of protein stability changes for single-site mutations using support vector machines. Proteins 2006 Mar 1;62(4):1125-1132. [doi: 10.1002/prot.20810] [Medline: 16372356]

38. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 2005 Jul 1;33:W306-W310. [doi: 10.1093/nar/gki375] [Medline: 15980478]

39. Chen CW, Lin J, Chu YW. iStable: off-the-shelf predictor integration for predicting protein stability changes. BMC Bioinformatics 2013;14(Suppl 2):S5. [doi: 10.1186/1471-2105-14-S2-S5] [Medline: 23369171]

40. Ashkenazy H, Abadi S, Martz E, et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic Acids Res 2016 Jul 8;44(W1):W344-W350. [doi: 10.1093/nar/gkw408] [Medline: 27166375]

41. Celniker G, Nimrod G, Ashkenazy H, et al. ConSurf: using evolutionary data to raise testable hypotheses about protein function. Isr J Chem 2013 Apr;53(3-4):199-206. [doi: 10.1002/ijch.201200096]

42. Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res 2010 Jul;38:W529-W533. [doi: 10.1093/nar/gkq399] [Medline: 20478830]

43. Hunter S, Apweiler R, Attwood TK, et al. InterPro: the integrative protein signature database. Nucleic Acids Res 2009 Jan;37:D211-D215. [doi: 10.1093/nar/gkn785] [Medline: 18940856]

44. Waterhouse A, Bertoni M, Bienert S, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res 2018 Jul 2;46(W1):W296-W303. [doi: 10.1093/nar/gky427] [Medline: 29788355]

45. Yang Y, Faraggi E, Zhao H, Zhou Y. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. Bioinformatics 2011 Aug 1;27(15):2076-2082. [doi: 10.1093/bioinformatics/btr350] [Medline: 21666270]

46. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr 1993 Apr 1;26(2):283-291. [doi: 10.1107/S0021889892009944]

47. Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR 1996 Dec;8(4):477-486. [doi: 10.1007/BF00228148] [Medline: 9008363]

48. Venselaar H, Te Beek TAH, Kuipers RKP, Hekkelman ML, Vriend G. Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. BMC Bioinformatics 2010 Nov 8;11:548. [doi: 10.1186/1471-2105-11-548] [Medline: 21059217]

49. Pavlenko E, Cabron AS, Arnold P, Dobert JP, Rose-John S, Zunke F. Functional characterization of colon cancer-associated mutations in ADAM17: modifications in the pro-domain interfere with trafficking and maturation. Int J Mol Sci 2019 May 4;20(9):2198. [doi: 10.3390/ijms20092198] [Medline: 31060243]

50. Sun P, Lu X, Xu C, Wang Y, Sun W, Xi J. CD-sACE2 inclusion compounds: an effective treatment for coronavirus disease 2019 (COVID-19). J Med Virol 2020 Oct;92(10):1721-1723. [doi: 10.1002/jmv.25804] [Medline: 32232976]

51. Chaudhary M. COVID-19 susceptibility: potential of ACE2 polymorphisms. Egypt J Med Hum Genet 2020;21(1):54. [doi: 10.1186/s43042-020-00099-9] [Medline: 38624559]

52. Abdulrazzaq MM, Ramaha NTA, Hameed AA, et al. Consequential advancements of self-supervised learning (SSL) in deep learning contexts. Mathematics 2024 Mar 3;12(5):758. [doi: 10.3390/math12050758]

53. Abbas SR, Abbas Z, Zahir A, Lee SW. Federated learning in smart healthcare: a comprehensive review on privacy, security, and predictive analytics with IoT integration. Healthcare (Basel) 2024 Dec 22;12(24):2587. [doi: 10.3390/healthcare12242587]

Abbreviations

ACE2: angiotensin-converting enzyme 2
ADAM: A disintegrin and metalloprotease domain
HOPE: Have (y)Our Protein Explained
nsSNP: nonsynonymous single-nucleotide polymorphism
PANTHER: Protein Analysis Through Evolutionary Relationships
PhD-SNP: Predictor of Human Deleterious Single Nucleotide Polymorphisms
PROCHECK: Programs to Check the Stereochemical Quality of Protein Structures
PROVEAN: Protein Variation Effect Analyzer
SIFT: Sorting Intolerant From Tolerant
SNAP2: Screening for Non-Acceptable Polymorphisms 2
SNP: single-nucleotide polymorphism
SNP&GO: Single Nucleotide Polymorphisms and Gene Ontology
SVM: support vector machine
TMPRSS2: transmembrane protease, serine 2
TNF-α: tumor necrosis factor alpha

XSL•FO
RenderX

XSL•FO

RenderX

Investigating Associations Between Prognostic Factors in Gliomas: Unsupervised Multiple Correspondence Analysis

Maria Eduarda Goes Job[1]; Heidge Fukumasu[1], PhD; Tathiane Maistro Malta[2], PhD; Pedro Luiz Porfirio Xavier[1], PhD

[1]Laboratory of Comparative and Translational Oncology, Department of Veterinary Medicine, School of Animal Science and Food Engineering, University of Sao Paulo, Avenida Duque de Caxias, 225, Jardim Elite, Pirassununga, Brazil

[2]Cancer Epigenomics Laboratory, Department of Clinical Analysis, Toxicology and Food Sciences, School of Pharmaceutical Sciences of Ribeirao Preto, University of Sao Paulo, Ribeirão Preto, Brazil

Corresponding Author:
Pedro Luiz Porfirio Xavier, PhD
Laboratory of Comparative and Translational Oncology, Department of Veterinary Medicine, School of Animal Science and Food Engineering, University of Sao Paulo, Avenida Duque de Caxias, 225, Jardim Elite, Pirassununga, Brazil

Abstract

Background: Multiple correspondence analysis (MCA) is an unsupervised data science methodology that aims to identify and represent associations between categorical variables. Gliomas are an aggressive type of cancer characterized by diverse molecular and clinical features that serve as key prognostic factors. Thus, advanced computational approaches are essential to enhance the analysis and interpretation of the associations between clinical and molecular features in gliomas.

Objective: This study aims to apply MCA to identify associations between glioma prognostic factors and also explore their associations with stemness phenotype.

Methods: Clinical and molecular data from 448 patients with brain tumors were obtained from the Cancer Genome Atlas. The DNA methylation stemness index, derived from DNA methylation patterns, was built using a one-class logistic regression. Associations between variables were evaluated using the χ^2 test with k degrees of freedom, followed by analysis of the adjusted standardized residuals (ASRs >1.96 indicate a significant association between variables). MCA was used to uncover associations between glioma prognostic factors and stemness.

Results: Our analysis revealed significant associations among molecular and clinical characteristics in gliomas. Additionally, we demonstrated the capability of MCA to identify associations between stemness and these prognostic factors. Our results exhibited a strong association between higher DNA methylation stemness index and features related to poorer prognosis such as glioblastoma cancer type (ASR: 8.507), grade 4 (ASR: 8.507), isocitrate dehydrogenase wild type (ASR:15.904), unmethylated MGMT (methylguanine methyltransferase) Promoter (ASR: 9.983), and telomerase reverse transcriptase expression (ASR: 3.351), demonstrating the utility of MCA as an analytical tool for elucidating potential prognostic factors.

Conclusions: MCA is a valuable tool for understanding the complex interdependence of prognostic markers in gliomas. MCA facilitates the exploration of large-scale datasets and enhances the identification of significant associations.

KEYWORDS

brain tumors; bioinformatics; stemness; multiple correspondence analysis

Introduction

Cancer is a dynamic and heterogeneous disease characterized by several hallmarks controlling and contributing to its development and progression [1]. Cancer research continually generates large scales of data encompassing clinical information, genomic and transcriptomic profiles, prognostic and diagnostic markers, and therapeutic targets [2]. Different approaches have been used to study and associate all these variables to manage this complexity, aiming to reduce the dimensionality and enhance data interpretation and decision-making process. Several features used to study and classify the different types of cancer are based on categorical variables. For instance, the most widely used cancer staging system, TNM, is based on categorical variables, where "T" refers to the size of the primary tumor, "N" refers to the number of lymph nodes affected by cancer, and "M" refers to absence or presence of metastasis [3]. Thus, these biological and clinical variables interact, and their associations can be measured and diagnosticated using statistical tests such as Fisher exact tests and χ^2 tests. However, these approaches could not provide a global and comprehensive picture of the associations between these variables, particularly in datasets with a large number of categorical variables. Therefore, using multivariate and visual analysis methods can significantly improve the analysis and interpretation of associations between clinical and molecular cancer phenotypes.

Brain tumors are a particularly aggressive type of cancer, mostly due to local tissue damage and highly invasive growth. Gliomas, which originate from neuroglial stem cells or progenitor cells, account for 30% of primary brain tumors and 80% of malignant brain tumors [4]. This heterogeneous disease is histologically classified based on anaplasia criteria and predominant cell types such as oligodendroglioma, astrocytoma, and glioblastoma (GBM) [5]. Nevertheless, as further investigation aimed to elucidate the neuropathological mechanisms of gliomas, new variables are considered for characterizing this cancer tumor, leading to reclassifications based on mutational profiles, clinical data, and epigenetic factors [6]. This scenario resulted in different prognosis predictions, diagnosis determination, and treatment responses, contributing to an increasingly complex and stratified understanding of gliomas.

Stemness is a key phenotype of cancer stem cells (CSCs), related to tumor initiation and progression, therapy resistance, and metastasis [7]. CSCs are referred to as a subpopulation of tumor cells able to self-renew and differentiate into distinct cell lineages, enabling those cells to adapt to different environmental situations [8]. Moreover, recent studies have demonstrated associations between stemness features and different histologic classifications or prognostic factors of gliomas [9-11]. Therefore, providing a comprehensive visualization of the associations between clinical features and stemness in brain tumors could be valuable for identifying and determining potential prognostic and therapeutic markers.

Multiple correspondence analysis (MCA) is an unsupervised data science methodology that aims to observe and represent associations between variables disposed in contingency tables, visualizing these associations in a 2D perceptual map. This approach allows for the simultaneous visualization of the relationship between 2 or more characteristics [12]. MCA shares general characteristics, and it is an extension of principal component analysis which is effective in reducing data dimensionality. Thus, MCA can significantly reduce the workload and simplify statistical analysis in healthy research [13]. The results of MCA are typically interpreted in a 2D map, where the relative positions of categories of each variable and their distribution along the dimensions are analyzed. Categories that cluster together and are closer are more likely to be associated, providing key insights into the relationship [14]. Despite its applicability, rigor, and success in other disciplines such as Geography, Epidemiology, and Human Physiology, MCA remains underused in Oncology research and few studies are applying [12,14-16].

By using MCA, we aimed to gain a deeper understanding of the interdependence between stemness and prognostic factors. Our findings revealed associations among molecular and clinical characteristics and prognostic factors, as previously described by the literature [17]. Additionally, we demonstrated the capability of MCA to identify associations between stemness and these prognostic factors. Our results exhibited a strong association between higher stemness index and features related to poorer prognosis, demonstrating the utility of MCA as an analytical tool for elucidating oncological heterogeneity and may also offer a valuable strategy for therapeutic decision-making. This study highlights MCA as a powerful tool for overcoming the barrier of representing the heterogeneity and complexity of cancer variables, particularly in glioma.

Methods

Dataset of the Tumor Samples

Clinical and molecular information of a total of 448 patients with brain tumors was obtained from the Cancer Genome Atlas (TCGA). We tailored the dataset to contain only qualitative information, with 12 variables: cancer type, histology, grade, patient's vital status, IDH (isocitrate dehydrogenase) status, codeletion of chromosomes 1p and 19q arms, MGMT (methylguanine methyltransferase) gene methylation, telomerase reverse transcriptase (TERT) expression, gain of chromosome 19 and 20, chromosome 7 gain and chromosome 10 loss, ATRX (alpha thalassemia/mental retardation syndrome, X-linked) status, and GBM transcriptome subtypes. All categorical variables were selected based on their established role as prognostic factors for brain tumors.

DNA Methylation Stemness Index

The DNA methylation stemness index (mDNAsi) based on DNA methylation was built using a one-class logistic regression [18] on the pluripotent stem cell samples (embryonic stem cell and induced pluripotent stem cell) from the Progenitor Cell Biology Consortium dataset [19,20]. The algorithm was built and validated as described in the original paper [21]. The mDNAsi was applied in 381 samples from the TCGA database. Malta's model presented a high correlation among other CSC signatures, providing significant insights into the biological and clinical features of pan-cancer. The workflow to generate the mDNAsi is available in the original paper [21].

Multiple Correspondence Analysis

MCAs were conducted in the RStudio (version 4.3.1; Posit, PBC) environment using the packages FactoMineR (version 2.11; Institut Agro) [22] and cabootcrs (version 2.1.0; Cranfield University), for creating matrices for MCAs. Contingency tables for the categorical variables were generated, and associations between variables were assessed using a χ^2 test with k degrees of freedom. This was followed by the analysis of the adjusted standardized residuals (ASRs). The χ^2 test evaluates whether the observed associations between categorical variables are nonrandomly associated (P value <.05). ASRs higher than 1.96 indicate a significant association between variables in the matrix. To perform MCA, the categorical variables should not be randomly associated. To create the perceptual map, inertia was determined as the total χ^2 divided by the number of samples, resulting in the number of associations in the dataset. MCA was performed based on the binary matrices and row and column profiles were determined to demonstrate the influence of each category of variables on the others. Matrices were defined based on the row and column profiles. Eigenvalues were then extracted to represent the number of dimensions that could be captured in the analysis. Finally, the x- and y-axis coordinates of the perceptual map were determined, allowing the category of the variables to be represented and established. In MCA, the spatial distance between categories of different variables reflects their associations. Categories with high coordinates that are close in

space are directly associated, while categories presenting high coordinates but opposing coordinates are inversely associated.

Statistical Analysis

Fisher exact tests and χ^2 tests were performed using RStudio 4.3.1 environment and GraphPad Prism (version 10.3.0; Dotmatics, USA).

Ethical Considerations

The results published in this paper are in whole based upon data generated by the TCGA Research Network [23]. TCGA Ethics and Policies was originally published by the National Cancer Institute [24].

Results

MCA Can Identify Associations Between Different Variables of Gliomas and Patient Vital Status

To determine the suitability of glioma variables for MCA, we first evaluated whether categorical glioma variables were randomly or nonrandomly associated. This involved creating individual contingency tables for each pair of glioma variables (Multimedia Appendices 1-13). Then, we applied χ^2 tests for each contingency table to confirm nonrandom associations (P value <.05). We also confirmed the associations between categorical variables and patients' vital status using the Fisher exact test (P value <.05) (Multimedia Appendix 14). Based on the χ^2 test, the results indicated that only 2 categorical variables, gender and DAXX expression, were randomly associated, suggesting no significant association patterns between these

variables and the others. Consequently, gender and DAXX expression were excluded from further analysis.

In the subsequent analysis, we observed and measured the strength of associations between the patient vital status (0-alive; 1-dead) and different factors including cancer type, histology, grade, IDH status, 1p19q codeletion, MGMT promoter methylation, gain of chromosome (Chr) 7 and loss of Chr10 (7+/10–), co-gain of Chr19 and Chr20 (19+/20+), TERT expression, ATRX status, and transcriptome subtype, aiming to determine whether MCA could identify associations between prognostic factors for this disease. We used ASRs to assess these associations, considering a category of each variable to be associated with either alive or dead vital status when the ASR values were higher than 1.96. Patients' vital status classified as dead were associated with poorer prognostics factors such as GBMs, grade 4, IDH wild type, non-codeleted 1p19q, unmethylated MGMT promoter, gain of Chr7 and loss of Chr10, expression of TERT, ATRX wild type, and classical (CL) and mesenchymal (ME) transcriptome subtypes (Table 1). In contrast, patients classified as alive were linked to favorable prognostic variables, including oligoastrocytomas and oligodendrogliomas, grade 2, IDH mutant, codeleted 1p19q, methylated MGMT promoter, absence of combined Chr7+/Chr10– (chromosome 7 gain and 10 loss), lack of TERT expression, ATRX mutant, and the proneural (PN) and neural (NE) transcriptome subtypes (Table 1). Histological classification, grade, IDH status, and Chr7+/Chr10– were the most strongly associated features with patient vital status. These associations were further illustrated in a heatmap (Figure 1A-D).

Table . Table exhibiting the values of the adjusted standardized residuals. Categories of variables with values higher than 1.96 are considered associated. We could observe a strong association between poorer prognostic factors and dead vital status. In contrast, better prognostic factors were associated with alive vital status.

Glioma variables	Patient vital status		Categories associated with
	Alive	Dead	
Glioblastoma	—[a]	8.127	Dead
Oligoastrocytoma	2.64	—	Alive
Oligodendroglioma	3.309	—	Alive
Astrocytoma	1.756	—	Not associated
Grade 2	6.809	—	Alive
Grade 3	0.155	—	Not associated
Grade 4	—	8.127	Dead
IDH[b] wild type	—	8.804	Dead
IDH mutant	8.804	—	Alive
1p/19q codeletion	5.265	—	Alive
1p/19q non-codeletion	—	5.265	Dead
Methylated MGMT[c] promoter	5.26	—	Alive
Unmethylated MGMT promoter	—	5.26	Dead
No combined Chr7+/Chr10–[d]	5.756	—	Alive
Chr7+/Chr10–	—	5.756	Dead
Not expressed TERT[e]	3.078	—	Alive
Expressed TERT	—	3.078	Dead
ATRX[f] mutant	2.311	—	Alive
ATRX wild type	—	2.311	Dead
Proneural subtype	4.122	—	Alive
Neural subtype	3.593	—	Alive
Mesenchymal subtype	—	4.635	Dead
Classical subtype	—	4.852	Dead

[a]Not applicable.

[b]IDH: isocitrate dehydrogenase.

[c]MGMT: methylguanine methyltransferase.

[d]Chr7+/Chr10–: chromosome 7 gain and 10 loss.

[e]TERT: telomerase reverse transcriptase.

[f]ATRX: alpha thalassemia/mental retardation syndrome, X-linked.

Figure 1. Heatmap exhibiting the values of the adjusted standardized residuals. Categories of variables with values higher than 1.96 are associated. We could observe a strong association of (**A**) glioblastoma (8.127), (**B**) grade 4 (8.127), (**C**) IDH wild type (8.804), and (**D**) Chr7+/Chr10– (5.756) with dead vital status. Favorable prognostic factors including (**A**) oligoastrocytoma and oligodendroglioma, (**B**) grade 2, (**C**) IDH mutant, and (**D**) no combined copy number alterations were associated with alive vital status. Chr7+/Chr10–: chromosome 7 gain and 10 loss; IDH: isocitrate dehydrogenase.
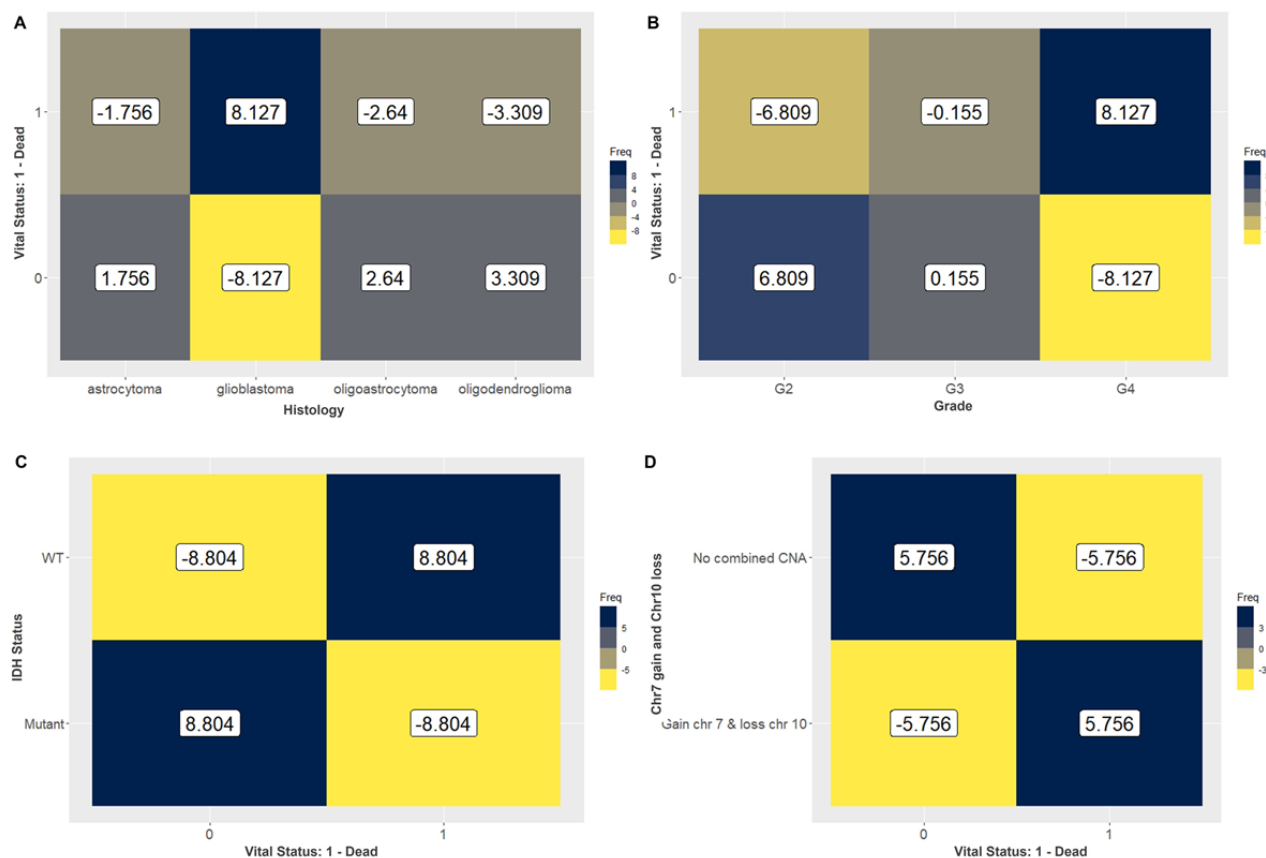
Using MCA, we observed that dimension 1 (x-axis) accounted for 33.71% of the variance, while dimension 2 (y-axis) accounted for 14.08%. The inertia (sum of the variances) for these 2 dimensions was 47.79%. The variance of the overall dimensions (17 dimensions) for the combinations of the variables is illustrated in Multimedia Appendix 15. The main idea was to present the percentage of explained variance for each dimension and not the influence of individual variables. The total inertia (sum of the variances) was 1.41.

The results obtained from the MCA were visualized in a 2D perceptual map (Figure 2), highlighting the associations between the categories of each variable. The coordinates of each category are detailed in Table 2. The perceptual map reveals that categories such as GBM, unmethylated MGMT promoter, IDH wild type, Chr7 gain and Chr10 loss, grade 4, GBM ATRX wild type, TERT expression, non-codel 1p.19q, and CL and ME

transcriptome subtypes are closely associated with dead vital status, appearing along the positive x-axis (dimension 1). Conversely, categories like oligoastrocytomas and oligodendrogliomas, grade 2, IDH mutant, codel 1p19q, methylated MGMT promoter, no combined copy number alterations, no expression of TERT, ATRX mutant, and PN and NE transcriptome subtypes are closely associated with alive vital status, appearing along the negative x-axis (dimension 1) (Figure 2).

These findings highlight the utility and capacity of MCA in reducing data dimensionality and demonstrate that, in gliomas, variables interact cohesively. MCA allows us to further visualize these interactions on a global perceptual map, organizing the characteristics into distinct clusters that correspond to different prognostic profiles.

Figure 2. Multiple correspondence analysis (MCA) 2D perceptual map demonstrating the association between the categories of each categorical variable. Categories that are closely clustered are strongly associated with each other. Categories such as glioblastoma, unmethylated MGMT promoter, IDH wild type, chromosome 7 gain and 10 loss (Chr7+/Chr10–), grade 4, glioblastoma ATRX wild type, TERT expression, non-codel 1p.19q, CL and ME transcriptome subtypes are closely associated with dead vital status (1), appearing along the positive x-axis (dimension 1). ATRX: alpha thalassemia/mental retardation syndrome, X-linked; CL: classical; GBM: glioblastoma; IDH: isocitrate dehydrogenase; ME: mesenchymal; MGMT: methylguanine methyltransferase; NE: neural; PN: proneural; TERT: telomerase reverse transcriptase.
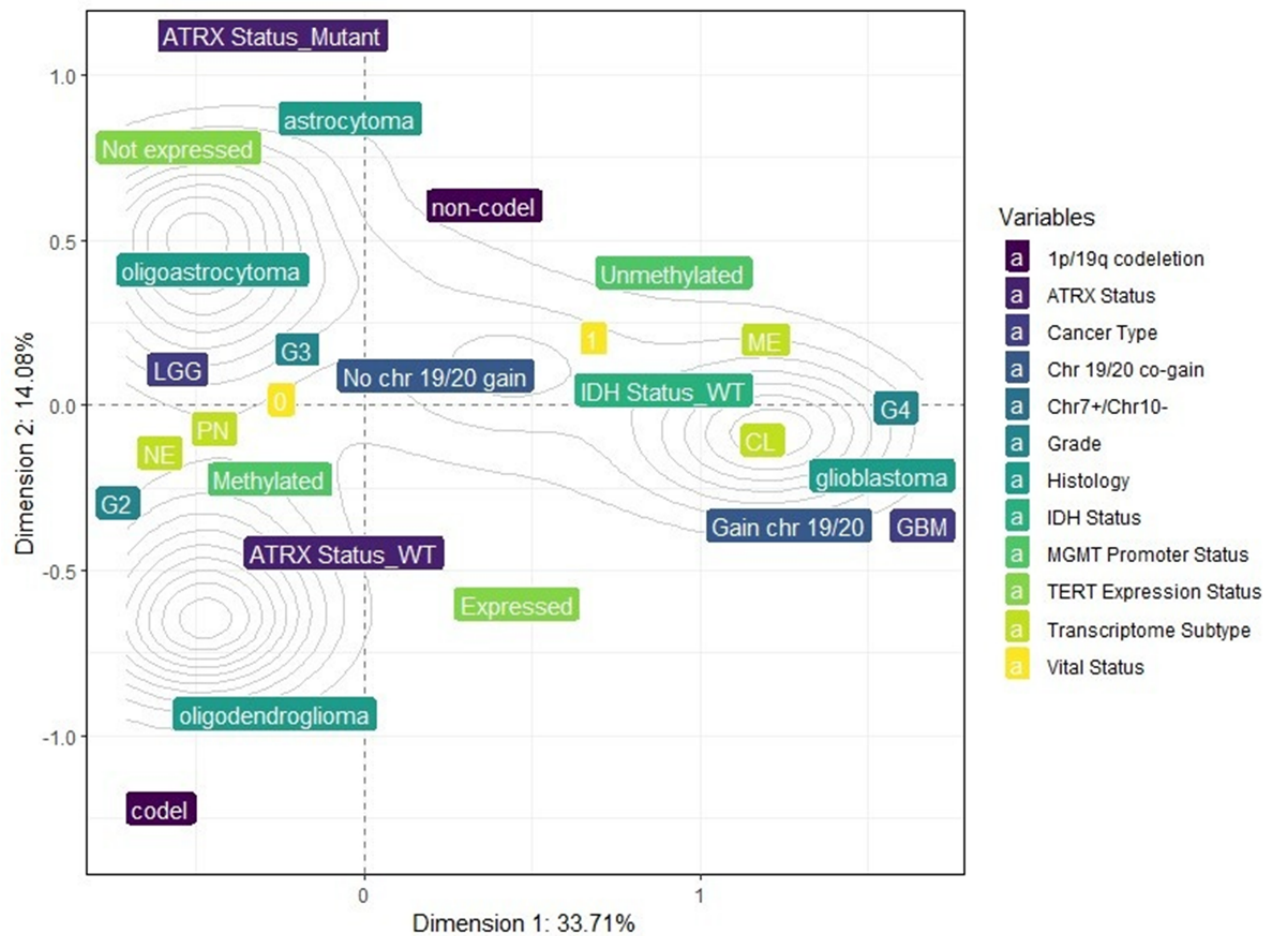
Table . Coordinates of each category compounding the perceptual map.

Category	Dimension 1 (x-axis)	Dimension 2 (y-axis)
GBM[a]	1.6650830	−0.0896760
Low-grade glioma	−0.4723301	0.0254382
Astrocytoma	−0.2672355	0.9527631
Glioblastoma	1.6650830	−0.0896760
Oligoastrocytoma	−0.5334711	0.3276318
Oligodendroglioma	−0.6011671	−0.9346433
Grade 2	−0.6611308	−0.1971919
Grade 3	−0.2970898	0.2320783
Grade 4	1.6650830	−0.0896760
0-Alive	−0.3185609	−0.0551369
1-Dead	0.7544862	0.1305874
IDH[b] mutant	−0.6734117	−0.0548104
IDH wild type	1.1888626	0.0967641
1p/19q codel	−0.6877365	−13.034.766
1p/19q non-codel	0.2750946	0.5213906
Methylated	−0.3429710	−0.1087842
Unmethylated	1.0048449	0.3187185
Chr7+/Chr10–[c]	1.4087248	−0.0210234
No combined Chr7+/Chr10–	−0.4205758	0.0062766
Chr 19/20 co-gain	1.4900007	−0.1295089
No Chr 19/20 co-gain	−0.0843397	0.0073307
Expressed TERT[d]	0.3715020	−0.6845760
Not expressed TERT	−0.4690682	0.8643636
ATRX[e] mutant	−0.6448249	1.0773395
ATRX wild type	0.2693572	−0.4500279
Classical	1.2675815	−0.0217510
Mesenchymal	1.0920361	0.2687642
Neural	−0.5475482	−0.0650952
Proneural	−0.5971662	−0.0604168

[a]GBM: glioblastoma.

[b]IDH: isocitrate dehydrogenase.

[c]Chr7+/Chr10–: chromosome 7 gain and 10 loss.

[d]TERT: telomerase reverse transcriptase.

[e]ATRX: Alpha Thalassemia/Mental Retardation Syndrome X-linked.

MCA Can Associate an Epigenetic Stemness Index (mDNAsi) as a Prognostic Factor in Gliomas

After demonstrating that MCA effectively reduces dimensionality and identifies associations between prognostic factors and clinical data in the glioma database, we proceeded to explore whether MCA could also associate these variables with stemness phenotype. For this analysis, we updated our database by including mDNAsi as a new variable, categorized into low, intermediate, and high levels of stemness. These categories were based on the DNA methylation index related to tumor pathology and clinical outcomes, as previously studied by [21].

First, we evaluated whether the categorical glioma variables were randomly or nonrandomly associated with mDNAsi by creating individual contingency tables for each pair of glioma variables and applying χ^2 tests (Multimedia Appendix 16). We also confirmed the associations between categorical variables using the Fisher exact test (P value <.05) (Multimedia

Appendix 17). All the variables were found to be suitable for MCA. Then, using ASR values to evaluate the strength of these associations, our results indicated strong associations between high mDNAsi levels and poor prognostic and clinical factors. Higher mDNAsi levels were associated with GBM, IDH wild-type, absence of 1p19q co-deletion, unmethylated MGMT promoter, TERT expression, grade 3 and 4, patient's vital status as dead, Chr7+/Chr10–, chromosomes 19/20 co-gain, ATRX

wildtype and ME and CL transcriptome subtypes (Table 3). Conversely, intermediate and lower levels of mDNAsi were associated with characteristics related to favorable prognosis, including oligodendroglioma, IDH mutant, 1p19q co-deletion, methylation of MGMT promoter, absence of TERT expression, grade 2, patient's vital status as alive, no combined copy number alteration, absence of chromosomes 19/20 co-gain, ATRX mutant, and PN and NE transcriptome subtypes (Table 3).

Table . Table exhibiting the values of the adjusted standardized residuals. Categories of variables with values higher than 1.96 are considered associated. We could observe a strong association between poorer prognostic factors and a higher stemness index (DNA methylation stemness index [mDNAsi]). In contrast, better prognostic factors were associated with lower stemness index.

Glioma Variables	mDNAsi			Categories associated with
	Low	Intermediate	High	
Glioblastoma	—[a]	—	8.507	High
Oligoastrocytoma	—	—	—	Not associated
Oligodendroglioma	3.949	—	—	Low
Astrocytoma	—	—	2.832	High
G2	3.279	4.057	—	Low and intermediate
G3	—	—	2.392	High
G4	—	—	8.507	High
IDH[b] wild type	—	—	15.904	High
IDH mutant	8.743	7.057	—	Low and intermediate
1p/19q codeletion	5.772	2.102	—	Low and intermediate
1p/19q non-codeletion	—	—	7.964	High
Methylated MGMT[c] promoter	5.944	3.961	—	Low and intermediate
Unmethylated MGMT promoter	—	—	9.983	High
No combined Chr7+/Chr10–[d]	6.436	5.927	—	Low and intermediate
Chr7+/Chr10–	—	—	12.433	High
Not expressed TERT[e]	—	3.216	—	Intermediate
Expressed TERT	—	—	3.351	High
ATRX[f] mutant	—	3.505	—	Intermediate
ATRX wild type	—	—	4.949	High
Proneural subtype	8.476	—	—	Low
Neural subtype	—	4.218	—	Intermediate
Mesenchymal subtype	—	—	4.771	High
Classical subtype	—	—	10.981	High

[a]Not applicable.

[b]IDH: isocitrate dehydrogenase.

[c]MGMT: methylguanine methyltransferase.

[d]Chr7+/Chr10–: chromosome 7 gain and 10 loss.

[e]TERT: telomerase reverse transcriptase.

[f]ATRX: Alpha Thalassemia/Mental Retardation Syndrome X-linked.

Using MCA, dimension 1 (x-axis) accounted for 28.7% of the variance, while dimension 2 (y-axis) accounted for 14.39%.

The inertia (sum of the variances) for these 2 dimensions was 43.09%. The variance of the overall dimensions (18 dimensions)

for the combinations of the variables is illustrated in Multimedia Appendix 18. The total inertia (sum of the variances) was 1.5. The 2D perceptual map exhibited the associations between the categories of each variable (Figure 3). The perceptual map reveals categories such as GBM, unmethylated MGMT promoter, IDH wild type, Chr7 gain and Chr10 loss, grade 4, GBM ATRX wild type, TERT expression, non-codel 1p.19q, and CL and ME transcriptome subtypes are closely associated

with high mDNAsi, appearing along the positive x-axis (dimension 1). Conversely, categories like oligoastrocytomas and oligodendrogliomas, grade 2, IDH mutant, codel 1p19q, methylated MGMT promoter, no combined copy number alterations, no expression of TERT, ATRX mutant, and PN and NE transcriptome subtypes are closely associated with alive vital status, appearing along the negative x-axis (dimension 1) (Figure 3).

Figure 3. Multiple correspondence analysis (MCA) 2D perceptual map demonstrating the association between the categories of each categorical variable. Categories that are closely clustered are strongly associated with each other. Categories such as glioblastoma, unmethylated MGMT promoter, IDH wild type, chromosome 7 gain and 10 loss (Chr7+/Chr10−), grade 4, glioblastoma ATRX wild type, TERT expression, non-codel 1p.19q, and CL and ME transcriptome subtypes are closely associated with high mDNAsi, appearing along the positive x-axis (dimension 1). ATRX: alpha thalassemia/mental retardation syndrome, X-linked; CL: classical; IDH: isocitrate dehydrogenase; mDNAsi: DNA methylation stemness index; ME: mesenchymal; MGMT: methylguanine methyltransferase; NE: neural; PN: proneural; TERT: telomerase reverse transcriptase.
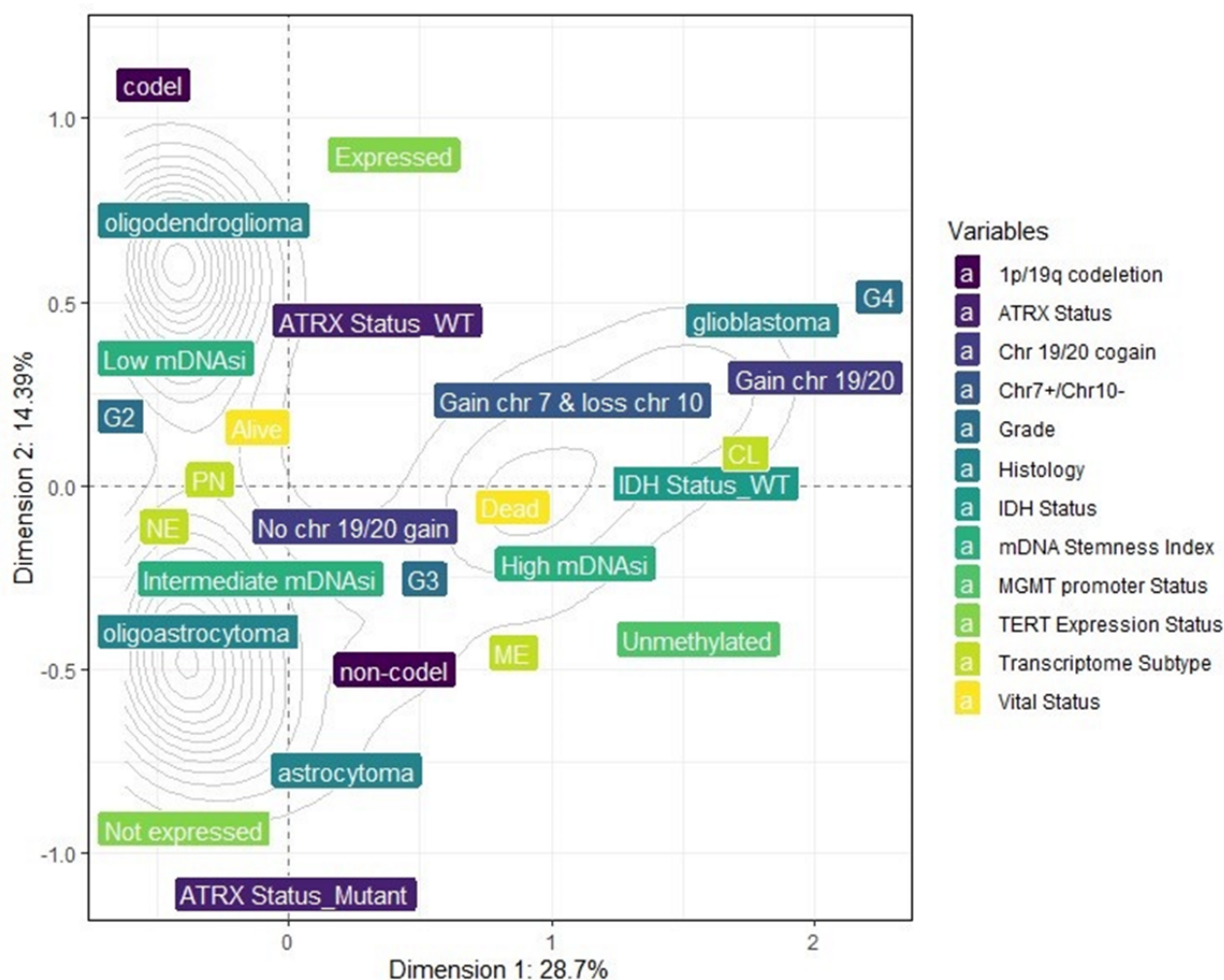
Discussion

Principal Findings

Multiple efforts have been made to explore the diversity of oncologic diseases, with significant contributions from genetics, cell and tissue biology, as well as computational and experimental technologies, providing a wealth of information on cancer manifestations. In the field of glioma research, emerging approaches have sought to clarify tumor pathology and grading through the introduction of novel types and subtypes, as well as by identifying molecular markers and genetic mutations that contribute to predicting diagnosis and

prognosis. However, it also results in an accumulation of extensive datasets, presenting challenges in interpretation and visualization regarding the associations between prognostic factors. In this study, we used MCA, an unsupervised data science approach, to establish statistical associations between different qualitative variables of gliomas. This method was able to reduce data dimensionality and represent it on a 2D perceptual map, revealing associations between various established glioma prognostic factors, including histological classification, IDH status, MGMT promoter methylation, and transcriptome subtypes. Furthermore, we associated these clinical and prognostic variables with an epigenetic-based stemness index

(mDNAsi), demonstrating that higher stemness levels were associated with poorer prognostic factors, providing a useful tool to associate prognostic markers in brain tumors.

Comparison to Prior Studies

Several clinical and molecular factors are considered in predicting the prognosis and survival of brain tumors, more specifically for gliomas. Beyond histological classification and tumor grade, genetic and molecular biomarkers have been incorporated as potential prognostic indicators. Thus, we first evaluated the ability of MCA to associate these consolidated prognostic variables with the patient's vital status. Our findings demonstrate that MCA effectively clusters poor prognostic factors with dead vital status. All these prognostic factors are well consolidated and associated with malignancy of gliomas. IDH mutation represents one of the main prognostic markers for gliomas [25]. It has been identified that one of the mechanisms given by this favorable outcome is the impaired production of nicotinamide adenine dinucleotide phosphate in Krebs cycle caused by IDH1 enzyme mutation that can sensitize tumor cells to chemotherapy and explain the favorable prognosis of patients with IDH mutation [25]. Likewise, co-deletion of 1p19q chromosome arms, especially when combined with other biomarkers such as IDH mutation and TERT expression, has been used as a predictive biomarker and recent studies investigated biological mechanisms to be significantly linked to genes involved in cell division, angiogenesis, and DNA repair responses [26]. Thus, we demonstrated that MCA was able to capture and associate key glioma hallmarks with patients' vital status, which was applied to different clinical variables.

Subsequently, we applied MCA to explore the association between high stemness levels (mDNAsi) and characteristics related to poor prognosis. Stemness has been considered an important phenotype in glioma malignancy and is potentially associated with CL genetic alterations, such as the gain of chromosome 7. Chromosome 7 harbors some key genes related to stemness, including Epidermal Growth Factor Receptor (EGFR), Mesenchymal-Epithelial Transition Factor (MET), and Homeobox A gene (HOXA). A study of 86 GBMs reported that EGFR amplification occurs with higher probability in samples that have a gain of chromosome 7 (82.1%) compared with those without it (66.7%) [27]. In addition, EGFR amplification is more prevalent in IDH-wildtype diffuse gliomas (66.0%) and GBM (85.5%) [28], which are also associated with poorer prognostic factors, consistent with our findings. High mDNAsi has been previously linked to EGFR mutations [21]. The HOXA and MET loci, also located on chromosome 7, have been implicated in stemness-related pathways. Notably, studies have demonstrated interactions between chromosome 7 gain and the expression of a stem cell-related HOX signature in GBMs [29]. Analysis of the MET gene at 7q31.2 revealed that gain occurs in 47% of primary and 44% of secondary GBMs, suggesting that this genetic alteration contributes to the pathogenesis of both GBM subtypes [30].

Overall, relatively few studies have used MCA to explore associations with cancer phenotypes. Previous studies have applied MCA to different approaches, such as analyzing prognosis low rectal cancer surgery [31], investigating the association between some types of cancer in rural or urban areas [15], examining the association between Traditional Chinese Medicine Syndrome and histopathology of colorectal cancer [32], assessing clinically relevant demographic variables across multiple gastrointestinal cancers [33], and the relationship between types of diagnostic classification in breast cancer [34]. Our study also highlights the utility of MCA in investigating associations within the context of brain tumors. MCA enables the investigation of the pattern among many categorical factors in gliomas, providing a powerful computational approach to identify and test prognostic variables. It was possible to visually and quantitatively represent the associations, which facilitates the identification of distinct patient clusters based on shared prognostic characteristics. Our findings were consistent with previous literature and emphasized stemness as an important phenotype for gliomas.

Limitations

Our study has inherent limitations. First, as a retrospective analysis of TCGA data, it is subject to selection bias. Second, we associated all the prognostic variables with patients' vital status, which may not be the most optimal variable for determining prognosis. For the future, we intend to improve our model validating its applicability in other prospective datasets. Third, the absence of therapy data is another limitation of this study. Finally, an intrinsic limitation of MCA is that retaining only 2 or 3 dimensions may not sufficiently capture all the significant features in the data. In our analysis, the percentage of explained inertia was approximately 40%. While there is not an accepted threshold for adequately explained inertia, common guidelines recommend retaining dimensions that represent over 70% of the inertia [35]. However, explained inertia in the range of 40%‐60% is often considered informative, and the interpretability and relevance of the patterns revealed by the dimensions are frequently more important than the exact percentage of inertia explained, especially in a complex heterogeneous disease such as brain tumors [36].

Conclusion and Future Perspectives

In conclusion, our findings suggest that MCA is a valuable tool for understanding the interdependence between prognostic markers in gliomas. MCA facilitates the exploration of a large-scale dataset and enhances the identification of associations. Considering the advances in computational oncology and the emergence of new oncological features, such as stemness phenotype, incorporating MCA into cancer research as an approach to exploring the complex heterogeneity of the oncologic field becomes a powerful tool for simplifying data management. It contributes to researchers statistically identifying associations between variables within extensive databases and improves the visual representation, leading to a deeper understanding of cancer findings.

Acknowledgments

Data Availability

The datasets generated or analyzed during this study are available at National Institutes of Health Genomic Data Commons (GDC) [37]. The workflow to generate the DNA methylation stemness index (mDNAsi) can be accessed at GitHub [38].

Authors' Contributions

MEGJ conducted the study, contributing to the acquisition of data, data analysis and interpretation, production of tables and figures, and wrote the first version of the manuscript. HF contributed to the interpretation and discussion of data and corrected the final version of the manuscript. TMM contributed to the acquisition, interpretation, and discussion of data, and corrected the final version of the manuscript. PLPX contributed to the concept and design of the study, data analysis and interpretation, funding, and corrected the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Individual contingency tables for cancer type.
[XLSX File, 29 KB - bioinform_v6i1e65645_app1.xlsx]

Multimedia Appendix 2
Individual contingency tables for histology.
[XLSX File, 32 KB - bioinform_v6i1e65645_app2.xlsx]

Multimedia Appendix 3
Individual contingency tables for grade.
[XLSX File, 27 KB - bioinform_v6i1e65645_app3.xlsx]

Multimedia Appendix 4
Individual contingency tables for gender.
[XLSX File, 24 KB - bioinform_v6i1e65645_app4.xlsx]

Multimedia Appendix 5
Individual contingency tables for vital status.
[XLSX File, 23 KB - bioinform_v6i1e65645_app5.xlsx]

Multimedia Appendix 6
Individual contingency tables for IDH (isocitrate dehydrogenase) status.
[XLSX File, 21 KB - bioinform_v6i1e65645_app6.xlsx]

Multimedia Appendix 7
Individual contingency tables for X1p.19q.codeletion.
[XLSX File, 20 KB - bioinform_v6i1e65645_app7.xlsx]

Multimedia Appendix 8
Individual contingency tables for MGMT (methylguanine methyltransferase) promoter.
[XLSX File, 18 KB - bioinform_v6i1e65645_app8.xlsx]

Multimedia Appendix 9
Individual contingency tables for Chr 7 gain and Chr 10 loss.
[XLSX File, 17 KB - bioinform_v6i1e65645_app9.xlsx]

Multimedia Appendix 10
Individual contingency tables for Chr 19/20 co-gain.
[XLSX File, 16 KB - bioinform_v6i1e65645_app10.xlsx]

Multimedia Appendix 11
Individual contingency tables for TERT (telomerase reverse transcriptase) expression status.
[XLSX File, 13 KB - bioinform_v6i1e65645_app11.xlsx]

Multimedia Appendix 12
Individual contingency tables for ATRX (Alpha Thalassemia/Mental Retardation Syndrome X-linkedalpha thalassemia/mental retardation syndrome, X-linked) status.
[XLSX File, 11 KB - bioinform_v6i1e65645_app12.xlsx]

Multimedia Appendix 13
Individual contingency tables for DAXX status.
[XLSX File, 10 KB - bioinform_v6i1e65645_app13.xlsx]

Multimedia Appendix 14
Fisher exact test and χ^2 test for vital status $\times$ glioma prognostic factors.
[XLSX File, 24 KB - bioinform_v6i1e65645_app14.xlsx]

Multimedia Appendix 15
Percentage of explained variances of the overall (17) dimensions.
[PNG File, 161 KB - bioinform_v6i1e65645_app15.png]

Multimedia Appendix 16
Individual contingency table for mDNAsi.
[XLSX File, 30 KB - bioinform_v6i1e65645_app16.xlsx]

Multimedia Appendix 17
Fisher exact test and χ^2 test for mDNAsi (DNA methylation stemness index) $\times$ glioma prognostic factors.
[XLSX File, 22 KB - bioinform_v6i1e65645_app17.xlsx]

Multimedia Appendix 18
Percentage of explained variances of the overall (18) dimensions.
[PNG File, 9 KB - bioinform_v6i1e65645_app18.png]

References

1. Hanahan D. Hallmarks of cancer: new dimensions. Cancer Discov 2022 Jan;12(1):31-46. [doi: 10.1158/2159-8290.CD-21-1059] [Medline: 35022204]
2. Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. Nat Rev Clin Oncol 2018 Feb;15(2):81-94. [doi: 10.1038/nrclinonc.2017.166] [Medline: 29115304]
3. Brierley J, O'Sullivan B, Asamura H, et al. Global consultation on cancer staging: promoting consistent understanding and use. Nat Rev Clin Oncol 2019 Dec;16(12):763-771. [doi: 10.1038/s41571-019-0253-x] [Medline: 31388125]
4. Weller M, Wick W, Aldape K, et al. Glioma. Nat Rev Dis Primers 2015 Jul 16;1:15017. [doi: 10.1038/nrdp.2015.17] [Medline: 27188790]
5. Louis DN, Ohgaki H, Wiestler OD, et al. The 2007 WHO classification of tumours of the central nervous system. Acta Neuropathol 2007 Aug;114(2):97-109. [doi: 10.1007/s00401-007-0243-4] [Medline: 17618441]
6. Louis DN, Perry A, Wesseling P, et al. The 2021 WHO classification of tumors of the central nervous system: a summary. Neuro Oncol 2021 Aug 2;23(8):1231-1251. [doi: 10.1093/neuonc/noab106] [Medline: 34185076]
7. Ayob AZ, Ramasamy TS. Cancer stem cells as key drivers of tumour progression. J Biomed Sci 2018 Mar 6;25(1):20. [doi: 10.1186/s12929-018-0426-4] [Medline: 29506506]
8. Batlle E, Clevers H. Cancer stem cells revisited. Nat Med 2017 Oct 6;23(10):1124-1134. [doi: 10.1038/nm.4409] [Medline: 28985214]
9. Wang Q, Hu B, Hu X, et al. Tumor evolution of glioma-intrinsic gene expression subtypes associates with immunological changes in the microenvironment. Cancer Cell 2017 Jul 10;32(1):42-56. [doi: 10.1016/j.ccell.2017.06.003] [Medline: 28697342]

10. Ortensi B, Setti M, Osti D, Pelicci G. Cancer stem cell contribution to glioblastoma invasiveness. Stem Cell Res Ther 2013 Feb 28;4(1):18. [doi: 10.1186/scrt166] [Medline: 23510696]

11. Tan J, Zhu H, Tang G, et al. Molecular subtypes based on the stemness index predict prognosis in glioma patients. Front Genet 2021;12:616507. [doi: 10.3389/fgene.2021.616507] [Medline: 33732284]

12. Sourial N, Wolfson C, Zhu B, et al. Correspondence analysis is a useful tool to uncover the relationships among categorical variables. J Clin Epidemiol 2010 Jun;63(6):638-646. [doi: 10.1016/j.jclinepi.2009.08.008] [Medline: 19896800]

13. Li BH, Sun ZQ, Dong SF. Correspondence analysis and its application in oncology. Commun Stat Theory Methods 2010 Mar 19;39(7):1229-1236. [doi: 10.1080/03610920902871446]

14. Costa PS, Santos NC, Cunha P, Cotter J, Sousa N. The use of multiple correspondence analysis to explore associations between categories of qualitative variables in healthy ageing. J Aging Res 2013;2013:302163. [doi: 10.1155/2013/302163] [Medline: 24222852]

15. Florensa D, Godoy P, Mateo J, et al. The use of multiple correspondence analysis to explore associations between categories of qualitative variables and cancer incidence. IEEE J Biomed Health Inform 2021 Sep;25(9):3659-3667. [doi: 10.1109/JBHI.2021.3073605] [Medline: 33857006]

16. van Horn A, Weitz CA, Olszowy KM, et al. Using multiple correspondence analysis to identify behaviour patterns associated with overweight and obesity in Vanuatu adults. Public Health Nutr 2019 Jun;22(9):1533-1544. [doi: 10.1017/S1368980019000302] [Medline: 30846019]

17. Śledzińska P, Bebyn MG, Furtak J, Kowalewski J, Lewandowska MA. Prognostic and predictive biomarkers in gliomas. Int J Mol Sci 2021 Sep 26;22(19):10373. [doi: 10.3390/ijms221910373] [Medline: 34638714]

18. Sokolov A, Paull EO, Stuart JM. ONE-class detection of cell states in tumor subtypes. Presented at: Proceedings of the Pacific Symposium; Jan 4-8, 2016; Kohala Coast, Hawaii, USA. [doi: 10.1142/9789814749411_0037]

19. Salomonis N, Dexheimer PJ, Omberg L, et al. Integrated genomic analysis of diverse induced pluripotent stem cells from the progenitor cell biology consortium. Stem Cell Rep 2016 Jul 12;7(1):110-125. [doi: 10.1016/j.stemcr.2016.05.006] [Medline: 27293150]

20. Daily K, Ho Sui SJ, Schriml LM, et al. Molecular, phenotypic, and sample-associated data to describe pluripotent stem cell lines and derivatives. Sci Data 2017 Mar 28;4:170030. [doi: 10.1038/sdata.2017.30] [Medline: 28350385]

21. Malta TM, Sokolov A, Gentles AJ, et al. Machine learning identifies stemness features associated with oncogenic dedifferentiation. Cell 2018 Apr 5;173(2):338-354. [doi: 10.1016/j.cell.2018.03.034] [Medline: 29625051]

22. Lê S, Josse J, Husson F. FactoMineR: an R package for multivariate analysis. J Stat Softw 2008 Mar;25(1):1-18. [doi: 10.18637/jss.v025.i01] [Medline: 27348562]

23. The Cancer Genome Atlas program (TCGA). Center for Cancer Genomics. URL: https://www.cancer.gov/tcga [accessed 2025-03-06]

24. The Cancer Genome Atlas program. National Cancer Institute. URL: https://www.cancer.gov/ccg/research/structural-genomics/tcga/history/policies/tcga-human-subjects-data-policies.pdf [accessed 2025-03-06]

25. Bleeker FE, Atai NA, Lamba S, et al. The prognostic IDH1(R132) mutation is associated with reduced NADP+-dependent IDH activity in glioblastoma. Acta Neuropathol 2010 Apr;119(4):487-494. [doi: 10.1007/s00401-010-0645-6] [Medline: 20127344]

26. Chai RC, Zhang KN, Chang YZ, et al. Systematically characterize the clinical and biological significances of 1p19q genes in 1p/19q non-codeletion glioma. Carcinogenesis 2019 Oct 16;40(10):1229-1239. [doi: 10.1093/carcin/bgz102] [Medline: 31157866]

27. McNulty SN, Cottrell CE, Vigh-Conrad KA, et al. Beyond sequence variation: assessment of copy number variation in adult glioblastoma through targeted tumor somatic profiling. Hum Pathol 2019 Apr;86:170-181. [doi: 10.1016/j.humpath.2018.12.004] [Medline: 30594748]

28. Wang H, Zhang X, Liu J, et al. Clinical roles of EGFR amplification in diffuse gliomas: a real-world study using the 2021 WHO classification of CNS tumors. Front Neurosci 2024;18:1308627. [doi: 10.3389/fnins.2024.1308627] [Medline: 38595969]

29. Kurscheid S, Bady P, Sciuscio D, et al. Chromosome 7 gain and DNA hypermethylation at the HOXA10 locus are associated with expression of a stem cell related HOX-signature in glioblastoma. Genome Biol 2015 Jan 27;16(1):16. [doi: 10.1186/s13059-015-0583-7] [Medline: 25622821]

30. Pierscianek D, Kim YH, Motomura K, et al. MET gain in diffuse astrocytomas is associated with poorer outcome. Brain Pathol 2013 Jan;23(1):13-18. [doi: 10.1111/j.1750-3639.2012.00609.x] [Medline: 22672415]

31. Mancini R, Pattaro G, Diodoro MG, et al. Tumor regression grade after neoadjuvant chemoradiation and surgery for low rectal cancer evaluated by multiple correspondence analysis: ten years as minimum follow-up. Clin Colorectal Cancer 2018 Mar;17(1):e13-e19. [doi: 10.1016/j.clcc.2017.06.004] [Medline: 28865674]

32. Wu T, Zhang S, Guo S, et al. Correspondence analysis between traditional Chinese medicine (TCM) syndrome differentiation and histopathology in colorectal cancer. Eur J Integr Med 2015 Aug;7(4):342-347. [doi: 10.1016/j.eujim.2015.07.003]

33. Kramer RJ, Rhodin KE, Therien A, et al. Unsupervised clustering using multiple correspondence analysis reveals clinically-relevant demographic variables across multiple gastrointestinal cancers. Surgical Oncology Insight 2024 Mar;1(1):100009. [doi: 10.1016/j.soi.2024.100009]

34.   Nadjib Bustan M, Arif Tiro M, Annas S. Correspondence analysis of breast cancer diagnosis classification. J Phys Conf Ser 2019 Jun 1;1244(1):012030. [doi: 10.1088/1742-6596/1244/1/012030]

35.   Higgs NT. Practical and innovative uses of correspondence analysis. R Stat Soc Ser D (The Statistician) 1991;40(2):183. [doi: 10.2307/2348490]

36.   Husson F, Leˆ S, Pagès J. Exploratory Multivariate Analysis by Example Using R: CRC Press; 2011.

37.   Machine learning identifies stemness features associated with oncogenic dedifferentiation. National Cancer Institute. URL: https://gdc.cancer.gov/about-data/publications/PanCanStemness-2018 [accessed 2025-03-06]

38.   PanCanStem: reproducing mrnasi from PMID: 29625051. GitHub. URL: https://github.com/ArtemSokolov/PanCanStem [accessed 2025-03-06]

Abbreviations

ASR: adjusted standardized residual
ATRX: alpha thalassemia/mental retardation syndrome, X-linked
Chr: chromosome
Chr7+/Chr10–: chromosome 7 gain and 10 loss
CL: classical
CSC: cancer stem cell
GMB: glioblastoma
IDH: isocitrate dehydrogenase
MCA: multiple correspondence analysis
mDNAsi: DNA methylation stemness index
ME: mesenchymal
MGMT: methylguanine methyltransferase
NE: neural
PR: proneural
TCGA: the Cancer Genome Atlas
TERT: telomerase reverse transcriptase

XSL•FO
RenderX

<u>Original Paper</u>

Decentralized Biobanking Apps for Patient Tracking of Biospecimen Research: Real-World Usability and Feasibility Study

William Sanchez[1], BA; Ananya Dewan[2], BA; Eve Budd[3]; M Eifler[1], BA, MFA; Robert C Miller[4,5], MD; Jeffery Kahn[6], PhD; Mario Macis[7], PhD; Marielle Gross[1,6], MD

[1]de-bi, co., Greencastle, PA, United States

[2]Johns Hopkins School of Medicine, Johns Hopkins University, Baltimore, MD, United States

[3]Harpur College of Arts and Sciences, State University of New York, Binghamton, NY, United States

[4]Faculty of Medicine, Mayo Clinic, Rochester, MN, United States

[5]School of Medicine, Indiana University Hospital, Indianapolis, IN, United States

[6]Johns Hopkins Berman Institute of Bioethics, Johns Hopkins University, Baltimore, MD, United States

[7]Carey School of Business, Johns Hopkins University, Baltimore, MD, United States

Corresponding Author:
Marielle Gross, MD
Johns Hopkins Berman Institute of Bioethics
Johns Hopkins University
1809 Ashland Ave.
Baltimore, PA, 17225
United States
Phone: 1 8135416103
Email: mariellesophiagross@gmail.com

Abstract

Background: Biobank privacy policies strip patient identifiers from donated specimens, undermining transparency, utility, and value for patients, scientists, and society. We are advancing decentralized biobanking apps that reconnect patients with biospecimens and facilitate engagement through a privacy-preserving nonfungible token (NFT) digital twin framework. The decentralized biobanking platform was first piloted for breast cancer biobank members.

Objective: This study aimed to demonstrate the technical feasibility of (1) patient-friendly biobanking apps, (2) integration with institutional biobanks, and (3) establishing the foundation of an NFT digital twin framework for decentralized biobanking.

Methods: We designed, developed, and deployed a decentralized biobanking mobile app for a feasibility pilot from 2021 to 2023 in the setting of a breast cancer biobank at a National Cancer Institute comprehensive cancer center. The Flutter app was integrated with the biobank's laboratory information management systems via an institutional review board–approved mechanism leveraging authorized, secure devices and anonymous ID codes and complemented with a nontransferable ERC-721 NFT representing the *soul-bound* connection between an individual and their specimens. Biowallet NFTs were held within a custodial wallet, whereas the user experiences simulated token-gated access to personalized feedback about collection and use of individual and collective deidentified specimens. Quantified app user journeys and NFT deployment data demonstrate technical feasibility complemented with design workshop feedback.

Results: The decentralized biobanking app incorporated key features: "biobank" (learn about biobanking), "biowallet" (track personal biospecimens), "labs" (follow research), and "profile" (share data and preferences). In total, 405 pilot participants downloaded the app, including 361 (89.1%) biobank members. A total of 4 central user journeys were captured. First, all app users were oriented to the ≥60,000-biospecimen collection, and 37.8% (153/405) completed research profiles, collectively enhancing annotations for 760 unused specimens. NFTs were minted for 94.6% (140/148) of app users with specimens at an average cost of US $4.51 (SD US $2.54; range US $1.84-$11.23) per token, projected to US $17,769.40 (SD US $159.52; range US $7265.62-$44,229.27) for the biobank population. In total, 89.3% (125/140) of the users successfully claimed NFTs during the pilot, thereby tracking 1812 personal specimens, including 202 (11.2%) distributed under 42 unique research protocols. Participants embraced the opportunity for direct feedback, community engagement, and potential health benefits, although user onboarding requires further refinement.

Conclusions: Decentralized biobanking apps demonstrate technical feasibility for empowering patients to track donated biospecimens via integration with institutional biobank infrastructure. Our pilot reveals potential to accelerate biomedical research through patient engagement; however, further development is needed to optimize the accessibility, efficiency, and scalability of platform design and blockchain elements, as well as a robust incentive and governance structure for decentralized biobanking.

Introduction

Background

University biobanks collect, store, and distribute biospecimens such as tissue and blood, capitalizing on leftover clinical materials from affiliated hospitals to drive biomedical science and drug discovery [1-3]. Standard operating procedure for most biobanks in academic medical centers includes prospective broad consent for nonspecific, future research [4] coupled with deidentification, whereby identifiers are stripped before specimen allocation [5]. In this setting, patients do not learn what becomes of their donations, and scientists lack access to the donor, linked specimens, and evolving clinical data [4,6]. This disconnect, though the by-product of policies designed to protect privacy while promoting learning, promulgates a biobank ecosystem that permits problematic gaps in recognition, reciprocity, and return of results [7,8]. Simultaneously, vast yet siloed specimen collections have accumulated across most US academic medical centers, a widely underused and unsustainable "treasure trove" wherein frozen assets lay hidden from patients and scientists for whom they may be most valuable [3,9]. The lack of an efficient market for ensuring the use of donated materials deepens the crisis of faith in public health institutions and has prompted attempts at marketplace solutions [10,11].

We are advancing *decentralized biobanking* as a software platform predicated on blockchain technology's democratic ethos, incentive alignment, transparency, and assurances of trust [12]. These key features are reflective of blockchains as permissionless, distributed, shared ledgers of digital transactions engineered to be mathematically concordant, accessible, and auditable [13], underscoring their first and most successful use to date for the creation of global digital currency such as Bitcoin, which makes them fit for purpose in efforts to decentralize ownership and governance of data through thoughtfully structured peer-to-peer networks [14]. One of the most promising innovations enabled by blockchains are nonfungible tokens (NFTs), digital record identifiers that serve as electronic deeds for provably unique digital or physical assets that may be represented "on-chain" [15]. The potential for blockchain and NFTs to play a role in restructuring control and ownership of data has been widely discussed, with several notable projects in the health care domain [16,17]. Although empowering patient ownership of health data is compelling in theory, full realization of such initiatives has been elusive in light of complex regulatory considerations, socioeconomic factors, and technical limitations for blockchain technologies and legacy systems [18,19].

Building on the success and diversity of blockchain applications for decentralized finance [20,21], decentralized biobanking applies human-centered design and innovative system mechanisms to empower patients to track donated biospecimens and engage in downstream research activities, outcomes, and products via a platform compatible with established privacy policies and workflows. Our approach provides patients with secure, direct access to personal specimen data housed in institutional databases via user-friendly mobile and web apps complemented with a privacy-preserving NFT digital twin framework [22]. This strategy may support stepwise adoption of increasingly autonomous and progressively decentralized collaborations among patients, scientists, and physicians in a dynamic biomedical metaverse, or "biomediverse."

Objectives

Successful implementation of decentralized biobanking will usher in a new standard for research transparency, foster institutional accountability to the patients and communities they serve, and create opportunities to unite siloed datasets, facilitate timely translation of precision medicine and enable structurally just marketplace solutions for improving efficiency and effectiveness in the management of one of our most precious human resources. In this paper, we explore the technical feasibility of decentralized biobanking through a description and quantitative analysis of a live pilot for a breast cancer biobank at a US academic medical center. We discuss system design, key features, and NFT functionality, illustrating how the platform provided transparency and recognition of patients' contributions to a real-world biobank.

Methods

Decentralized Biobanking System Design: NFT Digital Twin Framework

Decentralized biobanking builds digital bridges among patients, specimens, and scientists, connecting stakeholders based on real-world relationships predicated upon transactions within existing biobank infrastructure and research protocols (Figure 1). The system design represents all people, protocols, and assets in an NFT digital twin framework, creating a blockchain-backed overlay network on top of the established biospecimen ecosystem. Our approach presents a unique strategy for the progressive inclusion of patients, allowing for the implementation of a composable software platform with programmable, modular elements, mechanisms, and workflows that may be integrated with institutional biobank databases to provide durable transparency without requiring substantial time,

labor, or ongoing participation of physicians, biobankers, and scientists. This framework applies privacy by design throughout the engineering process, implementing techniques such as data minimization and innovative system architectures to ensure compliance with established biospecimen collection and research protocols, institutional policies, and data structures. The core benefits of our approach are use case agnostic and can be applied for all biobanks, research protocols, and institutions with minor modifications at each new site.

Figure 1. Decentralized biobanking system design—nonfungible token (NFT) framework and software applications uniting patients, specimens, and scientists. This system diagram illustrates key entities of biobanking connected via a specimen supply chain (red arrow) yet presently lacking a unified platform for collaboration. The proposed decentralized biobanking NFT digital twin framework is designed to integrate with this established infrastructure, mapping the stakeholders, specimens, and studies in the biobanking ecosystem and enabling applications whereby they may be united for mutually beneficial collaboration, data exchange, and value-building activities.
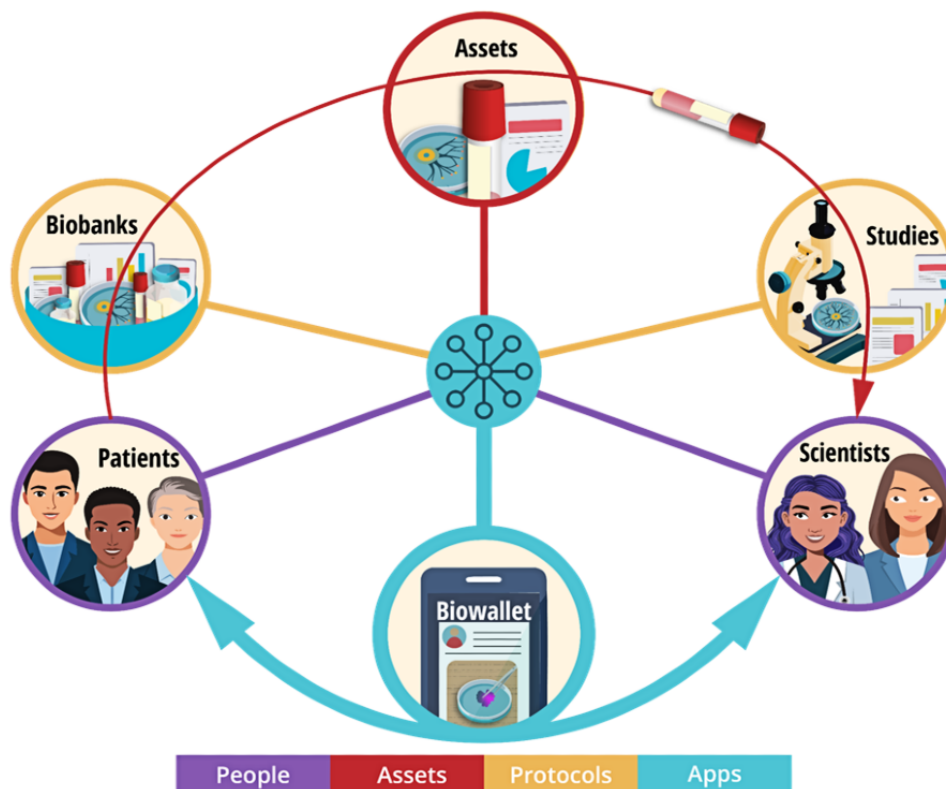
Pilot Setting

The Breast Disease Research Repository (BDRR; STUDY19060196) is a large breast cancer biobank platform at the intersection of the University of Pittsburgh, the University of Pittsburgh Medical Center, and Hillman Cancer Center that served as the pilot study use case. Broad prospective consent for the BDRR is embedded in the breast cancer service line, for example, concurrently with surgical consent. Once consented, "leftovers" from any clinical procedures may be collected by the biobank without further notice or engagement. From 2006 to 2023, more than 10,000 patients consented for the BDRR and specimens were collected from 4000 participants to date. In total, approximately 61,000 specimens were collected, and 6000 were distributed for research, with a mix of fresh and frozen distributions. The biobank operates via a hub-and-spoke model, allocating specimens chiefly to local investigators under designated research or subbiobanking protocols (eg, a flagship patient-derived organoid biobank that grows and distributes copies of living 3D cell cultures [approximately n=300]).

Requirement Gathering

Foundational surveys, semistructured interviews, community engagement, and stakeholder alignment activities with populations with breast cancer, physicians, advocates, and scientists informed our approach to designing a biobanking app for patients [23]. Broadly, we found that patients have an unmet demand for feedback about research on their specimens, with particular interest surrounding personal meaning or potential health benefits for the individual or their family members. For example, a survey respondent noted the following:

> *Giving patients access to this type of information could decrease the lethal lag between research findings and actual clinical practice.*

One patient advocacy leader captured this sentiment, noting the following:

> *We have been screaming for this, banging on pots and pans. Thank you for taking this on.*

Importantly, she alluded to the multifactorial challenge of enabling patients to track and learn about donated biospecimens [23], which would require novel, user-friendly interface designs as well as system architectures and pilot protocols compliant with regulatory norms, compatible with established workflows, and acceptable within the institutional milieu.

Thus, we interacted extensively with the breast cancer service line, the institutional biobanking platform, and institutional review board (IRB) and Office of Human Research Protections leadership, as well as research scientists, clinical and teaching faculty, IT staff, technology transfer teams, and cross-disciplinary institutional leadership. Concurrently, the ethnography of the specimen procurement supply chain allowed us to map the breast cancer biobank ecosystem [23]. We examined all contexts along the data pipeline, from population-level breast cancer screening to diagnostic biopsies and surgical treatments, clinical pathology, and specimen accessioning through the biobanking platform, where it may be stored for future use in –80 °C freezers or distributed fresh for next-generation biobanking applications such as patient-derived organoids, multi-omics, and high-throughput testing. Given the well-documented challenges for biobank sustainability, we took special interest in learning about economic and logistical challenges pertaining to this sector. Regulatory considerations, operational feasibility, and economic analyses will be reported elsewhere [23].

Prototyping

The first decentralized biobanking prototype established the proof of concept, leveraging ERC-721 NFTs to keep patients connected to donated specimens throughout the research life cycle. The NFT platform was integrated with a novel mobile app for privacy-preserving collaboration among patients, scientists, and physicians in a model breast cancer organoid ecosystem. A second prototype advanced a comprehensive NFT digital twin framework with ERC-1155 modeled using a publicly available real-world organoid biobank dataset (National Cancer Institute Human Cancer Models Initiative) [24,25]. This web-based prototype focused on generating value for scientists, illustrating potential to enhance efficiency, effectiveness, and impact of biospecimen research. Third, no-code front-end mobile app prototypes were developed to demonstrate, test, and refine user interfaces and experiences for the engagement of donors in biobanking.

User Interface and User Experience

We drafted wireframes using anonymous model biospecimen information from the institutional biobank database. App design processes sought to minimize cognitive effort for mobile app users, maximize accessibility across ages and educational levels, and adhere to rigorous privacy standards and customs in accordance with the established biospecimen collection protocols. We progressively simplified and iterated display text and content to make it as concise and concrete as possible and unified across decentralized biobanking app interfaces. To facilitate navigation, we streamlined presentation of content in each of the 4 core interfaces using accordion elements complemented with individual cards for each biospecimen, with pop-ups to guide transitions within and across interfaces. Unified color schemes, fonts, and item designs adhered to predetermined themes with a standardized format that was gradually refined.

The designs were tested and validated via further research surveys and interviews. Immersive design workshops solidified core app requirements. Initial usability testing included online and in-person sessions with clickable prototypes and functional prototype demonstrations followed by usability testing and cognitive walk-throughs on users' personal devices.

Front-End Development and Testing

Finalized mobile app designs were developed using Flutter so that iOS and Android users could participate in the pilot. The apps were tested and deployed to Apple TestFlight and the Google Play Store, allowing for download directly to participants' personal devices. From August 2022 to January 2023, feedback from 110 unique individuals was incorporated, including 45 (40.9%) BDRR members, 28 (25.5%) who downloaded and tested the app on their personal devices, and 14 (12.7%) who viewed personalized biospecimen content within the app interface. The result was a validated app facilitating interaction between donors and biospecimens within the breast cancer biobank, personalized collection content, and mappings from biobank database details.

Blockchain Development

Initial decentralized biobanking prototypes were developed experimenting with different tokenization strategies using Ethereum's ERC-721 and ERC-1155 NFT standards for mapping dynamic relationships among patients, biospecimens, physicians, scientists, and corresponding biobanking and research protocols. However, variable costs of transaction fees (known as gas fees) on the Ethereum network and high friction for blockchain onboarding were major limitations for implementing a real-world pilot.

These constraints informed the design of a functional, blockchain-backed prototype suitable for the pilot population and setting, leveraging a fit-for-purpose blend of centralized and decentralized applications that would enable patients to track and learn about donated specimens appropriate to the highest-order objectives for the first live pilot of decentralized biobanking technology.

A nontransferable ERC-721 NFT, also referred to as a "soul-bound token" [26], was developed to represent each donor's immutable, inherently unique connection to their personal biospecimens. This token [26] was held within a single *externally owned account* that served as a custodial wallet. Of note, our previous decentralized biobanking prototype for organoid research networks, as described elsewhere, used ERC-1155 to advance a comprehensive digital twin ecosystem with NFTs representing patients, specimens, multigenerational derivatives (eg, patient-derived organoids), scientists, and physicians, as well as externally owned biobanker accounts, demonstrating the potential for a sophisticated solution [25]. However, while using the ERC-1155 standard would have offered savings for deploying multiple token collections representative of the entire biobanking ecosystem, applying them to a single soul-bound token collection for this use case would have yielded no additional benefits while adding unnecessary complexity [25].

Each biowallet NFT served as a customized yet anonymous "token of appreciation" for specimen donation coupled with a front-end user experience simulating token-gated access to personal biobank data. This token-gated process was performed manually, minting the tokens individually via the smart contract

interface on Etherscan. Subsequently, the token metadata and transaction details were stored within a secure, IRB-approved database for the eligible user. This created a digital honest broker mechanism for managing in-app participant-specimen engagement without requiring further humans in the loop or revealing donor names or other personally identifiable information to third parties.

System Architecture

The decentralized biobanking pilot system incorporated 3 core components: an *app* overlying *institutional biobank and research infrastructure* with a blockchain-backed *NFT digital twin framework* (Figure 2).

The app used an *n*-tier architecture pattern with interconnected workflows across distinct, modular components with varying responsibilities (Table 1). Our user-friendly mobile app, available on Android and iOS, was powered by applications built using Amazon Web Services. During this initial pilot phase, our system relied on external services and data sources that were not yet directly integrated with our deployed technology. Our NFT framework consisted of an ERC-721 smart contract designed to mint nontransferable, soul-bound biowallet tokens that were deployed to the Ethereum mainnet. Deidentified biospecimen data were provided by biobank personnel to authorized study team members, who would use a secure device to import the records into the pilot system's database. Both required manual processes for pilot implementation.

Figure 2. System architecture diagram—decentralized biobanking pilot app for breast cancer biobank. This system architecture diagram incorporates the decentralized biobanking mobile app powered by internal components that handle business logic, data storage, and data integrations built on a cloud-based infrastructure using Amazon Web Services (AWS); this is flanked by corresponding elements connected via secure authorized access devices for interacting with the nonfungible token (NFT) digital framework's biowallet tokens deployed on Ethereum and institutional data sources from the Breast Disease Research Repository and Institute of Precision Medicine organoid biobank.
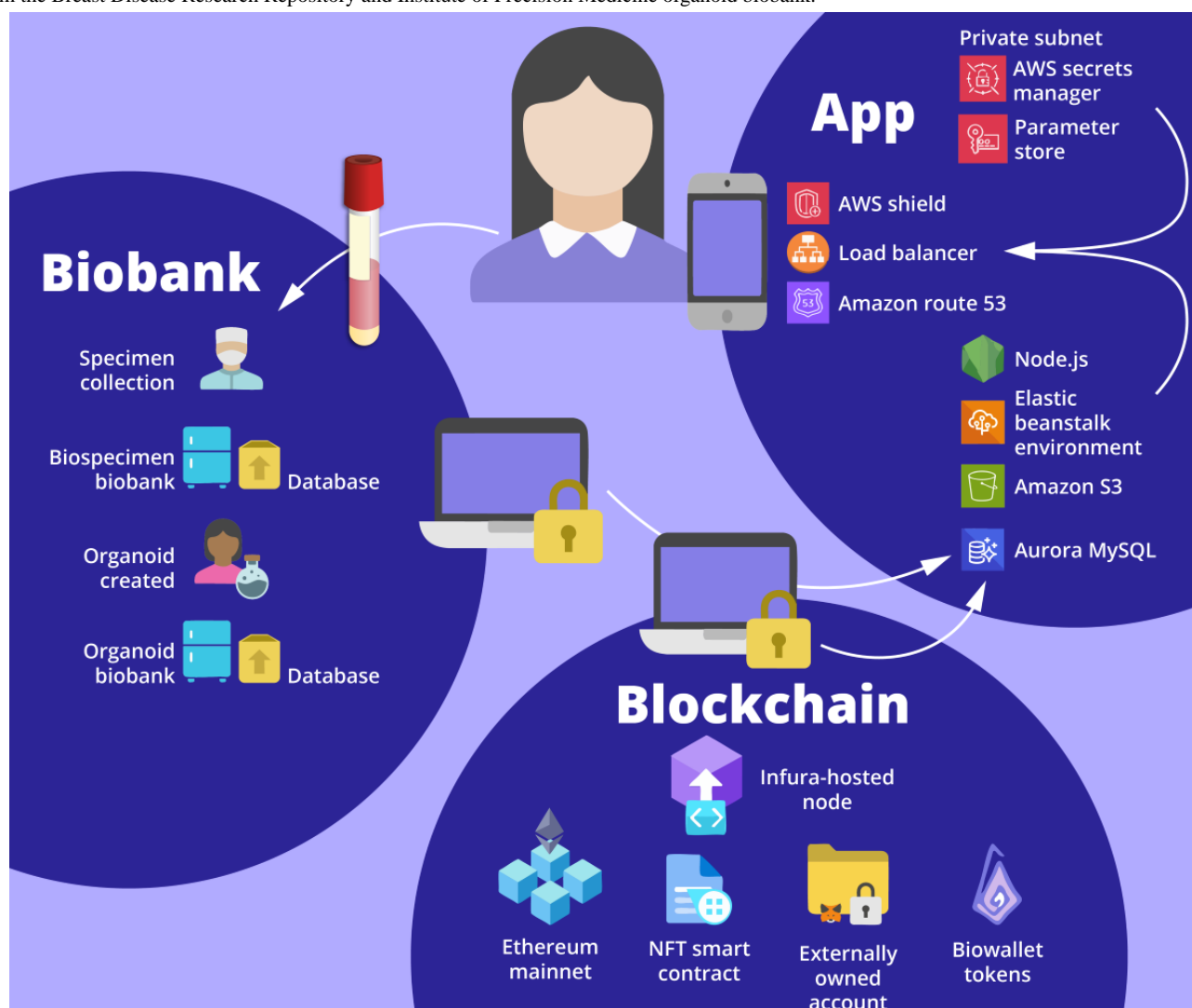
Table 1. Key details of the decentralized biobanking pilot system architecture.

Component	Technical details
App	• Presentation tier: the Flutter mobile app built and deployed using Android Studio (Google) and Xcode (Apple Inc) to enable download to Android and iOS devices. The app provided front-end user interfaces for patients, enabling dynamic interactions, user inputs, and the presentation of queried information from institutional data sources through the app tier. Google's Firebase Authentication services manage account creation and management, encrypting data in transit using HTTPS and at rest using the scrypt standard cryptographic protocol. Passwords are stored securely using encryption, salting, and 1-way hashing following NIST[a] 800-63b recommendations. • App tier: used a Node.js (OpenJS Foundation) server to enable all core functionality and logic of the app, including specimen tracking with enhanced transparency into biobank activities and subsequent research. This layer is also responsible for enforcing security and access rules, handling connectivity to and communication with data sources and external services, and processing data to return to the presentation layer. Deployed on AWS[b] Elastic Beanstalk, the app instances sit behind load balancers for scalability, running in private subnets. • Data tier: hosted by an Amazon Aurora database cluster using the MySQL engine. It hosts a secure, highly available database that stores and retrieves the information necessary for the app to run. This includes donated sample records housed on the BIOS[c] and corresponding biospecimen freezer repositories across 4 physical locations of the Pitt Biospecimen Core, as well as unique cryptographic IDs from Firebase and claimed biowallet NFTs[d] to establish privacy-preserving data linkages between donors and their deidentified biospecimens. As noted in the presentation tier, user credentials for accessing the app are stored separately on secure Firebase servers. • Infrastructure tier: referenced within the app and data tiers, our AWS cloud infrastructure provides the foundation for networking and security, ensuring availability, scalability, and interoperability across system components Multimedia Appendix 1.
Blockchain	• NFT framework: an ERC-721 smart contract designed to mint nontransferable, soul-bound biowallet tokens was deployed to the Ethereum mainnet via a transaction sent to an Infura-hosted node from a local Node.js runtime environment using Hardhat. The overarching framework incorporates NFTs representing all stakeholders, specimens, and protocols, allowing for composable layers of complexity, utility, and value to be built upon the PIO[e] architecture.
Biobank	• Institutional biospecimen and research databases: biobank personnel provided access to deidentified biospecimen data via OneDrive Microsoft Excel (Microsoft Corp) files to an authorized study team member, who would use a secure device to import the updated records into the Aurora database. Similarly, Microsoft Excel files containing biobank (BDRR[f]) registered members were provided by research staff as exported from OnCore. In addition, imaging and research data from an organoid biobank "spoke" were shared via OneDrive, and curated representative datasets were hosted on Dropbox (Dropbox, Inc).

[a]NIST: National Institute of Standards and Technology.

[b]AWS: Amazon Web Services.

[c]BIOS: Biospecimen Inventory and Operations System.

[d]NFT: nonfungible token.

[e]PIO: programmed input-output.

[f]BDRR: Breast Disease Research Repository.

Pilot Study

Participants were recruited via electronic and paper fliers for "Decentralized Biobanking "de-bi": An App for Patient Feedback from Biobank Research Donation" (STUDY22020035). The pilot aimed to recruit 300 participants over 6 to 12 months. App download invites were distributed via email with Apple and Android instructions. IT support was provided as needed, with real-time bug fixes and improvements based on user feedback. App interfaces, design, and features were iterated in monthly sprints. Participatory research, user-centered design, and usability testing, as well as quantitative and qualitative assessments of patient, physician, and scientist acceptability, will be reported elsewhere. NFT minting for pilot performance took place from March 7, 2023, to May 8, 2023. Multimedia Appendix 2 details the pilot recruitment to sample tracking process.

Data Sources and Analysis

The technical data reviewed included conceptual models, technical diagrams, product feature documentation, and screenshots of user journeys as experienced by decentralized biobanking pilot participants using the Flutter app. We also consider biospecimen collection data from the institutional Biospecimen Inventory and Operations System via Microsoft Excel (Microsoft Corp) exports, in-app activity data recorded in a MySQL database, and blockchain transactions on the Ethereum network accessed via Etherscan. Technical feasibility was assessed from feature requirements, interface designs, and quantifiable user experiences from the live implementation. To further evaluate pilot outcomes, we provide simple descriptive statistics from the quantitative datasets and comparative cost analyses for alternative NFT design strategies calculated using values from tokens minted during the pilot. Patient experiences were captured via written feedback from a co-design workshop during the app development phase and a usability workshop session held with pilot participants.

Ethical Considerations

Research was performed under IRB-approved human subjects research protocols and a Quality Improvement protocol (Textbox 1 provides protocol numbers, titles, and approving body).

Participants provided informed consent or the equivalent, in accordance with respective protocols. Conflict of interest disclosures were included in consent documents and verbal disclosures were provided for all online and in-person encounters. All data reported here are either de-identified or anonymized and privacy-by-design was utilized within the de-bi app to maintain confidentiality of participant identities.

Participants were not compensated for participation in the biobank, stakeholder interviews, quality improvement activities or de-bi app pilot study (STUDY19060196, IRB00019273, QRC 3958 and STUDY22020035, respectively). Our foundational research protocol (STUDY22010118) provided $10 gift cards for surveys, with an additional $20 for those who completed follow-up interviews.

Textbox 1. Human participants and quality improvement protocols for technology feasibility.

- STUDY22010118: patient views, preferences and engagement in next-generation biobank research (University of Pittsburgh)
- IRB00019273: nonfungible tokens for ethical, efficient and effective use of biosamples (Johns Hopkins University)
- STUDY19060196: Breast Disease Research Repository: tissue and bodily fluid and medical information acquisition protocol (04-162; Hillman Cancer Center)
- QRC 3958: patient-facing biobank platform development Quality Improvement proposal for Beckwith award–breast cancer supply chain analysis, biobank token model development, and initial pre-pilot testing with University of Pittsburgh Medical Center patients (University of Pittsburgh Medical Center)
- STUDY22020035: decentralized biobanking "de-bi": exploring patients interests in feedback, education, follow-up, engagement and tokens of appreciation regarding biobank donation via mobile and web applications (University of Pittsburgh)

Results

Prepilot Results

A co-design session (n=15) was conducted before the pilot to characterize patient preferences and areas of confusion. This session was one in a series of extensive participatory design sessions, which we have reported elsewhere [23]. Participants were most excited about decentralized biobanking for feedback and recognition ("to see my own cells+know how those cells are advancing science"), community-engaged research ("to connect with others through this app"), and precision medicine potential ("to get helpful results regarding my health"), suggesting acceptance of our vision and overall approach. At the conclusion of this phase, there was still confusion surrounding logistics and governance ("how we *find* our samples and approve their use"), technical concepts ("Why NFT's?"), and unanswered big-picture questions ("Short+long-term—who benefits from this?") regarding the decentralized biobanking platform. Table 2 provides a thematic overview and representative quotes.

Table 2. A thematic overview of participant feedback gathered through a prepilot co-design session.

Theme	Prepilot participant feedback
Aspects participants were "most excited about"	
Personalized feedback and recognition	• "The opportunity to see my own cells+know how those cells are advancing science and clinical care." • "Having knowledge about [sample] types, research and current news about my tumors." • "To be able to follow where my personal donation goes, and what they are doing with it, and what they get out of it."
Community-engaged research	• "Great for mutation studies with multiple primary cancer+tumors." • "Keeping up to date with genetic mutation research." • "I'm excited to connect with others through this app." • "That patients who invest their tissue in research are able to connect as co-investigators."
Potential health benefits	• "I'm excited about the idea that there may be more ways to care for my family—better research practices may enable the medical field to work smarter—maybe ensuring that my children don't need surgery, chemo, etc." • "I am very excited for anything that can improve my health and outcome (and of others)." • "Being able to get helpful results regarding my health." • "I'm excited about the possibility to know how my tissue reacted to a treatment." • "Patient access to personal info/data; Personalized medicine potential."
Aspects participants "still found confusing"	
Big picture	• "Why do people still get cancer, dammit!" • "I don't understand 1) How this may really help me+my family, 2) Short+long-term—who benefits from this? 3) Where does the $ come from? 4) What are we giving up/sacrificing by saying 'yes.'" • "How will Dr. utilize?"
Logistics and governance	• "I don't understand how we find our samples and approve their use—I also don't understand what studies we could 'suggest' or enable through the samples we have provided." • "How likely is it that my samples will be used?" • "Can you use it [de-bi app] even if your surgery already happened?" • "How to get my tissue submitted to researchers."
Unclear technical terms and concepts	• "Not really sure what an organoid is—is it a picture/video of my actual cells or is it a model of my cells?" • "Why NFT's?" • "I am still learning about NFTs and how they will help breast cancer patients." • "How will patients interpret data—will it be translated?"

Overall Pilot Results

Overview

Over 10 weeks of active recruitment (February 16 to April 30, 2023), 1080 unique participants enrolled in the decentralized biobanking pilot, including 9.54% (930/9750) of confirmed biobank members (Multimedia Appendix 3). Approximately 600 app invites were distributed, and 405 participants downloaded and completed app registration, including 361 (89.1%) biobank members. All app users were female (405/405, 100%), and the mean age was 56 (SD 12.8; range 18-87) years, making them younger than both the broader biobank membership and decentralized biobanking pilot participants (mean ages of 64, SD 13.6 and 58, SD 13.1 years, respectively). Multimedia Appendices 4 and 5 detail pilot participant and app user characteristics relative to those of the overall biobank membership. There were 4 key features of the piloted app, as shown in the user journey map (Figure 3). Biobank, biowallet, and profile features and quantified user journeys are illustrated in subsequent Journey sections, and laboratory features and respective user journeys for that context are also described in detail elsewhere.

Figure 3. Decentralized biobanking platform user journey. The user journey map demonstrates the status quo of the patient experience with biobank donation as well as the 4 key features of the decentralized biobanking mobile app that was piloted for a large breast cancer biobank member population from January 2023 to May 2023. Each of the columns represents primary activities within the different core screens of the decentralized biobanking mobile app, which the invited participants downloaded to personal iOS and Android devices. The Biobank, Biowallet, and Profile sections are illustrated with key activities and features. The Lab section on the far right is illustrated, although the journey for the community engagement feature is outside the scope of this study and is addressed elsewhere. NFT: nonfungible token.
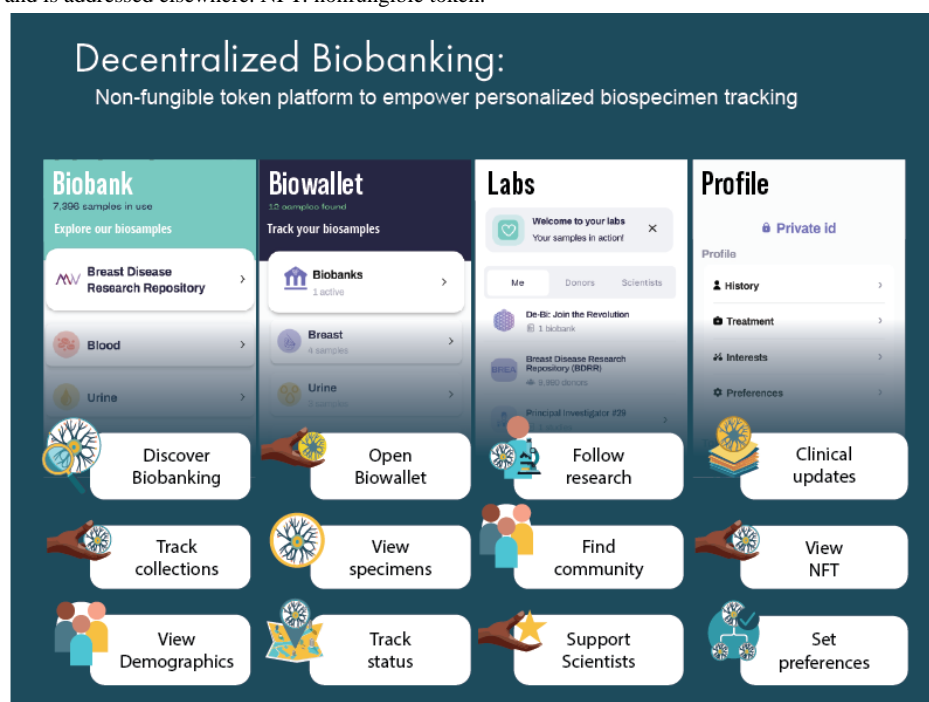
Journey 1: App Onboarding and Biowallet NFT Minting Process

Upon downloading the app, users entered their name and birth date, triggering verification of biobank membership and sample collections, with "biowallet NFT" minting, if applicable, serving as a digital representation of membership in the biobank donor community, delivering a user experience of a token-gated bridge between the user's app and specimen data, if available (Figure 4).

Figure 4. Opening a biowallet—simulation of token-gated specimen access. The process of opening a biowallet required participants to enter their name and date of birth, triggering the system to match participants to corresponding members in the biobank (Breast Disease Research Repository). Once specimen status was established, biowallet nonfungible tokens were minted, specimens were linked to the account, and email notifications indicated to participants that their biowallet was available.
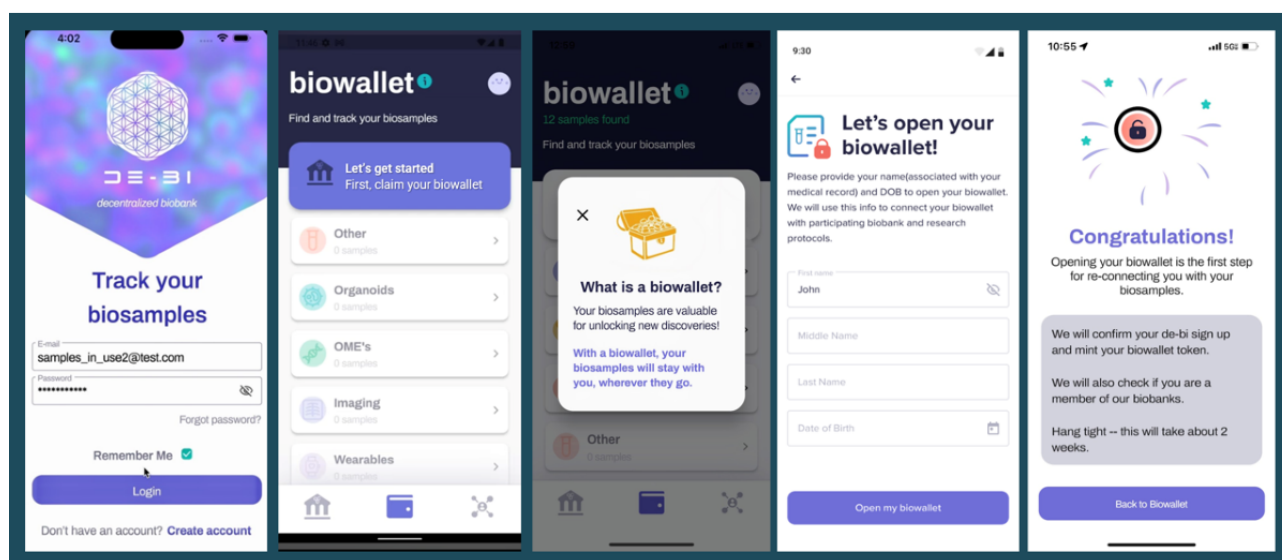
Simulated Token Gating Workflows

Once users entered their name and date of birth into the decentralized biobanking app, a manual, coordinated effort involving biobank personnel and authorized study team members verified each user's biobank consent and matched donors to their respective biospecimens via a unique anonymous study ID linked to a Firebase (Google) unique ID associated with their decentralized biobanking app account. During this process, study team members would also mint a unique biowallet token for each verified donor with specimens. These tokens were held in a custodial wallet, but each token identifier was linked to donor records within the Amazon Aurora database to establish a second privacy-preserving mechanism for data linkage.

Firebase established the functional linkage to allow for proper access control and permission management within the app for this pilot, whereas the biowallet NFTs and the act of claiming were representative as a proof of concept as well as a token of appreciation for participating donors. This decision was made to limit excess complexity related to using web3 technologies

as a barrier to participation for this population while providing a comprehensible introduction to the concept of NFTs for establishing relationships between donors and their samples. Our aim was to ensure that donors were not excluded from engaging with the platform based on the extent of their blockchain expertise.

Various criteria for minting Biowallet tokens were considered for entire pilot and biobank deployment. Using variation in token minting costs observed throughout the pilot study to model minimum, average, and maximum costs (US $1.84, US $4.51, and US $11.23, respectively), the selected model, minting tokens for all 272 pilot participants coenrolled in the biobank with one or more specimens collected, was projected to cost US $1226.72 (SD US $41.91; range US $500.48-$3054.56, Figure 5A, left). Extended entire biobank implementation, this model is projected to cost US $17,769.40 (SD US $159.52; range US $7265.62-$44,229.27; Figure 5A, right). Other models, such as specimen distribution to a research protocol or biobank membership were also considered.

Figure 5. Nonfungible token (NFT) minting costs and calculations for the breast cancer biobank pilot. (A) Pilot implementation—comparison of biowallet token minting criteria for the total cost of pilot deployment. Cost analysis used variation in token minting costs observed throughout the pilot study to model minimum, average, and maximum costs (US $1.84, US $4.51, and US $11.23, respectively). *Selected token minting criteria for the decentralized biobanking pilot. (B) Transaction costs in US $ and ether (ETH) are illustrated for 151 NFTs minted during the decentralized biobanking pilot. (C) Timeline mapping variable cost of biowallet minting events and cumulative costs of minting 151 NFT biowallet tokens throughout the decentralized biobanking pilot.

Token Minting Costs

The cost of deployment of the biowallet NFT protocol on Ethereum was US $223.52. A total of 151 biowallet tokens were minted for US $680.49 at an average of US $4.51 per token (SD US $2.54; range US $1.84-$11.23; Figure 5B. Biowallet tokens could be requested by decentralized biobanking pilot

participants who downloaded the app and had one or more specimens collected (148/405, 36.5%). For context, procurement, processing, storing, and disbursement of biospecimens in this institutional biobanking platform costs an estimated US $1600 per case.

Biowallet tokens could be requested by decentralized biobanking pilot participants who downloaded the app and had one or more specimens collected (148/405, 36.5%). During the pilot, 140 total tokens were requested and minted for eligible participants. Minting events varied in cost based on fluctuating transaction fees and the number of participants who had requested biowallet tokens since the last token minting event. For instance, minting events ranged from US $3.11 for minting one token, to US $288.52 for minting 80 tokens in the first batch (Figure 5C).

Journey 2: Biobank Orientation and Research Profile

After requesting a *biowallet*, users were directed to visit the *biobank*, where they were oriented and learned about the overall biobank inventory and activities, including demographics of the consented donor population, framed as "biobank members"; informed consent content; principal investigators; and respective biobank operations and research activities for entire specimen collection (Figure 6). We included education about research protocol development, IRB oversight, procedures for specimen allocation, and investigator- and protocol-level transactions. The biobank displayed 60,973 biospecimens from 3940 unique donors collected from February 1995 to May 2023 and updated on a regular basis, with 318 new specimens added during the pilot. The feature tracked collection and distribution totals for the biobank, with breakdowns for each specimen type (Table 3).

The "profile" allowed participants to enter clinical history and treatments relevant for research on their specimens. We also assessed research interests, privacy preferences, engagement interest, and willingness to donate additional specimens to scientists as needed. In total, 37.8% (153/405) of the app users completed one or more portions of the profile, including 37.1% (134/361) of the biobank members. The profile also displayed the random "Private ID" number, which enabled users to remain deidentified while linking to their respective specimens. During the pilot, we experimented with the naming conventions, location, and order of presentation of biobank and profile features to assess impact on participants' understanding of the biobank environment, affordances, constraints, and opportunities presented by the decentralized biobanking platform.

Nearly all participants who filled out the research profile (151/153, 98.7%) added one or more clinical details (eg, familial history of breast cancer; Multimedia Appendix 6). Profiles were completed by 39.9% (59/148) of the participants with samples, collectively annotating 886 specimens, including 760 (85.8%) available for future use, 36 (4.1%) "on hold" for a designated protocol, and 90 (10.2%) that were distributed for research, with information that was not contained within the institutional biobank database. In addition, participants added preferences regarding specimen use, willingness to provide further data and specimen donations, and future research engagement.

Figure 6. Biobank orientation journey, illustrating the biobank screen and user workflow introducing app users to biobank processes, what it means to be a biobank member, and regularly updated snapshots of investigator activities, protocols, and specimen allocations, at the level of the overall bank. The biobank also linked to participant's personal research profile, where they could provide key clinical details, interests, and preferences related to research on their specimens.
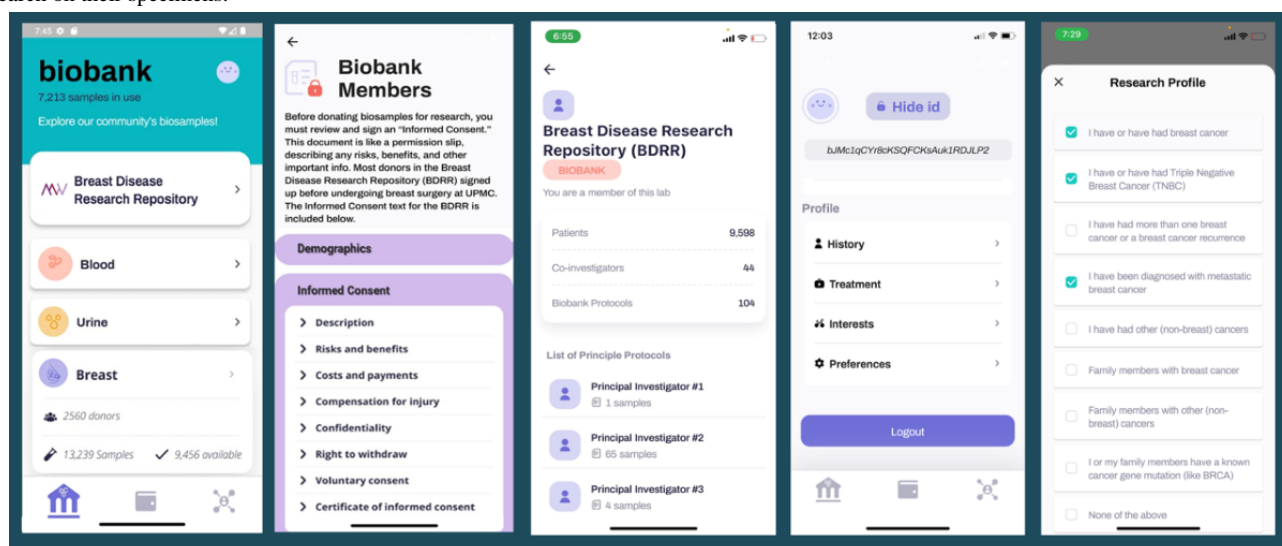
Table 3. Decentralized biobanking pilot population, app user and token claiming overview.

Pilot population	App users, n (%)[a]	Token claimed, n (%)[b]
Total (N=1080)	405 (37.5)	130 (12.04)
Biobank members (n=930)[c]	361 (38.82)	128 (13.76)
Biobank members with specimens (n=272)[d]	148 (54.41)	125 (46)
Collected specimens (n=3904)[d]	2133 (54.64)	1812 (46.41)
Biobank members with specimens in use (n=165)[d,e]	88 (53.33)	74 (44.85)
Fresh (n=90)	46 (51.11)	40 (44.44)
Frozen (n=100)	50 (50)	40 (40)
Specimens in use (n=377)[d,e]	202 (53.58)	177 (46.95)
Fresh (n=195)	104 (53.33)	95 (48.72)
Frozen (n=182)	98 (53.85)	82 (45.05)
Number of donors with specimens available (n=242)[d]	132 (54.55)	110 (45.45)
Breast (n=147)	81 (55.1)	67 (45.58)
Blood (n=185)	97 (52.43)	82 (44.32)
Urine (n=166)	91 (54.82)	80 (48.19)
Specimens available (n=3309)[d]	1757 (53.1)	1522 (46)
Breast (n=345)	205 (59.42)	178 (51.59)
Blood (n=2172)	1145 (52.72)	988 (45.55)
Urine (n=783)	406 (51.85)	355 (45.34)

[a]Specimen values and donor counts for all app engaged participants with specimens collected.

[b]Specimen values and donor counts for all app engaged participants with specimens collected who claimed biowallet tokens during the pilot study.

[c]Donor counts for all biobank consented pilot participants.

[d]Specimen values and donor counts for all biobank consented pilot participants with one or more specimens collected.

[e]Specimens considered in use if distributed to a research protocol as of May 4, 2023. A total of 218 specimens among all pilot participants with collected specimens designated "on hold" for future research use are not shown.

Journey 3: Claiming and Viewing the Biowallet NFT

Overview

Linking app accounts to biospecimen data occurred offline and took up to 2 weeks supported by software scripts and manual processes, including checks for false mismatches (eg, due to typos). Once biowallet NFTs were available, email notifications prompted participants to log in to their decentralized biobanking app to open their biowallet and access their personalized biospecimen data.

Once claimed, the "Biowallet token" appeared on the bottom of the screen with a link to view the corresponding Ethereum transaction data (Figure 7). The profile screen showed how patients could add clinical details that are not in the biobank database, making their biospecimens more readily discoverable by prospective users, reducing reliance on third-party chart review during study planning. The biowallet NFT signified membership in a collective committed to breast cancer research. Once claimed, the individual's unique biowallet NFT could be viewed via an in-app Etherscan display. The app user experience represented this process as a symbolic "token of appreciation" as a form of reciprocity for biobank contributions.

XSL·FO
RenderX

Figure 7. Claiming and viewing the biowallet nonfungible token (NFT). The figure illustrates the biowallet NFT claiming process, first showing the appearance of the biowallet when the token is available to be claimed. Next, the claiming process is shown, which invites donors to "join the revolution!" Once claimed, the user's personal NFT is represented on the profile page, which is connected via a hyperlink and an in-app display of the Etherscan view of the NFT, a customized biowallet logo for the pilot, and corresponding blockchain transaction data.
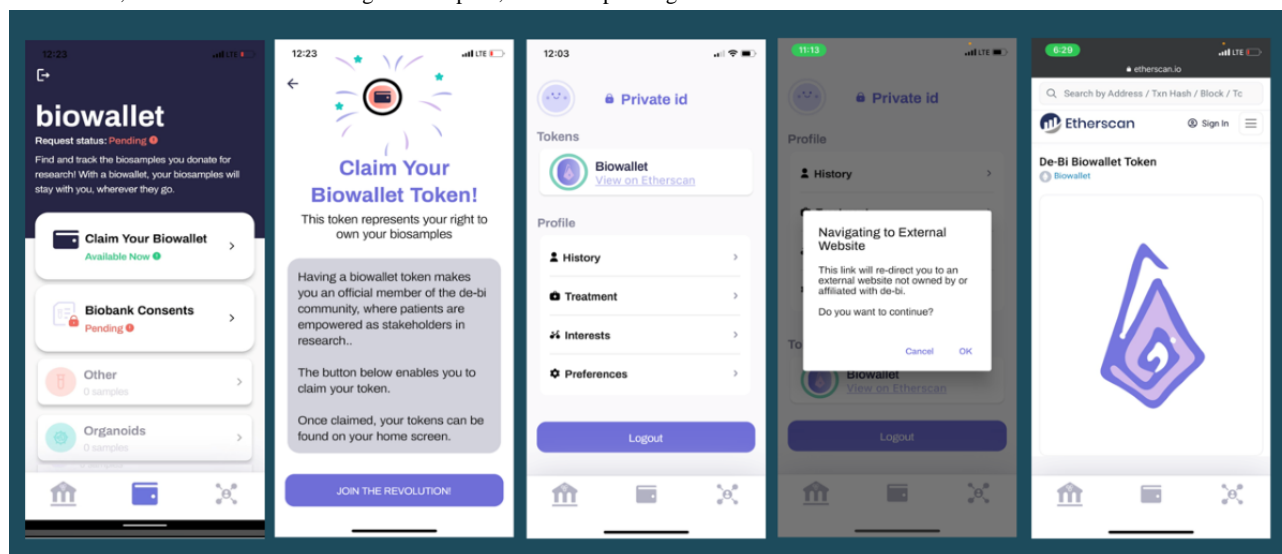
Proof of Concept for Blockchain-Backed Biobanking App

The initial round of minting included "tokens of appreciation" for participants who were active in the demonstration phase of the app design and usability testing. The blockchain mechanism was initially tested with 4 test mints followed by minting "tokens of appreciation" for 7 demonstration phase participants. In total, 71% (5/7) of the demonstration users successfully completed the token minting claiming process, illustrating the use of the "biowallet" NFT as a representation of the individual's membership in the biobank community. After validating functional integration of the blockchain simulation, eligibility for biowallet tokens was limited to those with confirmed specimens in the breast cancer biobank, enabling us to simulate use of the NFTs to establish token-gated access to deidentified specimen accounts.

Of 148 app users with specimens, 140 (94.6%) initiated the biowallet token minting process during the pilot. Of 140 tokens minted, 125 (89.3%) were claimed by users, with an average of 10 (median 1, IQR 1-5, range 0-100) days between token minting and token claiming (Figure 8). Compared to individuals who did not claim their biowallet, those who did claim their biowallet were slightly younger (average of 58.9, SD 10.8 vs 61.9, SD 14.3 years) and had a similar time since biobank consent (7.8, SD 5.0 years since consent for claimants vs 7.7, SD 5.3 years for nonclaimants; Multimedia Appendix 7).

Figure 8. Nonfungible token claiming details for the decentralized biobanking breast cancer biobank pilot. Participant engagement and timing illustrates (A) interest in biospecimen tracking and receptiveness to email notification to facilitate the token claiming process and (B) the effective reconnecting specimens to participants that occurred during the pilot as tokens were claimed. In total, 89.3% (125/140) of tokens minted for app users with specimens were claimed during the pilot. Tokens were considered unclaimed after ~2.5 months following the final token minting event. A total of 15 participants had not yet claimed their token as of the conclusion of the pilot.
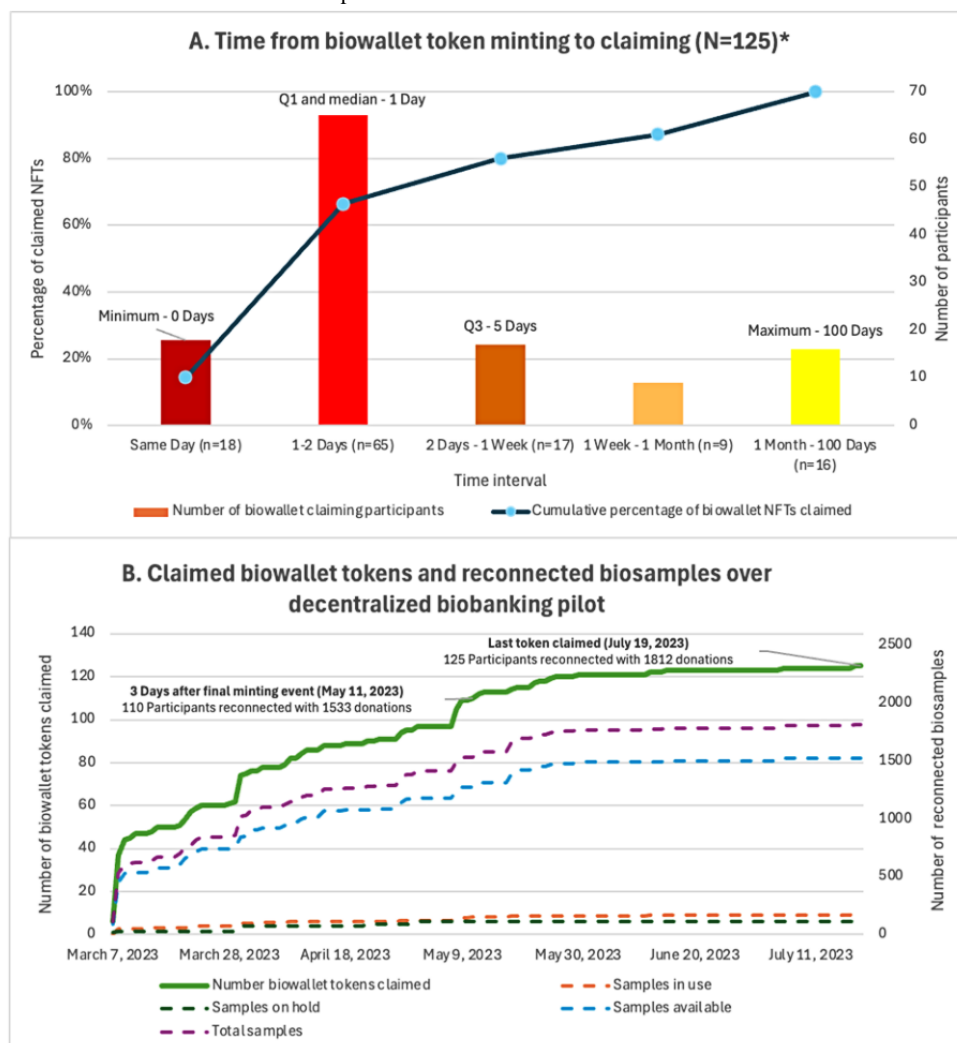
Ethnography of the US cancer specimen supply chain, including engagement with industry and academic stakeholders, generated the following conservative estimates for the commercial value of cancer tissue, blood, and urine specimens with well-annotated clinical data: US $1000 for cancer tissue, US $500 for blood, and US $300 for urine. Hypothetically, this equates to US $1 million of "available" specimens being populated into app users' biowallets during the pilot. Similarly, this corresponds with a total value of approximately US $30 million for unused specimens in frozen storage, with roughly US $7000 in value per specimen contributor. Additional details of the scalability and economic feasibility of the proposed blockchain solution will be addressed elsewhere.

Journey 4: Viewing Personal Specimen Details

The "biowallet" was where participants could view details about when they consented for biobank donation (Figure 9). Once linkage between the user's app and respective biobank data was established, individuals were able to track and learn about their own biospecimens. Details available via an interactive accordion feature included their biosample collection date, sample type and medium, if and when each sample was shared for a

particular research protocol, and similar sample-level information within the institutional database. The biowallet also includes a taxonomy of physical and digital biospecimen data types that may, in the future, be trackable by individual participants.

Further details regarding specimen distribution and availability were indicated via additional pop-ups, providing users with an opportunity to navigate to an app-based laboratory. Here app users could learn how many donors had contributed specimens of similar types, or had specimens distributed to the same research protocol. Of the biobank members using the app, 41% (148/361) had their "biowallet" populated with a total of 2113 specimens (mean 14.4, SD 12.1; range 1-84), including 1414 (66.9%) blood specimens, 419 (19.8%) urine specimens, and 296 (14%) breast tissues. In total, 70.9% (105/148) of sample holders had one or more breast tissue specimens. A total of 59.5% (88/148) had one or more specimens "in use" (mean 2.3, SD 1.6 per person; range 1-8), 40.5% (60/148) of the participants with specimens had none "in use," and 4.7% (7/148) of the participants had specimens "on hold" (mean 24.9, SD 16.3; range 10-61). Individuals who had no specimens available

received a digital biobank membership card (Figure 9, panel 2) and in-app text notifying the participant that no specimens had been located (yet), with a range of possible explanations.

Collectively, 202 of app users' specimens were "in use," including 104 (51.5%) that were delivered "fresh" the day of donation (eg, for organoid development) and 98 (4%) from a frozen collection. A total of 8.2% (174/2113) were "on hold"

for a designated study, and 83.15% (1757/2113) were "available." App users' specimens were distributed to 22 different investigators under 42 research protocols. Between February 15, 2023, and May 4, 2023, users donated 39 new specimens, which appeared on the app, 2 (5%) of which were distributed fresh. In addition, 18% (7/39) were distributed from frozen storage, and 54% (21/39) were placed "on hold" during the pilot.

Figure 9. Biowallet sample tracking journey. This figure illustrates the participant experience learning about their personal specimen donations via an interactive biowallet landing page. Pop-up and accordion features enabled participants to learn about their specimens, including the type, collection date, distribution to a research protocol versus availability for future use, and explore further details about similar donations and distributions.
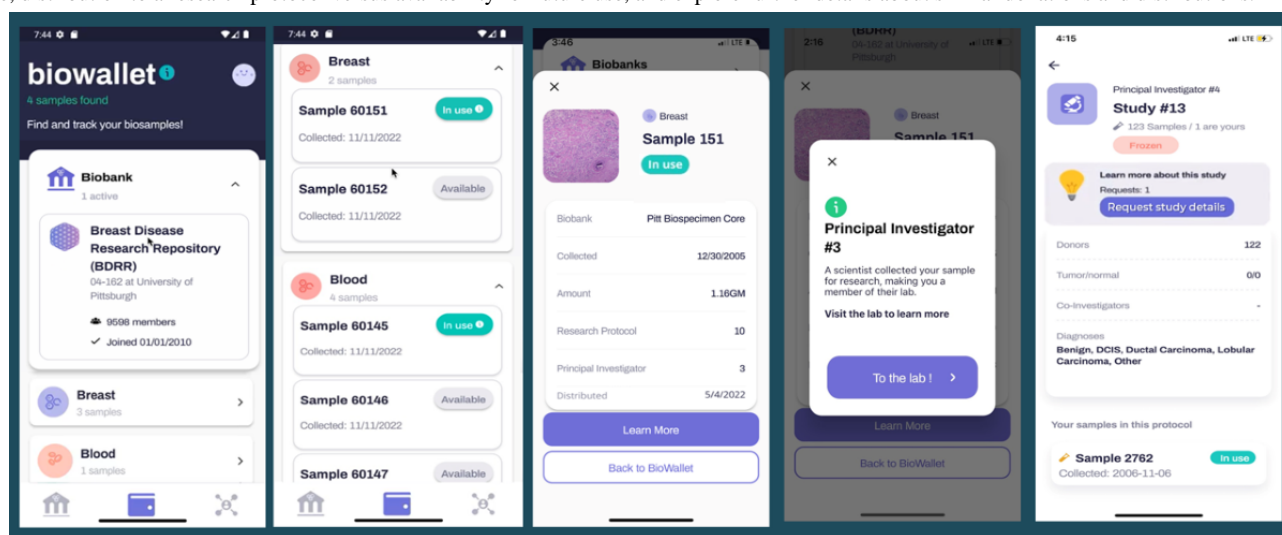
Participant Feedback During the Pilot

During the pilot, cognitive walk-throughs with participants illuminated areas of interest along with potential opportunities for design improvement. Key areas of excitement included seeing how their samples were used. One participant stated the following:

I will [otherwise] never know anything about my cells.

Areas for improvement included improving technological accessibility (eg, making it iPad compatible) and clarifying the information presented (eg, "Will there be a way to learn more about each study?"). Table 4 provides a detailed thematic overview and representative quotes.

Table 4. A thematic overview of participant feedback gathered through cognitive walk-throughs conducted during the pilot.

Theme	Pilot participant feedback
Things they liked	
Big picture	• "This is cool on so many levels." • "Incredible concept to learn about." • "There are endless possibilities and uses for this." • "There is hope for others by giving my cells."
Personalized feedback	• "I can't wait to see what's being done with my samples!" • "Loved the idea of having access to my tissue info+how the two cancers are connected." • "I will [otherwise] never know anything about my cells." • "I'll get to see the process."
Empowerment	• "Information I could never access before." • "Give patients more control and information." • "Profile preferences—great idea." • "Private ID+Ability to connect w/ others in similar diagnosis."
User interfaces and user experience	• "Menus under biowallet are clear+concise." • "Look of the app." • "Easy to navigate." • "Easy to use/menus good." • "Love the status of 'in use' and 'available.'"
Things they did not like or that did not meet their expectations	
Information provided	• "Where are investigators that have my tissue or samples." • "Unclear when no samples (needs explanation)." • "Will there be a way to learn more about each study?" • "I need a little more background before fooling around with the app."
Accessibility	• "Needed tutorial." • "Are there options for people who do not have email on their phone." • "Under personal history, other than TNBC (triple negative breast cancer) other breast cancers should be identified." • "Need to be able to use on an iPAD for larger screen." • "Possible to put app on android tablet?" • "Being older I'm not a techie and it takes a while."
Functionality and user navigation	• "Biowallet should be first icon." • "Make biobank/wallet first tab." • "Add search bar in connect." • "Some functions are more intuitive than others—more prompts are needed." • "What was the purpose behind 'home' icon community samples."

Discussion

Principal Findings

The decentralized biobanking pilot demonstrated the technical feasibility of design, development, and implementation of a user-friendly app to deliver transparency and engagement for donors to a well-established biospecimen collection protocol at a US academic medical center. Over 400 participants downloaded and tested the decentralized biobanking app during the pilot, asserting interest in tracking their biospecimens, demonstrating the usability of a patient interface for institutional biobanking data. "Biowallet" tokens (ERC-721) were minted for app users with confirmed specimens, and 89.3% (125/140) successfully claimed their NFTs on the app, with over half (72/125, 57.6%) of the population achieving the task within 1 day of token minting.

Pilot participants' biowallet token claiming process symbolically asserted their right to know what happens to their inherently unique biospecimens, to which they are immutably linked via a nontransferable, one-of-a-kind relationship. The user experience simulated an NFT-gated process, functionally reconnecting app users to >1800 deidentified specimens, providing visibility of affiliated community members and related research activities all while preserving confidentiality. Critically, this was achievable with data architecture, interfaces, and workflows that maintained compliance with preexisting deidentification standards and specimen collection and distribution protocols.

Similarly, we showed how integration with institutional biobank infrastructure can passively provide transparency for donors without imposing undue burdens on investigators or relying on individual research programs to sustain community engagement. Transparency in biobanking has the potential to rebuild donor

trust in biobanks and improve accountability in biomedical research [27-29]. Consequently, transparency may be a driver to improve biobank donations, particularly among communities with historically rooted distrust of biomedical research [30,31]. The decentralized biobanking framework also allowed for the retrospective and prospective onboarding of donors, demonstrating the potential to convert existing biobanks to a progressively decentralized, patient-centered model.

Minting biowallet NFTs averaged US $4.51 (SD US $2.54; range US $1.84-$11.23) per token, with a projected total cost of US $17,769.40 (SD US $159.52) for all biobank members with specimens. Importantly, a 1-time minting expense of <US $5 per patient may be considered marginal, especially in view of the cost of specimen procurement, storage, and distribution. A workshop on biospecimen economics found the cost of operating a large biobank to be US $861 per patient [32]. The value of the specimens themselves is also substantial relative to minting expenses; academic researchers may pay up to US $200 per sample, whereas commercial entities may pay up to US $20,000 per sample [32]. When biospecimens are converted into living models (eg, organoids), the expenses of both processing and development increase, but the value is multiplied several-fold as 1-mL aliquots of the model may cost upward of several thousand dollars per copy for academic and commercial users alike [33,34].

Importantly, we also demonstrated how empowering patients may in turn help scientists by allowing them to annotate their biospecimens with relevant data that may not be represented in the institutional biobank database or may be otherwise not directly available to prospective or current specimen users. Over 37% (150/405) of the participants demonstrated how longitudinal donor involvement might be leveraged to improve biosample curation and discoverability, creating opportunities to enrich research; link siloed datasets; and drive more efficient, community-driven use of biobank resources. Enhanced annotation of biospecimens with clinical data reflects increasing demand among the biobanking community to gain more contextual biospecimen data [35]. Project LUNGBANK is an example of ongoing efforts to provide more comprehensive clinical data to enrich biospecimens [36]. In LUNGBANK, clinically relevant findings collected through manual chart review of patient medical records were used to annotate biospecimens [36]. For the decentralized biobanking app, more intuitive, strategic placement of the profile feature and improved framing of its functionality and benefits for donors and scientists will be essential to optimize the utility of this feature.

Although relatively limited in functionality compared to the NFT framework advanced in our preclinical prototypes, the blockchain aspect of the piloted app was significant for several reasons. First, it represents the first time that most of our participants, including several octogenarians, had ever interacted with blockchain technologies. Second, persistence in overcoming the friction of onboarding related to the blockchain elements served as further evidence of the high value that patients place on tracking their specimens, to the point that they were willing to participate in a cumbersome, multistage process that, in some cases, took weeks. Third, the blockchain aspect of the piloted app remains a permanent, institution-agnostic record of the relationship between specific donors and their respective biospecimens, highlighting the potential to reunite individuals with these deeply personal assets, with yet unmet potential for assurances of trust and shared rewards of research. Finally, the biowallet NFT represents a foundational gateway to a composable and progressively decentralized biobanking ecosystem. That which starts with 1 biowallet token per participant who contributes specimens may be built upon in a stepwise manner, forging an interconnected overlay network that recognizes and unlocks value across today's siloed biobank landscape.

Limitations

The pilot relied on manual data workflows to enable demonstration of a functional decentralized biobanking platform without requiring full integration of the patient-facing apps with the enterprise system. Such manual workflows are impractical for sustainability and scalability. The exponential growth of health information and advanced computing makes workflow automation increasingly fundamental [37]. Thus, application programming interface (API) integration and automated processes will be necessary for future apps. In view of the volume of requests received during the pilot as well as interest in expanding the program to other institutional biobanks, hospital leadership approved API development to facilitate such integrations for the next stages of the pilot program. In addition to being essential for technical feasibility, this approval was critical as it demonstrated that the manual aspects of our workflows were not material for the acceptability of our strategy for reconnecting donors with their deidentified specimens within institutional biobanks.

Notifications based on in-app activity event triggers were not fully implemented during the pilot, and a number of manual steps were required, including substantial coordination across study team members and email-based messaging to notify participants about critical changes such as token availability and biosample status updates. Automated communications must be incorporated into future pilots with accommodation for a range of patient preferences and values. Subsequent development will also make a web-based version to avoid exclusion of participants for whom smartphone apps may not be preferred or accessible, particularly with respect to age and household income [38].

Furthermore, the piloted app interfaces and user journeys were designed for patient users, whereas engagement with physicians, biobankers, and scientists occurred via alternative channels (eg, email and institutional platforms). This limited the functionality and value within the app as research content was high level, limited to the scope of the biobank database. Ongoing work is advancing real-world applications of decentralized biobanking for scientists and other stakeholders within the NFT digital twin ecosystem. Inclusion of professional users directly within the decentralized biobanking platform will be key for unlocking the ongoing value and network effects of our framework.

Regarding the blockchain elements, the high and highly variable costs of token mints on Ethereum illustrate the importance of more cost-efficient strategies, such as layer-2 solutions, for full-scale implementation. Importantly, our focus on the primary

NFT digital twin framework centers the stakeholders and their relational mappings within the ecosystem. This allowed us to focus on tokenizing the individual participants, in this case, 1 token per biospecimen donor rather than 1 per biospecimen, which would have increased costs 10- to 20-fold. This was sensible, especially considering limitations on functionality of a specimen-representing NFT in the setting of our pilot app; that is, it was not necessary to tokenize specimens for implementing transparency and our study did not provide additional permissions relevant to potential tokenized specimen utility for shared governance or profit sharing regarding the underlying biobank assets. Moreover, ensuring the long-term economic sustainability of biobanks is already a salient concern, with high costs driven by human resources, equipment, and sample handling [39-41]. Cost-effectiveness will be essential for broader adoption of decentralized biobanking technology, and blockchain solutions in themselves must be complemented with social, cultural, and legal innovations to enact meaningful progress [40,42,43].

In addition, NFTs were minted for individual participants, and personal NFTs were rendered via an in-app Etherscan display, although the token-gated aspect of the app leveraged Firebase Unique Identifiers rather than NFTs to minimize complexity and potential points of failure. Simulation of the user interface and user experience of blockchain interactions was necessary to overcome barriers to onboarding inherent to contemporary avoidances and constraints of decentralized apps, particularly as our patient population was older and almost exclusively from non–digital native generations and many were actively grappling with cancer. This was especially critical given concurrent educational barriers surrounding the simultaneous introduction of patients to both biobanking and blockchain for the first time. For example, a knowledge assessment on biobanking administered to biospecimen donors found that approximately half of all questions were answered either incorrectly or with "I don't know." Similarly, most patients we engaged with during app design, development, and pilot-testing were not familiar with the term "biobank," illustrating the fundamental challenge of delivering a patient-friendly biobanking app. These findings underscore the gap between providing information during the prospective informed consent process and achieving true comprehension via enduring transparency and ongoing feedback [44,45]. To this end, we prioritized orientation to biobanking and developed lexicon and app design features that make data within biobank databases accessible to donors via a decentralized biobanking platform that coheres with the ethos of decentralization at its core.

For future implementations, we aim to advance blockchain-backed solutions with seamless onboarding experiences through the exploration of newer standards such as ERC-4337 for account abstraction, which awards the programmable flexibility to remove complex barriers to entry such as the current requirement for users to create their own third-party wallets to interact with the decentralized app. Advancement of these technologies may provide seamless integration of decentralized biobanking platforms with both institutional databases and blockchain overlay networks, with future potential to unite participants, specimens, and scientists across various institutions. Transparency and engagement in biospecimen management is a necessary step toward institutional transformation to achieve community partnership, shared decisions, and progressive democratization. More research is needed to test our hypotheses about the role of blockchain technology in a comprehensive and universal decentralized biobanking solution [46].

The success of our pilot inspired potential to revolutionize biobanking via a decentralized platform but also revealed challenges and limitations for current biospecimen collection workflows, standard operating procedures, and data management strategies [47]. Implementation of transparency for past, present, and future biospecimen collection and distribution will require innovative system designs that overcome idiosyncrasies of individual biobank databases coupled with incentive structures and governance models that promote trust and ensure that biobanking practice optimizes individual and collective interests for patients, scientists, and society [48-50]. While the principles and techniques demonstrated in this study theoretically translate to any other research biobanking context, our technical approach must be validated across a variety of clinical and socioeconomic settings, institutional and regional cultures, and biomedical research contexts.

Critically, this pilot addressed a single, disease-focused university biobank with a largely White, female, and geographically localized population. Technology acceptance must be confirmed for diverse patients, diseases, and contexts [51]. Both iOS and Android users were included, yet some did not use smartphones, and others preferred not to download apps. We have since developed a web-based platform, expanding availability to anyone with internet access, although disparities persist. Ongoing research is exploring the impact of age, race, time elapsed since surgery, and stage of disease on technology acceptability, as well as how to optimize recruitment and trustworthiness for underserved populations [51,52]. Current work is also addressing populations such as those with prostate and lung cancer in which male individuals are more heavily represented, and we have incorporated socioeconomic assessments into our data collection to ensure that we advance solutions that are broadly accessible and applicable, especially for economically and educationally marginalized groups.

Looking ahead beyond feasibility, the practical implementation of scalable, decentralized biobanking solutions requires technical enhancements to overcome the discussed challenges and limitations of this pilot. User interfaces must prioritize usability, comprehensibility, and accessibility by leveraging new standards for account abstraction to reduce the complexity of interacting with blockchain components in our solution. Similarly, ongoing research should inform iterative refinement of different strategies for effective presentation of research-related information curated for diverse patient populations. Efforts toward long-term sustainability should include app cost optimization techniques such as deployment on layer-2 networks for major reductions in blockchain transaction costs and the automation of key workflows and processes through proper integration with institutional software and databases. Because each new environment can be quite nuanced, the application of our technology to new use cases will still require custom

configurations when onboarding, but some of these efforts may be streamlined by standardizing integration patterns with widely used laboratory information management systems and research tools.

Finally, our privacy-by-design approach requires due diligence in execution to mitigate risks to users. Abiding by security best practices in development and thorough vulnerability testing are essential measures in protecting against critical security risks. Intentional disaster recovery plans with detailed incident response protocols for specific events are important for prompt threat containment, recovery of system resources with minimal downtime, and communication to affected users and stakeholders. Proactive preparation to set up comprehensive monitoring, automated backups with manual snapshots across system resources and environments, and pre-emptively programmed functionality for pausing and redeploying compromised system components or deployed smart contracts are crucial for the effective execution of incident response plans.

Conclusions

This pilot demonstrates the technical capacity and resources for a functional decentralized biobanking software app that empowers patients to track specimens donated to a real-world breast cancer research biobank with a novel implementation of blockchain technology. The patient-friendly mobile app renders institutional biobank inventory and transactions in a meaningful, personalized biowallet context, providing a rewarding user experience. We demonstrated the app's readiness for API integrations, which would allow for sustainable and scalable implementation across multiple biobank protocols by seamlessly and dynamically displaying biobanking activities to donors. Pilot participants successfully claimed NFTs within the app, restoring provenance for personal biospecimens and related data. This advancement introduces a new paradigm for ethical biobanking, fostering donor engagement and inclusion in personalized research networks appropriate to contemporary learning health systems and mobile computing capabilities while maintaining deidentification and compliance with established protocols.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

WS, RCM, JK, MM, and MG are shareholders in de-bi, co, a company created to advance decentralized biobanking technology to empower patients, accelerate science, and realize precision medicine, and ME is a consultant thereof. The pilot study described in this paper was not sponsored by de-bi, co; however, detailed conflict of interest disclosures were incorporated into all informed consent forms and quality improvement procedures, and verbal disclosures were made throughout pilot engagement. Conflicts of interest were disclosed in accordance with the pilot protocol under Conflict Management Plans, and the principal investigator on the decentralized biobanking app pilot protocol was nonconflicted.

Multimedia Appendix 1
Architecture details of the deployed cloud infrastructure, encompassing networking setup, front-end and back-end deployment configurations, database architecture, continuous integration and continuous delivery processes, domain management, and monitoring systems.
[DOCX File , 17 KB - bioinform_v6i1e70463_app1.docx]

Multimedia Appendix 2

Workflow for the de-bi pilot, including consent for the pilot, downloading the app, and minting biowallet tokens to link personal biospecimen data to their biowallet.

[DOCX File , 179 KB - bioinform_v6i1e70463_app2.docx]

Multimedia Appendix 3

Decentralized biobanking pilot study participation rates among eligible biobank members by age, race, and time from initial biobank consent.

[DOCX File , 31 KB - bioinform_v6i1e70463_app3.docx]

Multimedia Appendix 4

Decentralized biobanking pilot study and breast cancer biobank age distributions, and comparison of pilot enrollment rates by time from initial biobank consent.

[DOCX File , 136 KB - bioinform_v6i1e70463_app4.docx]

Multimedia Appendix 5

Demographics of the breast cancer biobank and decentralized biobanking pilot populations.

[DOCX File , 35 KB - bioinform_v6i1e70463_app5.docx]

Multimedia Appendix 6

Decentralized biobanking pilot participant age ranges and engagement metrics among app onboarded participants by sample collection status, including age, biobank membership, years since initial biobank consent, and research profile completion rates.

[DOCX File , 2726 KB - bioinform_v6i1e70463_app6.docx]

Multimedia Appendix 7

Characteristics of pilot participants who did versus did not complete research profiles and claim biowallet tokens on decentralized biobanking app.

[DOCX File , 15 KB - bioinform_v6i1e70463_app7.docx]

References

1. Kongsholm NC, Christensen ST, Hermann JR, Larsen LA, Minssen T, Pedersen LB, et al. Challenges for the sustainability of university-run biobanks. Biopreserv Biobank 2018 Aug;16(4):312-321. [doi: 10.1089/bio.2018.0054] [Medline: 30016130]
2. Kinkorová J. Biobanks in the era of personalized medicine: objectives, challenges, and innovation: overview. EPMA J 2015;7(1):4 [FREE Full text] [doi: 10.1186/s13167-016-0053-7] [Medline: 26904153]
3. Rush A, Matzke L, Cooper S, Gedye C, Byrne JA, Watson PH. Research perspective on utilizing and valuing tumor biobanks. Biopreserv Biobank 2019 Jun;17(3):219-229. [doi: 10.1089/bio.2018.0099] [Medline: 30575428]
4. Klingstrom T, Bongcam-Rudloff E, Reichel J. Legal and ethical compliance when sharing biospecimen. Brief Funct Genomics 2018 Jan 01;17(1):1-7 [FREE Full text] [doi: 10.1093/bfgp/elx008] [Medline: 28460118]
5. Hallinan D, Friedewald M. Open consent, biobanking and data protection law: can open consent be 'informed' under the forthcoming data protection regulation? Life Sci Soc Policy 2015 Jan 24;11(1):1 [FREE Full text] [doi: 10.1186/s40504-014-0020-9] [Medline: 26085311]
6. Sobel ME, Dreyfus JC, Dillehay McKillip K, Kolarcik C, Muller WA, Scott MJ, et al. Return of individual research results: a guide for biomedical researchers utilizing human biospecimens. Am J Pathol 2020 May;190(5):918-933 [FREE Full text] [doi: 10.1016/j.ajpath.2020.01.014] [Medline: 32201265]
7. Elger BS, De Clercq E. Returning results: let's be honest!. Genet Test Mol Biomarkers 2017 Mar;21(3):134-139 [FREE Full text] [doi: 10.1089/gtmb.2016.0395] [Medline: 28306398]
8. Wolf SM. Return of results in genomic biobank research: ethics matters. Genet Med 2013 Feb;15(2):157-159 [FREE Full text] [doi: 10.1038/gim.2012.162] [Medline: 23386184]
9. Scudellari M. Biobank managers bemoan underuse of collected samples. Nat Med 2013 Mar;19(3):253. [doi: 10.1038/nm0313-253a] [Medline: 23467224]
10. AminoChain homepage. AminoChain. URL: https://aminochain.io/ [accessed 2025-03-27]
11. iSpecimen. URL: https://www.ispecimen.com/ [accessed 2025-03-27]
12. Hasselgren A, Hanssen Rensaa JA, Kralevska K, Gligoroski D, Faxvaag A. Blockchain for increased trust in virtual health care: proof-of-concept study. J Med Internet Res 2021 Jul 30;23(7):e28496 [FREE Full text] [doi: 10.2196/28496] [Medline: 34328437]
13. Velmovitsky PE, Bublitz FM, Fadrique LX, Morita PP. Blockchain applications in health care and public health: increased transparency. JMIR Med Inform 2021 Jun 08;9(6):e20713 [FREE Full text] [doi: 10.2196/20713] [Medline: 34100768]

14. Bayyapu S. Blockchain healthcare: redefining data ownership and trust in the medical ecosystem. Int J Adv Res Eng Technol 2020 Nov;11(11):2748-2755 [FREE Full text]

15. Alshater MM, Nasrallah N, Khoury R, Joshipura M. Deciphering the world of NFTs: a scholarly review of trends, challenges, and opportunities. Electron Commer Res 2024 Jul 30. [doi: 10.1007/s10660-024-09881-y]

16. Gross M, Hood AJ, Sanchez WL. Blockchain technology for ethical data practices: decentralized biobanking pilot study. Am J Bioeth 2023 Nov 25;23(11):60-63. [doi: 10.1080/15265161.2023.2256286] [Medline: 37879029]

17. Mamo N, Martin GM, Desira M, Ellul B, Ebejer JP. Dwarna: a blockchain solution for dynamic consent in biobanking. Eur J Hum Genet 2020 May;28(5):609-626 [FREE Full text] [doi: 10.1038/s41431-019-0560-9] [Medline: 31844175]

18. McGhin T, Choo KR, Liu CZ, He D. Blockchain in healthcare applications: research challenges and opportunities. J Netw Comput Appl 2019 Jun;135:62-75. [doi: 10.1016/j.jnca.2019.02.027]

19. Attaran M. Blockchain technology in healthcare: challenges and opportunities. Int J Healthc Manag 2020 Nov 08;15(1):70-83. [doi: 10.1080/20479700.2020.1843887]

20. Schär F. Decentralized finance: on blockchain- and smart contract-based financial markets. Fed Reserve Bank St. Louis Rev 2021 Apr 15:153-174 [FREE Full text] [doi: 10.20955/r.103.153-74]

21. Chen Y, Bellavitis C. Blockchain disruption and decentralized finance: the rise of decentralized business models. J Bus Ventur Insights 2020 Jun;13:e00151. [doi: 10.1016/j.jbvi.2019.e00151]

22. Sanchez W, Linder L, Miller RC, Hood A, Gross MS. Non-fungible tokens for organoids: decentralized biobanking to empower patients in biospecimen research. Blockchain Healthc Today 2024;7 [FREE Full text] [doi: 10.30953/bhty.v7.303] [Medline: 38715762]

23. Dewan A, Eifler M, Hood A, Sanchez W, Gross M. Building a decentralized biobanking app for research transparency and patient engagement: participatory design study. JMIR Hum Factors 2025 Mar 05;12:e59485 [FREE Full text] [doi: 10.2196/59485] [Medline: 40053747]

24. Human Cancer Models Initiative (HCMI). National Institutes of Health National Cancer Institute Center for Cancer Genomics. URL: https://www.cancer.gov/ccg/research/functional-genomics/hcmi [accessed 2025-03-27]

25. Sanchez W, Dewan A, Budd E, Eifler M, Miller RC, Kahn J, et al. Decentralized biobanking applications empower personalized tracking of biospecimen research: technology feasibility. JMIR Bioinform Biotechnol 2025 Apr 14:70463. [doi: 10.2196/70463]

26. Singh P, Sagar S, Singh S, Alshahrani HM, Getahun M, Soufiene BO. Blockchain-enabled verification of medical records using soul-bound tokens and cloud computing. Sci Rep 2024 Oct 22;14(1):24830. [doi: 10.1038/s41598-024-75708-3] [Medline: 39438519]

27. Gille F, Axler R, Blasimme A. Transparency about governance contributes to biobanks' trustworthiness: call for action. Biopreserv Biobank 2021 Feb 01;19(1):83-85. [doi: 10.1089/bio.2020.0057] [Medline: 33124891]

28. Weil CJ, Nanyonga S, Hermes A, McCarthy A, Gross M, Nansumba H, et al. Experts speak forum: community engagement in research biobanking. Biopreserv Biobank 2024 Oct 01;22(5):535-539. [doi: 10.1089/bio.2024.0131] [Medline: 39431940]

29. Gross MS, Hood AJ, Miller RC. Nonfungible tokens as a blockchain solution to ethical challenges for the secondary use of biospecimens: viewpoint. JMIR Bioinform Biotechnol 2021 Oct 22;2(1):e29905 [FREE Full text] [doi: 10.2196/29905] [Medline: 38943235]

30. Statler M, Wall BM, Richardson JW, Jones RA, Kools S. African American perceptions of participating in health research despite historical mistrust. ANS Adv Nurs Sci 2023;46(1):41-58. [doi: 10.1097/ANS.0000000000000435] [Medline: 35984948]

31. Scharff DP, Mathews KJ, Jackson P, Hoffsuemmer J, Martin E, Edwards D. More than Tuskegee: understanding mistrust about research participation. J Health Care Poor Underserved 2010 Aug;21(3):879-897 [FREE Full text] [doi: 10.1353/hpu.0.0323] [Medline: 20693733]

32. Compton CC. Making economic sense of cancer biospecimen banks. Clin Transl Sci 2009 Jun 29;2(3):172-174 [FREE Full text] [doi: 10.1111/j.1752-8062.2008.00108.x] [Medline: 20443887]

33. HUB Organoids homepage. HUB Organoids. URL: https://www.huborganoids.nl/ [accessed 2025-03-27]

34. American Type Culture Collection homepage. American Type Culture Collection. URL: https://www.atcc.org/ [accessed 2025-03-27]

35. Reihs R, Proynova R, Maqsood S, Ataian M, Lablans M, Quinlan PR, et al. BBMRI-ERIC negotiator: implementing efficient access to biobanks. Biopreserv Biobank 2021 Oct 01;19(5):414-421. [doi: 10.1089/bio.2020.0144] [Medline: 34182766]

36. Ceker D, Baysungur V, Evman S, Kolbas I, Gordebil A, Nalbantoglu S, et al. LUNGBANK: a novel biorepository strategy tailored for comprehensive multi-omics analysis and P-medicine applications in lung cancer. Research Square Preprint posted online on January 24, 2024 [FREE Full text] [doi: 10.21203/rs.3.rs-3816689/v1]

37. Zayas-Cabán T, Okubo TH, Posnack S. Priorities to accelerate workflow automation in health care. J Am Med Inform Assoc 2022 Dec 13;30(1):195-201 [FREE Full text] [doi: 10.1093/jamia/ocac197] [Medline: 36259967]

38. Sidoti O, Dawson W, Gelles-Watnick R, Favereio M, Atske S, Radde K, et al. Mobile fact sheet. Pew Research Center. 2024 Nov 13. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/ [accessed 2025-03-27]

39. Doucet M, Yuille M, Georghiou L, Dagher G. Biobank sustainability: current status and future prospects. J Biorepository Sci Appl Med 2017 Jan;Volume 5:1-7. [doi: 10.2147/bsam.s100899]

XSL•FO

RenderX

40. Odeh H, Miranda L, Rao A, Vaught J, Greenman H, McLean J, et al. The biobank economic modeling tool (BEMT): online financial planning to facilitate biobank sustainability. Biopreserv Biobank 2015 Dec;13(6):421-429 [FREE Full text] [doi: 10.1089/bio.2015.0089] [Medline: 26697911]

41. Simeon-Dubach D, Henderson MK. Sustainability in biobanking. Biopreserv Biobank 2014 Oct;12(5):287-291. [doi: 10.1089/bio.2014.1251] [Medline: 25314050]

42. Racine V. Can blockchain solve the dilemma in the ethics of genomic biobanks? Sci Eng Ethics 2021 Jun 01;27(3):35. [doi: 10.1007/s11948-021-00311-y] [Medline: 34061257]

43. Sabharwal K, Hutler B, Eifler M, Gross M. Decentralized biobanking for transparency, accountability, and engagement in biospecimen donation. J Health Care Law Policy 2025 [FREE Full text]

44. Kasperbauer TJ, Schmidt KK, Thomas A, Perkins SM, Schwartz PH. Incorporating biobank consent into a healthcare setting: challenges for patient understanding. AJOB Empir Bioeth 2021 Dec 04;12(2):113-122 [FREE Full text] [doi: 10.1080/23294515.2020.1851313] [Medline: 33275086]

45. Dewan A, Eifler M, Hood A, Sanchez W, Gross M. Building a decentralized biobanking app for research transparency and patient engagement: participatory design study. JMIR Hum Factors 2025 Mar 05;12:e59485 [FREE Full text] [doi: 10.2196/59485] [Medline: 40053747]

46. El-Gazzar R, Stendal K. Blockchain in health care: hope or hype? J Med Internet Res 2020 Jul 10;22(7):e17199 [FREE Full text] [doi: 10.2196/17199] [Medline: 32673219]

47. Ellis H, Joshi MB, Lynn AJ, Walden A. Consensus-driven development of a terminology for biobanking, the Duke experience. Biopreserv Biobank 2017 Apr;15(2):126-133 [FREE Full text] [doi: 10.1089/bio.2016.0092] [Medline: 28338350]

48. Rogers J, Carolin T, Vaught J, Compton C. Biobankonomics: a taxonomy for evaluating the economic benefits of standardized centralized human biobanking for translational research. J Natl Cancer Inst Monogr 2011;2011(42):32-38. [doi: 10.1093/jncimonographs/lgr010] [Medline: 21672893]

49. Catchpoole D. 'Biohoarding': treasures not seen, stories not told. J Health Serv Res Policy 2016 Apr 05;21(2):140-142. [doi: 10.1177/1355819615599014] [Medline: 26248620]

50. Dewan A, Rubin JC, Gross MS. Informed consensus: the future of respect for persons in biomedical research. Am J Bioeth 2025 (forthcoming). [doi: 10.1080/15265161.2025.2470695]

51. Hiatt RA, Kobetz EN, Paskett ED. Catchment areas, community outreach and engagement revisited: the 2021 guidelines for cancer center support grants from the National Cancer Institute. Cancer Prev Res (Phila) 2022 Jun 02;15(6):349-354. [doi: 10.1158/1940-6207.CAPR-22-0034] [Medline: 35652232]

52. Wilkowska W, Ziefle M. Perception of privacy and security for acceptance of e-health technologies: exploratory analysis for diverse user groups. In: Proceedings of the 5th International Conference on Pervasive Computing Technologies for Healthcare and Workshops. 2011 Presented at: PervasiveHealth 2011; May 23-26, 2011; Dublin, Ireland. [doi: 10.4108/icst.pervasivehealth.2011.246027]

Abbreviations

API: application programming interface
BDRR: Breast Disease Research Repository
IRB: institutional review board
NFT: nonfungible token

XSL•FO

RenderX

Conversational Artificial Intelligence for Integrating Social Determinants, Genomics, and Clinical Data in Precision Medicine: Development and Implementation Study of the AI-HOPE-PM System

Ei-Wen Yang[1]*, PhD; Brigette Waldrup[2], BS; Enrique Velazquez-Villarreal[2,3]*, MPH, MS, MD, PhD

[1]Polyagent, San Francisco, CA, United States

[2]Beckman Research Institute, Duarte, CA, United States

[3]City Of Hope National Medical Center, 1500 E Duarte Road, Duarte, CA, United States

*these authors contributed equally

Corresponding Author:
Enrique Velazquez-Villarreal, MPH, MS, MD, PhD
Beckman Research Institute, Duarte, CA, United States

Abstract

Background: Integrating clinical, genomic, and social determinants of health (SDOH) data is essential for advancing precision medicine and addressing cancer health disparities. However, existing bioinformatics tools often lack the flexibility to perform equity-driven analyses or require significant programming expertise.

Objective: We developed AI-HOPE-PM (Artificial Intelligence Agent for High-Optimization and Precision Medicine in Population Metrics), a conversational artificial intelligence system designed to enable natural language–driven, multidimensional cancer analysis. This study describes the development, implementation, and application of AI-HOPE-PM to support hypothesis testing that integrates genomic, clinical, and SDOH data.

Methods: AI-HOPE-PM leverages large language models and Python-based statistical scripts to convert user-defined natural language queries into executable workflows. It was evaluated using curated colorectal cancer datasets from The Cancer Genome Atlas and cBioPortal, enriched with harmonized SDOH variables. Accuracy of natural language interpretation, run time efficiency, and usability were benchmarked against cBioPortal and UCSC Xena.

Results: AI-HOPE-PM successfully supported case-control stratification, survival modeling, and odds ratio analysis using natural language prompts. In colorectal cancer case studies, the system revealed significant disparities in progression-free survival and treatment access based on financial strain, health care access, food insecurity, and social support, demonstrating the importance of integrating SDOH in cancer research. Benchmark testing showed faster task execution compared to existing platforms, and the system achieved 92.5% accuracy in parsing biomedical queries.

Conclusions: AI-HOPE-PM lowers technical barriers to integrative cancer research by enabling real-time, user-friendly exploration of clinical, genomic, and SDOH data. It expands on prior work by incorporating equity metrics into precision oncology workflows and offers a scalable tool for supporting disparities-focused translational research. Five videos are included as multimedia appendices to demonstrate platform functionality in real-world scenarios.

KEYWORDS

artificial intelligence; LLM; AI agent; bioinformatics; cancer; genomics; precision medicine; social determinants of health; large language model

Introduction

Health care is being transformed by comprehensive precision medicine, which personalizes treatment based on individual differences in genetics, environment, and lifestyle [1,2]. Alongside this shift, there is growing recognition of the critical role social determinants of health (SDOH) play in shaping disease outcomes and access to care [2-5]. To advance both scientific discovery and health equity, integrating clinical, genomic, and SDOH data is imperative for uncovering disease mechanisms, enhancing treatment effectiveness, and reducing disparities—especially among underserved populations. However, several challenges remain: data silos, the need for specialized expertise in multiomics analysis, and the underrepresentation of diverse populations in existing datasets

all continue to hinder the equitable realization of precision medicine [6-9].

The complexity of cancer research workflows demands seamless integration of molecular profiles, clinical metadata, and population-level variables such as race, ethnicity, income, health literacy, and access to care. Although web-based tools like cBioPortal [10] and UALCAN [11] offer structured platforms for querying public cancer datasets such as The Cancer Genome Atlas (TCGA) [12], they operate within predefined analytical frameworks and require users to manually conduct multistep filtering, stratification, and statistical interpretation [13-16]. These limitations restrict the flexibility needed to explore hypothesis-driven, context-specific research questions—especially those involving SDOH variables critical for addressing health equity.

Meanwhile, emerging artificial intelligence (AI)–based tools like CellAgent [17] and AutoBA [18] have begun to explore the potential of large language models (LLMs) in bioinformatics

workflows [19-22]. However, these systems often focus solely on genomic data and lack the capacity to simultaneously integrate clinical and SDOH variables, thereby limiting their utility in advancing equitable biomedical research.

Motivated by these gaps, we introduced AI-HOPE-PM (Artificial Intelligence Agent for High-Optimization and Precision Medicine in Population Metrics), a novel LLM-powered conversational agent designed to democratize access to integrative bioinformatics analysis. AI-HOPE-PM allows users—regardless of technical background—to conduct robust, multidimensional cancer research using natural language queries. As illustrated in Figure 1, the platform employs natural language processing, retrieval-augmented generation, and Python-based bioinformatics pipelines to translate user queries into reproducible and explainable analyses. This includes case–control comparisons, survival modeling, and stratified multiomics analysis—all without requiring code or manual data preprocessing.

Figure 1. Overview of AI-HOPE-PM (Artificial Intelligence Agent for High-Optimization and Precision Medicine in Population Metrics) workflow. LLM: large language model; SDOH: social determinants of health; TCGA: The Cancer Genome Atlas.
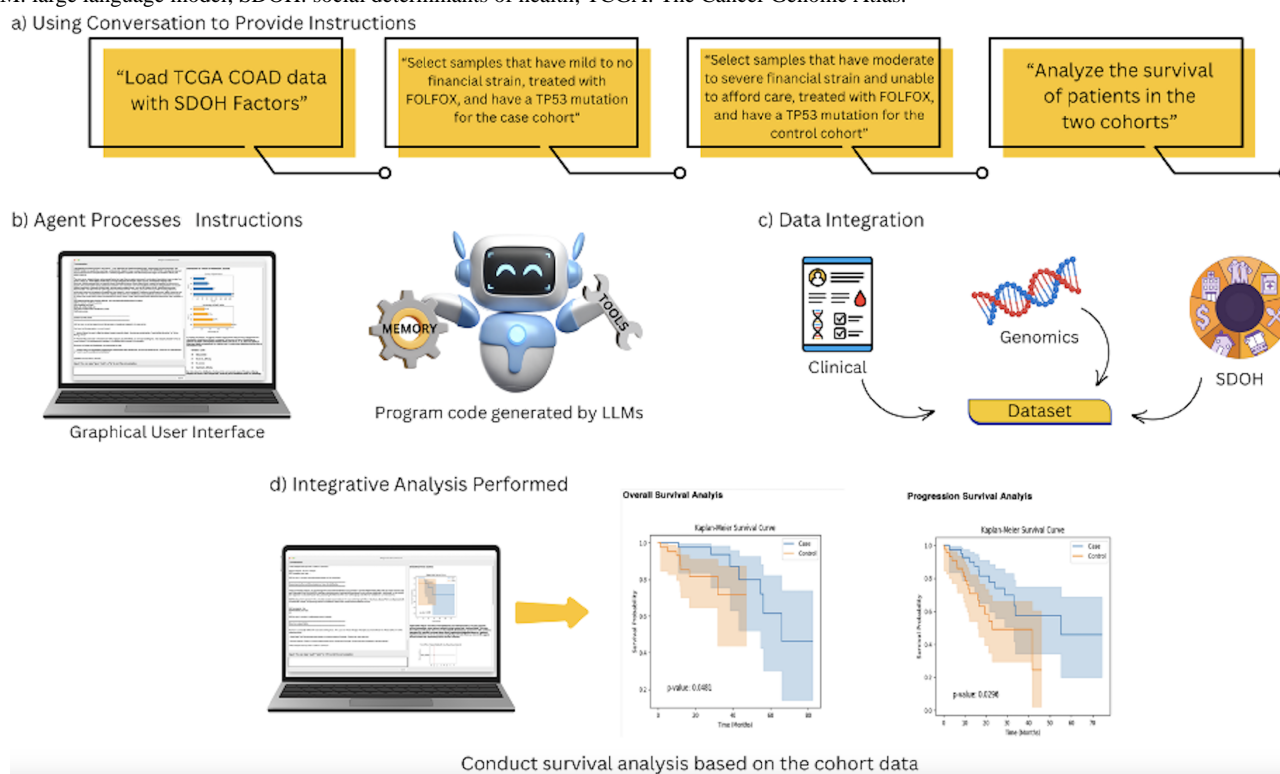
Unlike traditional graphical user interface (GUI) tools, AI-HOPE-PM supports complex, user-defined queries such as "Analyze FOLFOX-treated colorectal cancer (CRC) patients with TP53 mutations and varying levels of financial strain." The system autonomously identifies relevant data, filters patient cohorts, integrates clinical treatment and genomic mutation data with socioeconomic context, and generates statistical visualizations, survival curves, and interpretative summaries. By enabling real-time, dynamic exploration of clinical-genomic-SDOH interactions, AI-HOPE-PM simplifies complex workflows and enhances the translational relevance of precision oncology research. This work builds on our previously developed platform, AI-HOPE [23], a conversational

AI agent designed to support natural language-driven integration of clinical and genomic data for precision medicine research. While AI-HOPE demonstrated effective local analysis of structured datasets and addressed key bioinformatics needs, it did not incorporate SDOH or population-level variables critical to health equity research. AI-HOPE-PM extends this foundation by integrating SDOH data and supporting population-aware case-control analyses, enabling researchers to interrogate disparities across both molecular and social axes. To evaluate its performance, AI-HOPE-PM is being benchmarked against established tools such as cBioPortal and UCSC Xena [24]. The benchmarking involves assessing run time efficiency, reproducibility, and usability. In contrast to tools that require

step-by-step configuration, AI-HOPE-PM offers streamlined execution of advanced bioinformatics pipelines through LLM-guided user interaction, significantly lowering barriers to data exploration and hypothesis testing.

By bridging the gap between data complexity and user accessibility, AI-HOPE-PM offers a scalable, inclusive, and equitable AI framework for biomedical discovery. Its ability to integrate clinical, genomic, and SDOH variables addresses the long-standing need for tools that not only generate high-quality insights but also promote diversity and inclusiveness in biomedical research.

To address the limitations of current bioinformatics tools and advance equity in translational precision medicine, this study introduces AI-HOPE-PM—a novel conversational AI platform purpose-built to integrate clinical, genomic, and SDOH data through natural language interaction. The aim of this paper is to describe the development, implementation, and application of AI-HOPE-PM for multidimensional cancer analysis, with a focus on its ability to democratize data exploration, reduce technical barriers, and enable equity-driven hypothesis testing. Specifically, we demonstrate how AI-HOPE-PM enables real-time, case-control, and survival analyses that incorporate SDOH variables such as financial strain, food insecurity, health care access, and health literacy, alongside genomic and clinical features. By benchmarking its performance and illustrating its use through case studies in CRC, we highlight the platform's potential to accelerate disparities-focused research, improve biomarker discovery, and support inclusive precision oncology.

Methods

Development of AI-HOPE-PM and Data Sources

AI-HOPE-PM is a conversational AI platform designed to advance translational precision oncology by enabling users to perform integrative bioinformatics analyses through plain-language queries. The system is built on a retrieval-augmented generation framework—a method that enhances response accuracy by retrieving relevant information from structured datasets—and a fine-tuned biomedical LLM (LLaMA 3). Behind the scenes, the platform uses Python-based scripts to carry out statistical analyses and genomic data processing.

To enable robust analyses, we used curated multimodal datasets from TCGA, AACR Project GENIE, and cBioPortal. These datasets included harmonized clinical, genomic, and demographic variables. In addition, we generated synthetic SDOH variables using a validated Python script, guided by a literature-informed framework. These SDOH features included health care access, financial strain, food insecurity, social support, and health literacy. All datasets were preprocessed into standardized tab-delimited formats with annotated metadata describing each variable type. A full list of variables analyzed—including over 200 clinical, genomic, treatment, and SDOH fields—is publicly available [25], which also contains the source code, example queries, simulated data, and documentation for reproducing all analyses.

Workflow and Natural Language Interface

Users interact with AI-HOPE-PM via a GUI that accepts plaintext queries. The system interprets these queries using a natural language processing engine to define analytic tasks, including loading a dataset, stratifying cohorts based on genomic or SDOH features, and performing statistical analyses such as survival modeling or odds ratio testing. The resulting structured commands are executed programmatically, streamlining workflows that typically require multiple manual steps or coding expertise.

Evaluation and Validation of System Accuracy

We evaluated AI-HOPE-PM's query interpretation accuracy using 100 natural language prompts that reflected diverse real-world research scenarios in clinical genomics and health disparities. A team of expert reviewers established ground truth interpretations for each query to assess system performance. AI-HOPE-PM achieved an overall accuracy of 92.5%, with near-perfect accuracy (99.1%) for single-variable queries and strong performance (88.4%) for more complex, multivariable prompts. Most errors stemmed from ambiguous phrasing (eg, nonspecific end points), syntactic inconsistencies (eg, nested logic), or misalignment between user language and system variable mappings. To address these issues, AI-HOPE-PM integrates built-in clarification prompts and applies a domain-specific ontology to harmonize terminology and guide users toward more structured input. Future development will focus on improving the natural language understanding engine and refining internal mapping algorithms to further enhance accuracy and reproducibility.

To confirm the analytical fidelity of AI-HOPE-PM, we cross-validated its survival analyses, odds ratio outputs, and cohort stratifications against manually performed analyses previously published by our group using similar datasets and variables. These included studies investigating CRC disparities based on *TP53*, *APC*, and *KRAS* mutation status, treatment modality, and SDOH factors across TCGA and cBioPortal cohorts. The results generated by AI-HOPE-PM were consistent with those from traditional statistical pipelines in terms of hazard ratios, *P* values, and overall survival trends. This validation step supports the platform's accuracy in replicating established findings and reinforces its reliability as a tool for real-time, natural language–driven bioinformatics analyses.

Although we benchmarked AI-HOPE-PM against established platforms such as cBioPortal and UCSC Xena, it is important to acknowledge that these platforms function through traditional GUIs requiring multistep, click-based interactions. This structural difference makes direct comparisons with AI-HOPE-PM—an intelligent, conversational AI system—challenging. In cBioPortal and Xena, executing multilayered queries or stratified analyses may involve multiple browser windows, dropdown menus, and manual dataset subsetting. In contrast, AI-HOPE-PM enables users to perform similar tasks via a single plain-language prompt, streamlining the process and reducing complexity. While speed remains an advantage for AI-HOPE-PM, we also validated its outputs through comparisons with previously published manual analyses, ensuring consistency and analytical fidelity. This intelligent

design is intended to reduce the technical barrier for researchers and support scalable, real-time hypothesis generation in precision medicine Multimedia Appendix 1.

To ensure robustness in handling natural language variability, AI-HOPE-PM incorporates an interactive clarification mechanism that prompts users for additional input when queries are ambiguous or underspecified. Common edge cases include vague end points (eg, "better outcomes"), undefined comparison groups, or syntactic inconsistencies (eg, nested logic). In these instances, the system pauses execution and requests clarification through a structured prompt. Furthermore, AI-HOPE-PM uses a curated biomedical ontology to harmonize synonymous terms and align user inputs with internal variable definitions. These strategies support resilient query interpretation and maintain analytical fidelity across diverse and potentially ambiguous user queries (Multimedia Appendix 2).

Benchmarking and Comparative Analysis

To assess usability and speed, we benchmarked AI-HOPE-PM against existing platforms including cBioPortal and UCSC Xena. Biomedical researchers were asked to complete tasks such as dataset loading, filtering based on genomic or SDOH attributes, and initiating analyses. Task durations were measured using stopwatch protocols. AI-HOPE-PM consistently outperformed traditional tools in terms of execution time and ease of use, owing to its automation and intuitive language-driven interface.

To evaluate the capacity of AI-HOPE-PM to integrate and analyze SDOH alongside clinical and genomic data, we developed a set of simulated SDOH variables. These variables—including financial strain, food insecurity, social support, health literacy, and insurance access—were generated using a Python-based simulation framework informed by published epidemiological distributions and associations relevant to cancer outcomes. The simulation approach was designed to mirror the variability and prevalence observed in real-world populations [26], thereby enabling realistic case–control stratifications and hypothesis testing. Although these SDOH variables are simulated, they serve as a pragmatic proxy in the absence of widely available, high-quality, individual-level SDOH data within public genomic datasets. For full transparency and reproducibility, the simulation scripts are publicly available [25]. Future validation studies using empirical SDOH data from institutional and community-linked datasets are planned to further refine and expand the platform's capabilities.

Statistical Analysis and Report Generation

The platform supports several statistical methods commonly used in cancer genomics, including Kaplan-Meier survival analysis with log-rank testing, Cox proportional hazards regression, and odds ratio calculations for categorical comparisons. Output includes plots such as survival curves and forest plots, accompanied by narrative summaries that describe the findings in context. All outputs are backed by reproducible Python code logs, which are stored internally and can be exported for validation or inclusion in publications [13-15].

Usability Study and Accessibility

A formal usability study is underway to evaluate AI-HOPE-PM's effectiveness and accessibility for biomedical researchers. Participants are comparing its interface, output quality, and query interpretation capabilities with those of GUI-based tools and other AI-driven platforms. While we did not perform head-to-head comparisons with generative systems such as CellAgent or AutoBA due to differing scopes, AI-HOPE-PM's unique ability to integrate SDOH, clinical, and genomic data positions it as a novel tool for equitable and scalable precision medicine research.

To preliminarily assess usability, we conducted a small-scale case study involving six non-bioinformatician users, including oncology fellows and public health researchers. Participants were asked to perform common research tasks using AI-HOPE-PM—such as loading datasets, selecting cohorts by genomic and social variables, and running survival analyses—using only natural language queries. All users completed the tasks successfully, with positive feedback highlighting the intuitive interface, rapid execution, and elimination of the need for coding expertise. These findings provide initial validation of the platform's accessibility to diverse research users.

Ethical and Privacy Considerations

As with any LLM-based system, AI-HOPE-PM is susceptible to biases and potential hallucinations, particularly when interpreting complex or underspecified queries. To mitigate these risks, the system integrates domain-specific ontologies and harmonized variable dictionaries to reduce misinterpretation and support consistent query resolution. Additionally, the platform's built-in clarification prompts serve as a real-time validation mechanism, prompting users to confirm or refine ambiguous instructions. While simulated SDOH-genomic interactions provide a useful testing framework, future efforts will emphasize empirical validation using real-world datasets to reduce confounding.

To address privacy concerns when working with sensitive real-world SDOH variables—such as insurance status, ethnicity, and income—AI-HOPE-PM is designed to operate as a secure, local AI system deployed within institutional infrastructures. Unlike cloud-based models that may transmit data externally, AI-HOPE-PM processes all data on-site, minimizing the risk of exposure or unauthorized access. This local deployment model supports compliance with data protection regulations, including the Health Insurance Portability and Accountability Act and the General Data Protection Regulation, where applicable. In future iterations, we plan to integrate customizable privacy modules and access controls to align with institutional review board protocols and ensure ethical handling of sensitive population-level health data.

To mitigate the risk of hallucinations and enhance the reliability of AI-HOPE-PM's outputs, the platform incorporates several ethical safeguards. First, the system leverages domain-specific biomedical ontologies to align user queries with validated clinical and genomic concepts, reducing the likelihood of misinterpretation. Second, AI-HOPE-PM includes built-in

prompts that clarify ambiguous user input, supporting more accurate query resolution. We also plan to implement human-in-the-loop verification workflows and bias detection modules, which will allow researchers to review, confirm, and flag generated outputs prior to downstream use. These strategies collectively enhance interpretability, accountability, and user trust in the AI-driven analytical process.

Given the sensitivity of SDOH data, especially variables such as insurance status, ethnicity, or socioeconomic conditions, AI-HOPE-PM is currently designed as a locally deployed system to prevent data exposure through external servers or third-party services. This architecture ensures that no identifiable information is shared beyond institutional firewalls. For future deployments that may involve real-world SDOH data, we plan to incorporate privacy-preserving methods including data deidentification, access controls, and secure computation protocols. All future iterations will comply with established data protection regulations such as the Health Insurance Portability and Accountability Act and the General Data Protection Regulation, ensuring responsible and ethical use of sensitive population-level data.

Results

By converting natural language instructions into executable bioinformatics workflows, AI-HOPE-PM enabled seamless integration and analysis of clinical, genomic, and SDOH data within CRC datasets. The platform's ability to interpret user queries and automate complex analyses demonstrated its effectiveness in supporting multidimensional, translational cancer research. Through its intuitive conversational interface, the system dynamically classified patient samples into case and control cohorts based on user-defined criteria. These criteria encompassed gene mutation status, treatment regimens, SDOH attributes, and demographic variables, facilitating highly

customizable stratifications. The system autonomously performed statistical analyses—including prevalence estimation, odds ratio tests, and survival modeling—and generated comprehensive visualizations and interpretable reports.

In a prominent use case, AI-HOPE-PM analyzed data from the TCGA COAD dataset to investigate how financial strain affects outcomes among folinic acid, fluorouracil, and oxaliplatin (FOLFOX)–treated patients with CRC with *TP53* mutations (Figure 2). The analysis began by selecting the COAD dataset enriched with SDOH data, allowing users to explore attribute distributions such as financial strain. A bar chart visualization was generated, showing both the count and percentage distribution of financial strain levels across the dataset (Figure 2A). Based on user-defined filtering criteria—patients treated with FOLFOX and harboring *TP53* mutations—AI-HOPE-PM created two cohorts: a case cohort of 40 (10.9%) patients reporting mild or no financial issues and a control cohort of 43 (11.7%) patients experiencing moderate to severe financial strain, including those unable to afford care. Pie charts illustrated the proportional distribution of these cohorts within the total 366-sample dataset (Figure 2B). Once cohorts were defined, the user selected a survival analysis module. AI-HOPE-PM performed a Kaplan-Meier analysis to assess both overall and progression-free survival. The resulting survival plots demonstrated significantly shorter survival in the control group compared to the case group, with P values of .05 (overall survival) and .03 (progression-free survival), supported by CIs indicating statistical robustness (Figure 2C). These findings underscore AI-HOPE-PM's ability to integrate clinical, genomic, and SDOH data through natural language–guided workflows, enabling rapid identification of clinically meaningful disparities in treatment outcomes and survival. This functionality is further supported by the multimedia demonstration with a similar query [27].

Figure 2. AI-HOPE-PM (Artificial Intelligence Agent for High-Optimization and Precision Medicine in Population Metrics) analysis of folinic acid, fluorouracil, and oxaliplatin–treated patients with colorectal cancer with *TP53* mutations and varying levels of financial strain.
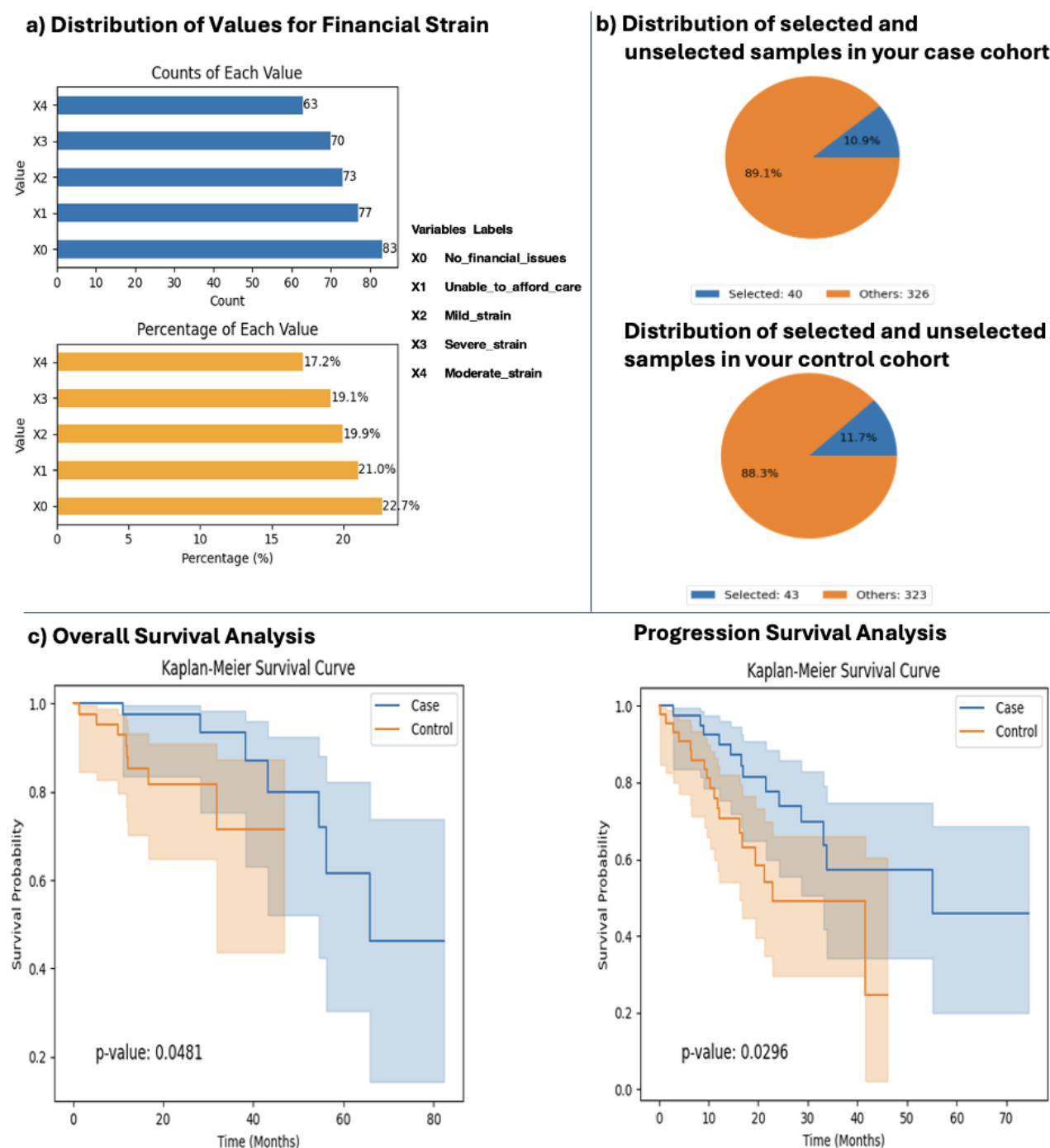
Another case study explored the impact of *APC* mutation status among patients with CRC treated with FOLFOX and reporting easy access to health care (Figure 3). The analysis began by selecting the COAD dataset enriched with SDOH, which enabled exploration of the distribution of health care access variables. A bar chart was generated to visualize both the count and percentage of patients stratified by their reported level of health care access (Figure 3A). AI-HOPE-PM then applied user-defined filters to create case and control cohorts. The case cohort consisted of 40 (10.9%) patients who had *APC* mutations, reported easy access to health care, and received FOLFOX treatment. The control cohort comprised 12 (3.3%) patients who met the same filtering criteria except they were *APC* wild-type.

Pie charts illustrated the proportional distribution of these cohorts out of the total 366 patients in the dataset (Figure 3B). After defining the cohorts, AI-HOPE-PM enabled the user to run a Kaplan-Meier survival analysis, which revealed that patients in the control group (*APC* wild-type) experienced significantly poorer progression-free survival, with a *P* value of .02, as shown in the survival plot (Figure 3C). This suggests a potential prognostic role of *APC* mutation status under standardized treatment and access conditions. Additionally, the system performed an odds ratio analysis to assess differences in ethnic representation between the cohorts. In this context, Hispanic/Latino identity was used as the comparative variable. The case cohort included 6 (15%) in-context Hispanic/Latino

patients and 34 out-of-context patients, while the control cohort included 3 (15%) in-context patients and 9 out-of-context patients. The resulting odds ratio was 0.529 (95% CI 0.11-2.541), indicating a lower—but not statistically significant—representation of Hispanic/Latino individuals in the control group (Figure 3D). Together, these results reinforce

AI-HOPE-PM's ability to integrate genomic, clinical, and SDOH variables and to highlight the importance of considering ancestral background and access to care when evaluating mutation-driven outcomes in precision oncology. As shown through the multimedia demonstration with a similar query [28], the platform effectively processes complex, user-defined inputs.

Figure 3. AI-HOPE-PM (Artificial Intelligence Agent for High-Optimization and Precision Medicine in Population Metrics) analysis of patients with colorectal cancer with and without *APC* mutations that have easy access to health care and treated with folinic acid, fluorouracil, and oxaliplatin.
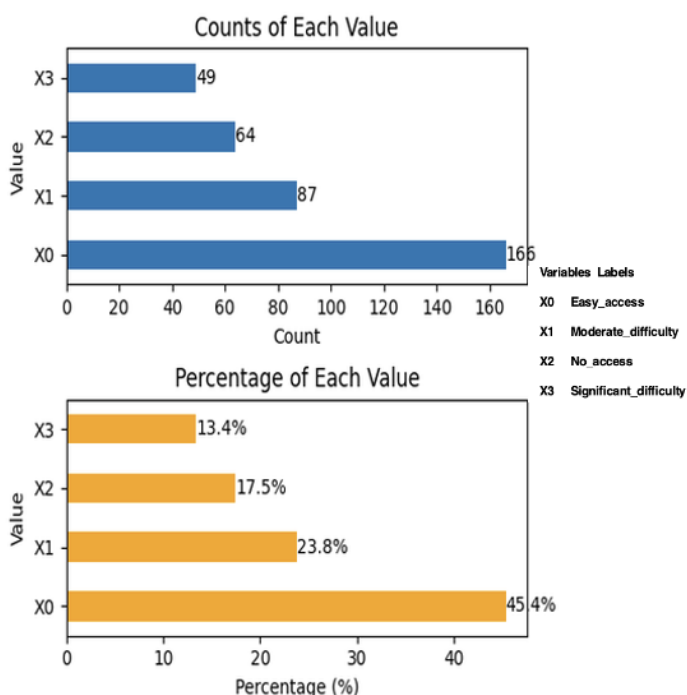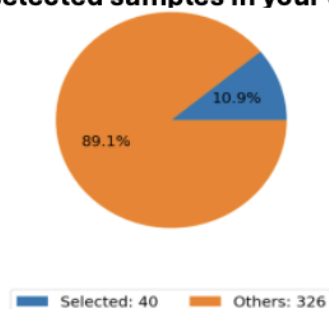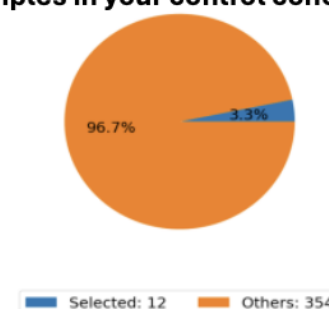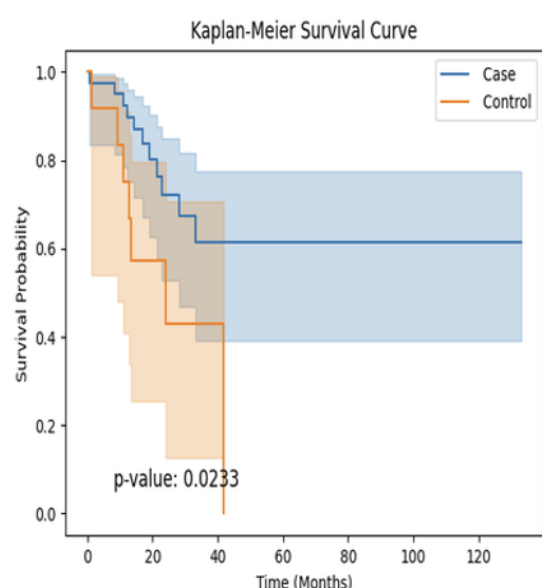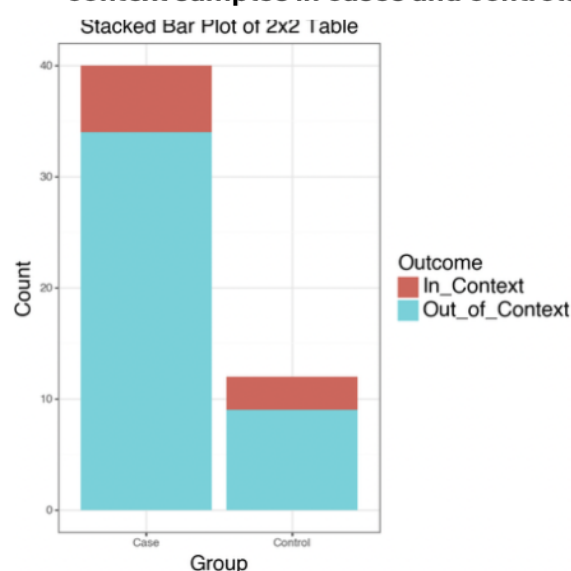
In a third application, AI-HOPE-PM examined patients with early-onset CRC (age <50 y) to evaluate the impact of social support on survival outcomes among those treated with FOLFOX chemotherapy (Multimedia Appendix 1). The user began by selecting the COAD dataset containing enriched SDOH data. This enabled visualization of patient-level attributes

such as age, treatment type, mutation status, and social support level. Histograms and bar plots provided an overview of the distribution of these variables across the cohort (Multimedia Appendix 1A). The case cohort was defined using user-set criteria: patients younger than 50 years, treated with FOLFOX, and classified as having strong or moderate social support,

resulting in 17 patients. This subset was visualized using pie charts to reflect the proportion of selected versus total samples (Multimedia Appendix 1B). In parallel, the control cohort was defined with the same criteria except for social support, selecting 14 patients who reported limited or no support. A similar pie chart depicted the sample distribution for the control group (Multimedia Appendix 1C). A Kaplan-Meier survival analysis was then conducted to assess overall and progression-free survival differences between the two groups. The survival curves revealed a statistically significant difference in progression-free survival (P=.02), with the control group experiencing poorer outcomes. Although the difference in overall survival did not reach statistical significance (P=.07), a trend toward worse survival in the control group was observed (Multimedia Appendix 1D). To further characterize these groups, an odds ratio analysis was performed using *TP53* mutation status as the comparative context. The case group had a lower—but not statistically significant—prevalence of *TP53* mutations, resulting in an odds ratio of 0.706 (95% CI 0.208-2.396; Multimedia Appendix 1E). These findings suggest that lower levels of social support may be linked to poorer progression-free survival and potentially associated with mutational profiles, reinforcing the importance of incorporating psychosocial variables in precision oncology—particularly in younger patients with CRC. The capabilities of AI-HOPE-PM are further demonstrated in the multimedia example of a comparable query [29].

In another analysis focused on food insecurity, AI-HOPE-PM investigated survival disparities and treatment access among patients with CRC with *APC* mutations (Multimedia Appendix 2). The analysis began with the selection of the COAD dataset integrated with SDOH, allowing the user to visualize variables such as food insecurity, treatment type, and *APC* mutation status. Histograms and bar plots summarized the distribution of these attributes across the cohort—highlighting proportions of food-insecure patients and chemotherapy exposure stratified by mutation status (Multimedia Appendix 2A). The case cohort was defined as patients reporting no food insecurity and having an *APC* mutation, yielding 245 samples. A pie chart illustrated the representation of this subset within the overall dataset (Multimedia Appendix 2B). The control cohort was established using the same criteria—*APC* mutation present—but selecting patients with moderate-to-severe food insecurity, resulting in 206 samples. A corresponding pie chart depicted this cohort's proportional distribution (Multimedia Appendix 2C). A Kaplan-Meier progression-free survival analysis was performed, stratifying patients by treatment type, specifically whether they received chemotherapy. Although exact P values were not displayed in the figure, the survival curves showed a clear separation, suggesting poorer outcomes in food-insecure patients not receiving chemotherapy (Multimedia Appendix 2D). These visual trends support the finding that food-insecure patients experienced significantly worse progression-free survival, as confirmed by a P value of .02 from the associated analysis. To further explore this disparity, an odds ratio analysis was conducted using TREATMENT_TYPE (chemotherapy vs nonchemotherapy) as the comparative context. The results revealed that food-insecure patients were less likely to receive chemotherapy, with an odds ratio of 0.356 (95% CI 0.136-1.186), indicating a potential treatment access gap

(Multimedia Appendix 2E). This analysis underscores the ability of AI-HOPE-PM to uncover how socioeconomic burden, in conjunction with genomic context, may modulate both treatment delivery and clinical outcomes in CRC. As illustrated through the multimedia demonstration of a similar query [30], the platform effectively interprets complex natural language inputs.

Sex-based disparities were explored in a separate analysis focusing on patients with CRC with limited health literacy who were treated with FOLFOX chemotherapy (Multimedia Appendix 3). AI-HOPE-PM utilized the COAD dataset enriched with SDOH and genomic annotations to assess the intersection of insurance status, tumor stage, and *KRAS* mutation status. Bar charts provided an overview of insurance coverage within the dataset, illustrating both the absolute counts and proportional distribution across different insurance categories (Multimedia Appendix 3A). To define the cohorts, the case group was filtered to include insured patients who had *KRAS* mutations, were diagnosed at stage I or II, and received leucovorin-based chemotherapy, yielding 31 samples. This subset was visualized using a pie chart to indicate its proportion out of the total 373 samples (Multimedia Appendix 3B). The control cohort applied identical clinical and molecular filters but included only uninsured patients, resulting in 30 samples (Multimedia Appendix 3C). This side-by-side comparison emphasizes how insurance coverage may influence patient stratification and treatment access, even under otherwise uniform clinical conditions. An odds ratio test was performed using *KRAS* mutation status as the defining context to examine mutation prevalence differences between insured and uninsured groups. A stacked bar chart visualized the distribution of in-context (*KRAS*-mutated) versus out-of-context samples in each group. The analysis revealed a modest difference in *KRAS* mutation representation, suggesting that financial access to care could intersect with genomic profiles in ways that warrant deeper investigation (Multimedia Appendix 3D). In a related sex-disparity analysis among patients with limited health literacy, AI-HOPE-PM defined a case cohort of 33 females and a control cohort of 41 males, both treated with FOLFOX. Odds ratio testing using *KRAS* mutation status showed that 30.3% of females and 56.1% of males were *KRAS*-mutated, yielding an odds ratio of 0.503 (95% CI 0.192-1.319; P=.24). Although not statistically significant, these findings suggest potential sex-based differences in *KRAS* mutation prevalence under constrained health literacy conditions and highlight the utility of AI-HOPE-PM for uncovering multidimensional disparities in cancer genomics and treatment. This process is illustrated through the multimedia demonstration of a similar query [31].

AI-HOPE-PM also facilitated analyses of nongenomic SDOH influences on CRC outcomes. In one study, the platform was used to explore how insurance status, treatment exposure, and clinical care setting affected survival among patients in the COAD dataset (Multimedia Appendix 4). The analysis began with the selection of a dataset enriched with SDOH attributes. Bar charts provided a comprehensive overview of insurance type distribution, showing both the absolute number of patients per insurance category and their relative proportions, offering insight into the socioeconomic landscape of the cohort (Multimedia Appendix 4A). Using this context, AI-HOPE-PM

defined a case cohort of 41 insured patients with the following characteristics: stage IV CRC, Hispanic/Latino ethnicity, FOLFOX treatment, and care received at a community oncology practice. A pie chart visualized the size of this cohort relative to the dataset (Multimedia Appendix 4B). A control cohort was generated using the same clinical and demographic criteria but restricted to uninsured patients, resulting in 22 samples. The corresponding pie chart highlighted the discrepancy in sample size and access between the insured and uninsured groups (Multimedia Appendix 4C). Following cohort definition, AI-HOPE-PM performed a Kaplan-Meier survival analysis to evaluate overall survival outcomes. The survival plots illustrated a clear separation between the two groups, with uninsured patients showing poorer survival outcomes, despite receiving similar treatments and having similar disease profiles (Multimedia Appendix 4D). While the figure does not specify *P* values or CIs, the divergence in survival curves strongly suggests a detrimental impact of lack of insurance on patient outcomes. These findings underscore the critical role of insurance coverage in modulating survival, even when controlling for genomic, treatment, and staging variables. As demonstrated in the multimedia example using a comparable query [26], the platform accurately handles complex, user-driven inputs.

This study complements other AI-HOPE-PM findings by leveraging its capacity to integrate SDOH with clinical and genomic data to uncover disparities in CRC care and outcomes. In one analysis, the system examined the relationship between moderate to severe financial strain and CRC screening adherence, revealing that patients experiencing economic hardship were significantly less likely to participate in screening programs, highlighting a critical barrier to early detection (Multimedia Appendix 5). The analysis began with the selection of the SocialFactors_COAD dataset, enabling structured visualization of variables such as *APC* mutation status and health care access levels. Bar plots showed both the frequency of *APC* mutations and the distribution of health care access categories within the full cohort (Multimedia Appendix 5A). A case cohort of 326 patients was created using filters for limited health care access; treatment with agents in fluorouracil, leucovorin, and oxaliplatin; and presence of *APC* mutations (mutation_status=1). A pie chart depicted their proportion relative to the total dataset (Multimedia Appendix 5B). A control cohort of 354 patients was defined using the same criteria except for *APC* wild-type status (mutation_status=0). Their distribution was similarly visualized (Multimedia Appendix 5C). A Kaplan-Meier progression-free survival analysis was then performed, stratified by chemotherapy treatment status and highlighting differences particularly among Hispanic/Latino patients. The survival curves revealed a noticeable separation between groups, suggesting a potential survival disadvantage linked to disparities in health care access and genomic background (Multimedia Appendix 5D). Additionally, an odds ratio analysis evaluated treatment disparities based on chemotherapy exposure across the defined cohorts. A bar plot illustrated differences in chemotherapy receipt, reinforcing how limited access to care and mutation status may jointly influence treatment pathways and clinical outcomes (Multimedia Appendix 5E). Other AI-HOPE-PM analyses supported these observations. One study found that

patients reporting low social support or isolation had higher rates of treatment discontinuation and worse survival outcomes, consistent with psychosocial oncology literature [24]. The platform also uncovered racial and ethnic disparities in progression-free survival, with non-Hispanic White patients demonstrating better outcomes than Black and Hispanic patients, even after adjusting for treatment type and disease stage. Collectively, these results underscore the value of incorporating SDOH variables into precision medicine frameworks, enabling AI-HOPE-PM to reveal systemic inequities that might otherwise be overlooked in genomic-only analyses. The multimedia demonstration of a similar query [29] highlights the platform's ability to interpret and execute complex, user-defined instructions.

AI-HOPE-PM demonstrated high computational efficiency, executing high-dimensional case-control studies involving over 10,000 patient records in under 1 minute. In a benchmark comparison, the platform required only 28.02 seconds to open the application, select a database, and filter a single data attribute—significantly faster than cBioPortal (58.01 s) and UCSC Xena (46.06 s). By automating the ingestion, filtering, analysis, and reporting stages, AI-HOPE-PM substantially reduced manual burden and turnaround time compared to conventional bioinformatics tools. This performance underscores its value as a scalable AI platform capable of delivering real-time, integrative data analysis to support precision oncology and health equity research.

In a comparative timing analysis, AI-HOPE-PM significantly outperformed established platforms such as cBioPortal and UCSC Xena in executing basic data query tasks. The standardized task—which included launching the application, selecting a dataset, and applying a filter based on a single data attribute—was completed in just 28.02 seconds using AI-HOPE-PM. In contrast, the same task required 58.01 seconds on cBioPortal and 46.06 seconds on UCSC Xena. These results underscore the efficiency advantages of AI-HOPE-PM's natural language–driven, automated workflow, which streamlines multistep analyses and reduces manual input time compared to traditional GUI-based platforms.

Discussion

Principal Findings

This study presents the development and application of AI-HOPE-PM, a conversational AI system designed to integrate clinical, genomic, and SDOH data for precision oncology research. AI-HOPE-PM addresses key limitations in existing bioinformatics tools by enabling users to pose natural language queries that are automatically translated into executable workflows. This allows for case-control stratification and hypothesis testing that include both molecular and nonmolecular variables.

In multiple CRC case studies, AI-HOPE-PM demonstrated the ability to reveal associations between genomic alterations (eg, *TP53* and *APC* mutations), treatment exposures (eg, FOLFOX chemotherapy), and SDOH variables such as financial strain, food insecurity, health care access, and social support. These

findings underscore the importance of contextualizing genomic data within broader socioeconomic and behavioral frameworks to better understand cancer disparities and inform population-relevant strategies.

Comparison to Prior Work

Traditional tools such as cBioPortal and UCSC Xena have facilitated broad access to public cancer genomic datasets, yet they require manual, multistep filtering and operate within fixed analytical frameworks. These platforms typically lack support for SDOH integration and require a certain level of technical expertise, limiting their accessibility for noncomputational researchers and clinicians. More recent tools like CellAgent [17] and AutoBA [18] have begun to explore the use of LLMs in biomedical contexts, but their scope is generally limited to genomic analysis and does not extend to the integration of clinical or social variables essential for advancing health equity.

Our group's prior work introduced AI-HOPE, a closed-system, LLM-driven conversational agent designed to enable integrative clinical and genomic data analyses through natural language interactions [23]. AI-HOPE allows users to perform association studies, prevalence assessments, and survival analyses on locally stored datasets while maintaining data security and interpretability. It demonstrated its capabilities by identifying well-documented associations in TCGA CRC datasets, such as the enrichment of *TP53* mutations in late-stage CRC and the association of *KRAS* mutations with poor progression-free survival in FOLFOX-treated patients. While AI-HOPE addressed the integration of clinical and genomic data, it was not explicitly designed to handle population-level equity metrics or SDOH variables.

AI-HOPE-PM builds on and significantly extends this foundation by incorporating SDOH dimensions—such as financial strain, health care access, food insecurity, and health literacy—into its analytical framework. This addition allows researchers to study cancer outcomes in a more holistic context, bridging molecular findings with real-world social environments. Furthermore, AI-HOPE-PM expands the scope of natural language query handling to accommodate multimodal stratification involving genomic, clinical, and social parameters, which is essential for addressing health disparities. By doing so, it complements AI-HOPE's functionality while introducing new capabilities that are critical for equity-focused translational research.

Strengths and Limitations

A key strength of AI-HOPE-PM is its ability to perform integrative, user-defined analyses through natural language queries without requiring programming expertise. This significantly reduces technical barriers for clinician-scientists and public health researchers. Importantly, the platform enables the inclusion of SDOH variables—such as financial strain, health care access, and social support—that are often absent from traditional bioinformatics workflows. Its modular architecture supports rapid cohort definition, survival modeling, and odds ratio testing across large, harmonized datasets, allowing for real-time hypothesis generation and exploratory analysis.

However, several limitations should be acknowledged. First, while this study used harmonized and simulated SDOH variables to demonstrate the platform's functionality, the availability and quality of real-world, longitudinal SDOH data remain limited in many health care systems. This may affect the generalizability of findings and the real-world applicability of the platform. Future efforts will require integration with validated, longitudinal SDOH datasets—potentially through partnerships with clinical institutions and population health data repositories. Second, AI-HOPE-PM's current design is optimized for structured, publicly available datasets such as TCGA, cBioPortal, and AACR GENIE. As such, its adaptability to unstructured clinical data or eHealth records is limited. While this design choice enhances reproducibility and alignment with standardized biomedical ontologies, future work should explore interoperability with clinical informatics platforms and natural language extraction from eHealth records to expand usability in health care settings. Third, this study focused exclusively on CRC datasets. As a result, findings and workflows may not be immediately generalizable to other cancer types without retraining or additional customization of the AI system. Although the architecture is designed to be adaptable, validation on other tumor types and disease areas will be essential for broader adoption. Fourth, while benchmarking analyses demonstrated strong performance compared to tools like cBioPortal and UCSC Xena, formal usability testing and prospective validation in real-world clinical and research environments were not conducted. These are planned as part of future development phases and will be critical for refining the user interface, evaluating human-AI collaboration, and assessing clinical impact. By acknowledging and addressing these limitations, future iterations of AI-HOPE-PM can be improved to better support equitable, scalable, and clinically relevant precision medicine research.

A notable limitation of the current study is the use of simulated SDOH variables rather than real-world data. While these simulated features were generated to reflect established patterns from peer-reviewed literature and public health datasets, they cannot fully replicate the variability, context-dependence, or missingness typical of empirical SDOH data collected in clinical or community settings. This limitation may impact the external validity of some findings and restrict generalizability. To address this, we are actively pursuing collaborations with health systems and community-based data partners to incorporate validated, longitudinal SDOH datasets into future deployments of AI-HOPE-PM. This planned integration will enable more accurate assessment of equity-relevant outcomes and enhance the platform's application in real-world clinical research.

While AI-HOPE-PM achieved a high query interpretation accuracy of 92.5% during internal evaluation, several error modes were identified that merit consideration. The most frequent issues involved ambiguity in natural language input—particularly when users provided imprecise criteria for cohort selection or omitted critical parameters. Additionally, complex nested queries and nonstandard phrasing occasionally led to misinterpretation or partial execution. In a minority of cases, errors stemmed from misalignment between user terminology and the platform's internal ontology, particularly

for less common clinical or SDOH variables. To address these challenges, AI-HOPE-PM integrates clarification prompts that guide users toward more precise query formulation and supports synonym recognition for common variable names. Ongoing improvements include refining the language model's domain specificity and expanding the internal ontology to better accommodate diverse user inputs. These enhancements are essential for improving reproducibility and user experience in real-world settings.

A key limitation of this study is the use of simulated SDOH variables rather than real-world data. While simulation allowed us to prototype and evaluate the functionality of AI-HOPE-PM under controlled conditions, it does not fully capture the complexity, heterogeneity, or potential missingness often present in real clinical and social datasets. To address this limitation, we have developed and released an open-source Python script [25] that transparently outlines our simulation methodology. Additionally, we are actively working on the integration of real-world SDOH data through ongoing projects [32], which is sequencing and characterizing tumors from 500 Hispanic/Latino patients in the Los Angeles catchment area. These datasets will allow us to test AI-HOPE-PM's performance in real clinical environments and refine its capacity to analyze authentic, population-specific SDOH variables in future iterations.

To address this limitation, we acknowledge that the current evaluation of AI-HOPE-PM using 100 natural language queries—while carefully curated by physician-scientists, public health researchers, biostatisticians, and bioinformaticians to reflect real-world clinical and translational scenarios—represents an early validation phase. These queries were intentionally designed to ensure clinical accuracy, relevance, and internal consistency. However, we recognize the importance of expanding evaluation to include a broader and more diverse group of end users across different levels of expertise. Future iterations of AI-HOPE-PM will incorporate structured feedback from clinicians, public health researchers, and community health stakeholders. This participatory approach will help identify diverse interaction patterns, reduce potential biases, and enhance the platform's interpretive capacity over time.

Future Directions

Future development of AI-HOPE-PM will focus on several enhancements. First, expanding support for additional omics layers [32], including spatial biology [33] and single-cell [34-36], could improve the platform's applicability to emerging areas in systems oncology. Integration with federated learning frameworks may also enable secure, institution-specific model updates without compromising patient privacy. Moreover, enhancing the system's ability to handle longitudinal data, including treatment timelines and SDOH trajectories, will be critical for supporting causal inference and policy-relevant research [37-42].

Future iterations of AI-HOPE-PM will prioritize the integration of more inclusive and representative genomic datasets to enhance the platform's utility across diverse patient populations. While the current analyses rely on publicly available sources such as TCGA and cBioPortal—which are known to underrepresent racial and ethnic minorities—there have been

encouraging advances in improving dataset diversity, particularly in CRC cohorts submitted by major US cancer centers. Notably, several ongoing initiatives aim to sequence and characterize tumors from historically underrepresented populations, including Hispanic/Latino patients with CRC [32]. These datasets, once publicly released, will be incorporated into AI-HOPE-PM to improve its generalizability and relevance in addressing cancer health disparities. This aligns with our overarching mission to develop equity-focused precision oncology tools that are responsive to the needs of all communities.

In this study, benchmarking primarily focused on task completion time—measuring the duration to execute standard bioinformatics queries across AI-HOPE-PM, cBioPortal, and UCSC Xena. While AI-HOPE-PM demonstrated superior efficiency due to its natural language automation, we acknowledge that this assessment does not encompass analytical output comparison. Future benchmarking studies will evaluate not only speed but also reproducibility and concordance of statistical results, including survival curves, odds ratios, and subgroup analyses. This expanded evaluation will ensure that AI-HOPE-PM delivers results comparable in accuracy and robustness to established platforms, further supporting its utility for translational cancer research.

A preliminary usability assessment was conducted during an internal pilot deployment involving five clinician-scientists and three public health researchers. Participants were asked to complete common clinical-genomic queries using AI-HOPE-PM and provide structured feedback on system usability, interpretability of outputs, and ease of query formulation. Feedback indicated that users found the natural language interface intuitive and appreciated the automation of statistical analyses without coding. Suggestions for improvement included refining terminology prompts and expanding visualization customization. These insights have been incorporated into the current version of AI-HOPE-PM, and a formal usability study with a larger and more diverse cohort is currently underway to systematically evaluate accessibility, performance, and user satisfaction.

To enhance accessibility and promote broader adoption, particularly in resource-constrained environments, we are actively exploring deployment strategies that reduce local infrastructure requirements. Although the current AI-HOPE-PM system benefits from graphics processing unit acceleration for large-scale genomic analyses, the core functionalities—including query interpretation, basic statistical modeling, and report generation—can be executed on standard central processing unit-based systems. Additionally, we are developing a lightweight web-hosted version of the platform with backend support on scalable cloud infrastructure, enabling institutions with limited computational resources to access AI-HOPE-PM through a browser without the need for specialized hardware. Future iterations will also offer modular processing options that allow users to select compute-intensive features based on available resources.

User-centered evaluations—including usability studies with diverse researchers and clinicians—are planned to better

understand the platform's accessibility and impact in real-world settings. Additionally, collaborations with community-based research initiatives may help validate AI-HOPE-PM's role in addressing health disparities and improving equity in precision medicine.

AI-HOPE-PM was developed with scalability and accessibility in mind, including potential deployment in resource-constrained settings. The system can be installed and executed locally, eliminating the need for high-bandwidth internet or continuous cloud access. While graphics processing unit acceleration can enhance performance for large-scale queries, the platform's modular backend supports central processing units–only configurations for smaller datasets and standard analyses. Ongoing optimization efforts aim to further reduce computational overhead through lightweight LLM variants and model compression techniques. These features support broader adoption across diverse institutional environments, including low-resource clinical and research settings.

A key consideration for the broader adoption of AI-HOPE-PM is the potential for language bias and variability in natural language queries. While the current version of the platform is optimized for English-language input, this may limit accessibility for nonnative English speakers or introduce semantic variability that could affect interpretation. To mitigate this, AI-HOPE-PM employs a domain-specific ontology with synonym recognition and structured clarification prompts that guide users toward standardized, interpretable input. These features reduce the likelihood of misinterpretation and increase the reliability of query processing. Nonetheless, we recognize the importance of supporting diverse linguistic backgrounds in biomedical research. Future iterations of the platform will integrate multilingual capabilities and undergo structured usability evaluations in non–English-speaking populations to ensure equitable utility and minimize language-related inequities in research engagement.

Data Availability

The AI-HOPE-PM (Artificial Intelligence Agent for High-Optimization and Precision Medicine in Population Metrics) platform, along with demonstration datasets, can be accessed on GitHub [25].

Conflicts of Interest

None declared.

Multimedia Appendix 1
AI-HOPE-PM (Artificial Intelligence Agent for High-Optimization and Precision Medicine in Population Metrics) analysis of patients with early-onset colorectal cancer treated with folinic acid, fluorouracil, and oxaliplatin and varying levels of social support.
[DOCX File, 501 KB - bioinform_v6i1e76553_app1.docx]

Multimedia Appendix 2
AI-HOPE-PM (Artificial Intelligence Agent for High-Optimization and Precision Medicine in Population Metrics) analysis of patients with colorectal cancer with and without chemotherapy treatment, food security, and APC mutations.
[DOCX File, 360 KB - bioinform_v6i1e76553_app2.docx]

Multimedia Appendix 3
AI-HOPE-PM (Artificial Intelligence Agent for High-Optimization and Precision Medicine in Population Metrics) analysis of patients with colorectal cancer with KRAS mutations in the context of insurance coverage and tumor stage.
[DOCX File, 328 KB - bioinform_v6i1e76553_app3.docx]

Multimedia Appendix 4
AI-HOPE-PM (Artificial Intelligence Agent for High-Optimization and Precision Medicine in Population Metrics) analysis of survival outcomes in patients with colorectal cancer with different insurance and treatment profiles.
[DOCX File, 352 KB - bioinform_v6i1e76553_app4.docx]

Multimedia Appendix 5
AI-HOPE-PM (Artificial Intelligence Agent for High-Optimization and Precision Medicine in Population Metrics) stratification of patients with colorectal cancer by health care access, *APC* mutation, and ethnicity for survival and treatment disparity analysis.
[DOCX File, 385 KB - bioinform_v6i1e76553_app5.docx]

References

1.   Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med 2015 Feb 26;372(9):793-795. [doi: 10.1056/NEJMp1500523] [Medline: 25635347]

2. de Bono JS, Ashworth A. Translating cancer research into targeted therapeutics. Nature New Biol 2010 Sep 30;467(7315):543-549. [doi: 10.1038/nature09339] [Medline: 20882008]

3. Braveman P, Gottlieb L. The social determinants of health: it's time to consider the causes of the causes. Public Health Rep 2014;129 Suppl 2(Suppl 2):19-31. [doi: 10.1177/00333549141291S206] [Medline: 24385661]

4. Pinheiro LC, Reshetnyak E, Akinyemiju T, Phillips E, Safford MM. Social determinants of health and cancer mortality in the Reasons for Geographic and Racial Differences in Stroke (REGARDS) cohort study. Cancer 2022 Jan 1;128(1):122-130. [doi: 10.1002/cncr.33894] [Medline: 34478162]

5. Adler NE, Stewart J. Preface to the biology of disadvantage: socioeconomic status and health. Ann N Y Acad Sci 2010 Feb;1186(2):1-4. [doi: 10.1111/j.1749-6632.2009.05385.x] [Medline: 20201864]

6. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature New Biol 2016 Oct 13;538(7624):161-164. [doi: 10.1038/538161a] [Medline: 27734877]

7. Manrai AK, Funke BH, Rehm HL, et al. Genetic misdiagnoses and the potential for health disparities. N Engl J Med 2016 Aug 18;375(7):655-665. [doi: 10.1056/NEJMsa1507092] [Medline: 27532831]

8. Hamid JS, Hu P, Roslin NM, Ling V, Greenwood CMT, Beyene J. Data integration in genetics and genomics: methods and challenges. Hum Genomics Proteomics 2009 Jan 12;2009:869093. [doi: 10.4061/2009/869093] [Medline: 20948564]

9. Reif DM, White BC, Moore JH. Integrated analysis of genetic, genomic and proteomic data. Expert Rev Proteomics 2004 Jun;1(1):67-75. [doi: 10.1586/14789450.1.1.67] [Medline: 15966800]

10. Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov 2012 May;2(5):401-404. [doi: 10.1158/2159-8290.CD-12-0095] [Medline: 22588877]

11. Chandrashekar DS, Bashel B, Balasubramanya SAH, et al. UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses. Neoplasia 2017 Aug;19(8):649-658. [doi: 10.1016/j.neo.2017.05.002] [Medline: 28732212]

12. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn) 2015;19(1A):A68-A77. [doi: 10.5114/wo.2014.47136] [Medline: 25691825]

13. Monge C, Waldrup B, Carranza FG, Velazquez-Villarreal E. Molecular heterogeneity in early-onset colorectal cancer: pathway-specific insights in high-risk populations. Cancers (Basel) 2025 Apr 15;17(8):1325. [doi: 10.3390/cancers17081325] [Medline: 40282501]

14. Monge C, Waldrup B, Manjarrez S, Carranza FG, Velazquez-Villarreal E. Detecting PI3K and TP53 pathway disruptions in early-onset colorectal cancer among Hispanic/Latino patients. Cancer Med 2025 Apr;14(7):e70791. [doi: 10.1002/cam4.70791] [Medline: 40165548]

15. Monge C, Waldrup B, Carranza FG, Velazquez-Villarreal E. Ethnicity-specific molecular alterations in MAPK and JAK/STAT pathways in early-onset colorectal cancer. Cancers (Basel) 2025 Mar 25;17(7):1093. [doi: 10.3390/cancers17071093] [Medline: 40227607]

16. Zheng L, Lu J, Kong D, Zhan Y. A gene signature related to programmed cell death to predict immunotherapy response and prognosis in colon adenocarcinoma. PeerJ 2025;13:e18895. [doi: 10.7717/peerj.18895] [Medline: 39950044]

17. Xiao Y, Liu J, Zheng Y, et al. CellAgent: an LLM-driven multi-agent framework for automated single-cell data analysis. bioRxiv. Preprint posted online on May 5, 2020. [doi: 10.1101/2024.05.13.593861]

18. Zhou J, Zhang B, Li G, et al. An AI agent for fully automated multi‐omic analyses. Adv Sci (Weinh) 2024 Nov;11(44). [doi: 10.1002/advs.202407094]

19. Mandelbaum M, Levy-Erez D, Soffer S, Klang E, Levy-Mendelovich S. Artificial intelligence: large language models in pediatrics. What do we know so far? Isr Med Assoc J 2025 Mar;27(3):183-188. [Medline: 40134173]

20. Carlà MM, Gambini G, Giannuzzi F, Boselli F, De Luca L, Rizzo S. Testing the reliability of ChatGPT assistance for surgical choices in challenging glaucoma cases. J Pers Med 2025 Feb 28;15(3):97. [doi: 10.3390/jpm15030097] [Medline: 40137413]

21. Kopka M, von Kalckreuth N, Feufel MA. Accuracy of online symptom assessment applications, large language models, and laypeople for self-triage decisions. NPJ Digit Med 2025 Mar 25;8(1):178. [doi: 10.1038/s41746-025-01566-6] [Medline: 40133390]

22. Ishida S, Sato T, Honma T, Terayama K. Large language models open new way of AI-assisted molecule design for chemists. J Cheminform 2025 Mar 24;17(1):36. [doi: 10.1186/s13321-025-00984-8] [Medline: 40128788]

23. Yang EW, Velazquez-Villarreal E. AI-HOPE: an AI-driven conversational agent for enhanced clinical and genomic data integration in precision medicine research. Bioinformatics 2025 Jul 1;41(7):btaf359. [doi: 10.1093/bioinformatics/btaf359] [Medline: 40577785]

24. Goldman MJ, Craft B, Hastie M, et al. Visualizing and interpreting cancer genomics data via the Xena platform. Nat Biotechnol 2020 Jun;38(6):675-678. [doi: 10.1038/s41587-020-0546-8] [Medline: 32444850]

25. Velazquez Villarreal Lab. AI-HOPE-PM. GitHub. URL: https://github.com/Velazquez-Villarreal-Lab/AI-HOPE-PM [accessed 2024-06-22]

26. Hutchings H, Behinaein P, Enofe N, et al. Association of social determinants with patient-reported outcomes in patients with cancer. Cancers (Basel) 2024 Feb 29;16(5):1015. [doi: 10.3390/cancers16051015] [Medline: 38473374]

27.   YouTube video: Artificial intelligence agent – the Velazquez-Villarreal Lab at City of HOPE - AI-HOPE-PM example 4. YouTube Doctor Enrique Velazquez YouTube page. 2025 Jun 20. URL: https://youtu.be/4ZuD7-L2RuM [accessed 2025-06-22]

28.   YouTube video: Artificial intelligence agent – the Velazquez-Villarreal Lab at City of HOPE - AI-HOPE-PM example 1. Doctor Enrique Velazquez YouTube page. 2025 Jun 20. URL: https://youtu.be/senJauIc47M [accessed 2025-06-22]

29.   YouTube video: Artificial intelligence agent – the Velazquez-Villarreal Lab at City of Hope - AI-HOPE-PM example 3. Doctor Enrique Velazquez YouTube page. 2025 Jun 20. URL: https://youtu.be/wPk6Vh6n5xg [accessed 2025-06-22]

30.   YouTube video: Artificial intelligence agent – the Velazquez-Villarreal Lab at City of HOPE - AI-HOPE-PM example 2. Doctor Enrique Velazquez YouTube page. 2025 Jun 20. URL: https://youtu.be/b6us-TKWgDc [accessed 2025-06-22]

31.   YouTube video: Artificial intelligence agent – the Velazquez-Villarreal Lab at City of HOPE - AI-HOPE-PM example 5. Doctor Enrique Velazquez YouTube page. 2025 Jun 20. URL: https://youtu.be/XNi4bkWE_AI [accessed 2025-06-22]

32.   Waldrup B, Carranza F, Jin Y, et al. Integrative multi-omics profiling of colorectal cancer from a Hispanic/Latino cohort of patients. medRxiv. Preprint posted online on Nov 15, 2024. [doi: 10.1101/2024.11.03.24316599]

33.   Carranza FG, Diaz FC, Ninova M, Velazquez-Villarreal E. Current state and future prospects of spatial biology in colorectal cancer. Front Oncol 2024;14:1513821. [doi: 10.3389/fonc.2024.1513821] [Medline: 39711954]

34.   Velazquez-Villarreal EI, Maheshwari S, Sorenson J, et al. Single-cell sequencing of genomic DNA resolves sub-clonal heterogeneity in a melanoma cell line. Commun Biol 2020 Jun 25;3(1):318. [doi: 10.1038/s42003-020-1044-8] [Medline: 32587328]

35.   Amezquita RA, Lun ATL, Becht E, et al. Orchestrating single-cell analysis with bioconductor. Nat Methods 2020;17(2):137-145. [doi: 10.1038/s41592-019-0700-8] [Medline: 31792435]

36.   Heumos L, Schaar AC, Lance C, et al. Best practices for single-cell analysis across modalities. Nat Rev Genet 2023 Aug;24(8):550-572. [doi: 10.1038/s41576-023-00586-w] [Medline: 37002403]

37.   Yang EW, Waldrup B, Velazquez-Villarreal E. Precision oncology through dialogue: AI-HOPE-RTK-RAS integrates clinical and genomic insights into RTK-RAS alterations in colorectal cancer. Biomedicines 2025 Jul 28;13(8):1835. [doi: 10.3390/biomedicines13081835] [Medline: 40868090]

38.   Yang EW, Waldrup B, Velazquez-Villarreal E. Decoding the JAK-STAT axis in colorectal cancer with AI-HOPE-JAK-STAT: a conversational artificial intelligence approach to clinical-genomic integration. Cancers (Basel) 2025 Jul 17;17(14):2376. [doi: 10.3390/cancers17142376] [Medline: 40723258]

39.   Yang EW, Waldrup B, Velazquez-Villarreal E. AI-HOPE-TP53: a conversational artificial intelligence agent for pathway-centric analysis of TP53-driven molecular alterations in early-onset colorectal cancer. Cancers (Basel) 2025 Aug 31;17(17):2865. [doi: 10.3390/cancers17172865] [Medline: 40940961]

40.   Yang EW, Waldrup B, Velazquez-Villarreal E. AI-HOPE-TGFbeta: a conversational AI agent for integrative clinical and genomic analysis of TGF-β pathway alterations in colorectal cancer to advance precision medicine. AI 2025;6(7):137. [doi: 10.3390/ai6070137]

41.   Yang EW, Waldrup B, Velazquez-Villarreal E. From mutation to prognosis: AI-HOPE-PI3K enables artificial intelligence agent-driven integration of PI3K pathway data in colorectal cancer precision medicine. Int J Mol Sci 2025 Jul 5;26(13):6487. [doi: 10.3390/ijms26136487] [Medline: 40650262]

42.   Yang EW, Waldrup B, Velazquez-Villarreal E. Conversational AI agent for precision oncology: AI-HOPE-WNT integrates clinical and genomic data to investigate WNT pathway dysregulation in colorectal cancer. Front Artif Intell 2025;8:1624797. [doi: 10.3389/frai.2025.1624797] [Medline: 40860720]

Abbreviations

AI: artificial intelligence
AI-HOPE-PM: Artificial Intelligence Agent for High-Optimization and Precision Medicine in Population Metrics
CRC: colorectal cancer
FOLFOX: folinic acid, fluorouracil, and oxaliplatin
GUI: graphical user interface
LLM: large language model
SDOH: social determinants of health
TCGA: The Cancer Genome Atlas

XSL•FO

RenderX

Designing a Finite Element Model to Determine the Different Fixation Positions of Tracheal Catheters in the Oral Cavity for Minimizing the Risk of Oral Mucosal Pressure Injury: Comparison Study

Zhiwei Wang[1,2], MSc; Zhenghui Dong[1], MSc; Xiaoyan He[2], BSc; ZhenZhen Tao[3], MSc; Jinfang QI[1], MSc; Yatian Zhang[4], MSc; Xian Ma[4], MSc

[1]The Sixth Affiliated Hospital of Xinjiang Medical University, No. 39 Wuxing South Road, Tianshan District, Urumqi, Xinjiang, China

[2]Department of Critical Care Medicine, Qinghai Fifth People's Hospital, Xining, China

[3]Emergency Department, The Fourth Affiliated Hospital of Xinjiang Medical University, Xin Jiang, China

[4]School of Nursing, Xinjiang Medical University, Xin Jiang, China

Corresponding Author:
Zhenghui Dong, MSc
The Sixth Affiliated Hospital of Xinjiang Medical University, No. 39 Wuxing South Road, Tianshan District, Urumqi, Xinjiang, China

Abstract

Background: Despite being an important life-saving medical device to ensure smooth breathing in critically ill patients, the tracheal tube causes damage to the oral mucosa of patients during use, which increases not only the pain but also the risk of infection.

Objective: This study aimed to establish finite element models for different fixation positions of tracheal catheters in the oral cavity to identify the optimal fixation position that minimizes the risk of oral mucosal pressure injury.

Methods: Computed tomography data of the head and face from healthy male subjects were selected, and a 3D finite element model was created using Mimics 21 and Geomagic Wrap 2021 software. A pressure sensor was used to measure the actual pressure exerted by the oral soft tissue on the upper and lower lips, as well as the left and right mouth corners of the tracheal catheter. The generated model was imported into Ansys Workbench 22.0 software, where all materials were assigned appropriate values, and boundary conditions were established. Vertical loads of 2.6 N and 3.43 N were applied to the upper and lower lips, while horizontal loads of 1.76 N and 1.82 N were applied to the left and right corners of the mouth, respectively, to observe the stress distribution characteristics of the skin, mucosa, and muscle tissue in four fixation areas.

Results: The mean (SD) equivalent stress and shear stress of the skin and mucosal tissues were the lowest in the left mouth corner (28.42 [0.65] kPa and 6.58 [0.16] kPa, respectively) and progressively increased in the right mouth corner (30.72 [0.98] kPa and 7.05 [0.32] kPa, respectively), upper lip (35.20 [0.99] kPa and 7.70 [0.17] kPa, respectively), and lower lip (41.79 [0.48] kPa and 10.02 [0.44] kPa, respectively; $P<.001$ for both stresses). The equivalent stress and shear stress of the muscle tissue were the lowest in the right mouth angle (34.35 [0.52] kPa and 5.69 [0.29] kPa, respectively) and progressively increased in the left mouth corner (35.64 [1.18] kPa and 5.74 [0.30] kPa, respectively), upper lip (43.17 [0.58] kPa and 8.91 [0.55] kPa, respectively), and lower lip (43.17 [0.58] kPa and 11.96 [0.50] kPa, respectively; $P<.001$ for both stresses). The equivalent stress and shear stress of muscle tissues were significantly greater than those of skin and mucosal tissues in the four fixed positions, and the difference was statistically significant ($P<.05$).

Conclusions: Fixation of the tracheal catheter at the left and right oral corners results in the lowest equivalent and shear stresses, while the lower lip exhibited the highest stresses. We recommend minimizing the contact time and area of the lower lip during tracheal catheter fixation, and to alternately replace the contact area at the left and right oral corners to prevent oral mucosal pressure injuries.

(JMIR Bioinform Biotech 2025;6:e69298) doi:10.2196/69298

XSL•FO
RenderX

Introduction

The primary method of respiratory support for critically ill patients in the intensive care unit (ICU) is oral tube intubation, which ensures airway patency, increases ventilation volume, and enhances lung function. However, the use of oral tube intubation may lead to oral mucosal pressure injury (OMPI) due to excessive or prolonged pressure, friction, and shear forces [1]. OMPI can increase patient pain, elevate the risk of infection, impose a financial burden on health care, increase staff workload, and even result in medical disputes. The incidence of OMPI in patients in the ICU ranges from 2.95% to 49.2%, with different fixation positions and methods of tracheal catheterization influencing its occurrence [2]. While numerous factors contribute to OMPI, including patient-related factors, physiological conditions, the use of specific medications, and nursing-related aspects, there are limited reports addressing the mechanical factors that cause OMPI [3-5]. The International Guidelines for the Clinical Prevention and Treatment of Stress Injuries suggest that finite element models can be employed to evaluate mechanical factors by assessing stress distribution characteristics within tissue structures and predicting the risk of cellular and tissue damage [6].

The purposes of this study were to use the finite element theory contact algorithm to simulate and analyze the compression process of the oral soft tissue when the endotracheal tube is fixed in different fixed positions in the oral cavity, and to explore the stress distribution characteristics of the oral soft tissue under the force of the endotracheal tube. This would help to more realistically and accurately evaluate the actual force on the oral soft tissue structure and to clarify the reasonable fixed position of the endotracheal tube when it is fixed in the oral cavity, so as to prevent the occurrence of OMPI.

Methods

Finite Element Model

A finite element model of the tracheal catheter positioned at various locations within the mouth was established. The selected participant for the head and facial computed tomography scan was a 28-year-old male volunteer with a normal BMI, measuring 175 cm in height and weighing 72 kg. A total of 512 images, each with a thickness of 0.625 mm, were obtained. The DICOM format data were imported into the 3D reconstruction software Mimics (version 21.0; Materialise) and Geomagic Wrap (version 2021; Raindrop) for model fitting and structural segmentation, respectively. A resistive film pressure sensor was employed to measure the actual pressure exerted by the tracheal catheter in different areas of the patient's mouth, with each measurement being repeated 100 times to calculate an average value using the gravitational formula. Subsequently, using the measured pressures from solid models as the input data, the Ansys software (version 22.0; ANSYS) was used to import the optimized model, define material properties, remesh the model, and generate an accurate finite element model to conduct finite element analysis based on the defined elastic modulus, Poisson ratio, boundary conditions, and simulated loads for various tissues (skin mucosa and muscle tissue), as well as the tracheal catheter and bone [7,8]. The properties of each material are shown in Table 1; the skin and mucosa are set as nonlinear materials, and the bones are set as isotropic materials

Table . Material properties of the finite element model.

Material	Modulus of elasticity (Mpa）	Young modulus (Mpa）	Shear modulus (Mpa）	Poisson ratio (%）
Tracheal catheter	3	__a	1500	0.38
Skeleton	13,400	18,000	—	0.25
Muscle	0.045	0.25	—	0.49
Cutaneous mucosa	—	3	2	0.49

[a]not available.

Ethical Considerations

This study was approved by the Ethics Committee of the Sixth Affiliated Hospital of Xinjiang Medical University (approval number: LFYLLSC20220905-01). All procedures in this study are in line with the ethical standards of the Human Experiments Responsible Committee (Institution and State) and the Declaration of Helsinki.

Setting of Boundary Conditions

In this study, four models representing the upper lip, lower lip, left mouth corner, and right mouth corner were established. The fixed support areas of the models were designated as the top and bottom, allowing for rigid support to be simulated through fixed constraints. A sliding friction contact was implemented between the lip and the tracheal tube, with a friction coefficient set at 1 [9]. A bonded connection was established among the skin, mucous membrane, and muscle tissue. The model accounted for the effects of gravity in a vertical downward direction, with a gravitational acceleration of 9.8 m/s².

Measurement Indicators

The equivalent stress and shear stress of the skin mucosa and muscle tissue were measured under different fixed positions of the tracheal catheter within the mouth. The stress distribution characteristics of the pressure injury model were analyzed for the fixed positions of the upper lip, lower lip, left mouth corner, and right mouth corner. The stress measurement for each part was conducted 10 times to obtain an average value.

Statistical Analysis

Statistical analysis was performed using SPSS (version 25.0; IBM Corp). Measurement data were expressed as mean (SD). One-way ANOVA was employed for comparisons between

groups, while the *t* test was used for intragroup comparisons. A *P* value of less than .05 was considered statistically significant.

Results

Model Verification

A finite element model of the tracheal catheter was established with a total of 14,635 nodes and 8267 elements at various fixed positions within the oral cavity. This model included the ilium of the upper and lower jaws, as well as the skin, mucosa, and muscle tissues of the oral cavity. The extreme values and distribution trends of stress at the mouth angle and lower lip were consistent with the findings of Amrani et al [9], indicating the effectiveness of the modeling approach employed in this study.

Equivalent Stress

The equivalent stress of the skin mucosa was the lowest in the left mouth corner, and then progressively increased in the right mouth corner, upper lip, and lower lip. In contrast, the equivalent stress of muscle tissue was the highest in the right mouth corner, followed by the left mouth corner, upper lip, and lower lip. Notably, the equivalent stress of muscle tissue was significantly greater than that of the skin mucosal tissue (*P*<.001; Table 2).

Table . Comparison of equivalent stress results between skin mucosa and muscle tissue (kPa, n=10).

Position	Cutaneous mucosa, mean (SD)	Muscle tissue, mean (SD)	*t* test *(df)*	*P* value	95% CI
Upper lip	35.20 (0.99)	43.59 (0.84)	−20.371 (9)	<.001	−9.252 to −7.522
Lower lip	41.82 (0.92)	48.35 (0.92)	−15.927 (9)	<.001	−7.389 to −5.667
Left mouth corner	28.42 (0.65)	35.64 (1.18)	−16.924 (9)	<.001	−8.118 to −6.325
Right mouth corner	30.72 (0.99)	34.34 (0.38)	−10.789 (9)	<.001	−3.420 to −2.912
F_1-score	430.942	573.406	N/A[a]	N/A	N/A
P value	<.001	<.001	N/A	N/A	N/A

[a]not available.

Shear Stress

The shear stress of the skin mucosal tissue was the lowest in the left mouth corner, and progressively increased in the right mouth corner, upper lip, and lower lip. In contrast, the shear stress of the muscle tissue was the lowest in the right mouth corner, and progressively increased in the left mouth corner, upper lip, and lower lip. At the four fixed positions, the shear stress of the left and right oral muscle tissue was lower than that of the skin mucosa, while the shear stress of the upper and lower lip muscle tissue was higher than that of the skin mucosal tissue (*P*<.005; Table 3)

Table . Comparison of shear stress results between the skin mucosa and muscle tissue (kPa, n=10).

Position	Cutaneous mucosa, mean (SD)	Muscle tissue, mean (SD)	*t* test *(df)*	*P* value	95% CI
Upper lip	7.60 (0.21)	8.91 (0.39)	−8.959 (9)	<.001	−1.613 to −0.998
Lower lip	10.17 (0.16)	11.69 (0.78)	−5.057 (9)	<.001	−2.145 to −0.882
Left mouth corner	6.58 (0.17)	5.79 (0.33)	6.799 (9)	.001	0.543 to 1.030
Right mouth corner	7.45 (0.36)	5.69 (0.29)	11.972 (9)	<.001	1.450 to 2.068
F_1-score	244.363	126.411	N/A[a]	N/A	N/A
P value	<.001	<.001	N/A	N/A	N/A

[a]not available.

Comparison of Equivalent Stress and Shear Stress in the Mucosal Tissue of the Upper and Lower Lips and the Left and Right Mouth Corners

Equivalent stress was found to be lower in the upper lip compared to the lower lip, and the left mouth corner exhibited lower stress than the right mouth corner (*P*<.001; Table 4-5). In terms of shear stress, the upper lip also showed significantly lower values than the lower lip (*P*<.001;Table5), while the left mouth corner had lower shear stress than the right mouth corner (*P*<.001; Table 5).

Table . Comparison of the results of equivalent stress and shear force in the left and right mouth corners (kPa, n=10).

Position	Left side mouth corner, mean (SD)	Right side mouth corner, mean (SD)	t test (df)	P value	95% CI
Equivalent stress	28.42 (0.65)	30.72 (0.99)	−6.160 (9)	<.001	−3.094 to −1.520
Shear stress	6.58 (0.17)	7.45 (0.36)	−6.984 (9)	<.001	−1.125 to −0.605

Table . Comparison of the results of equivalent stress and shear force in the skin mucosal tissue of the upper and lower lip (kPa, n=10).

Position	Upper lip, mean (SD)	Lower lip, mean (SD)	t test (df)	P value	95% CI
Equivalent stress	35.20 (0.99)	41.82 (0.92)	−15.472 (9)	<.001	−7.519 to 5.721
Shear stress	7.60 (0.21)	10.17 (0.16)	−16.769 (9)	<.001	−2.931 to −2.279

Comparison of Equivalent Stress and Shear Stress in the Muscle Tissue of the Upper and Lower Lips and Left and Right Mouth Corners

The equivalent stress was the lower in the upper lip than in the lower lip (*P*<.001), and higher in the left mouth corner than in the right mouth corner (*P*=.004; Table 6). The shear stress was lower in the upper lip than in the lower lip (*P*<.001), and lower in the left mouth angle than in the right mouth angle (*P*=.298; Table 7)

Table . Comparison of equivalent stress and shear force results in the left and right mouth corners (kPa, n=10).

Position	Left side mouth corner, mean (SD)	Right side mouth corner, mean (SD)	t test (df)	P value	95% CI
Equivalent stress	35.64 (1.18)	34.34 (0.38)	3.308 (9)	.004	0.474 to 2.124
Shear stress	5.74 (0.30)	5.69 (0.29)	1.071 (9)	.50	−0.221 to 0.435

Table . Comparison of equivalent stress and shear force results in the muscle tissue of the upper and lower lips (kPa, n=10).

Position	Upper lip, mean (SD)	Lower lip, mean (SD)	t test (df)	P value	95% CI
Equivalent stress	43.59 (0.84)	48.35 (0.92)	−12.115 (9)	<.001	−5.587 to −3.935
Shear stress	8.91 (0.39)	11.69 (0.78)	−12.477 (9)	<.001	−3.561 to −2.545

Stress Distribution Rules of the Four Groups of Models

The equivalent stress range of the skin mucosa and muscle tissue gradually extends from the stress center to the periphery. In this study, the application direction of the forces on the upper and lower lips is vertical, with the maximum peak values of both equivalent stress and shear stress occurring at the stress point and subsequently radiating outward in the vertical direction. Conversely, the forces applied at the left and right mouth corners are horizontal, causing the stress range to spread horizontally, with the highest stress values appearing at the direct contact point between the tracheal catheter and the mucosal tissue. The distribution of shear stress is centered on the soft tissue stress point and encompasses the entire lip, mandibular region, and both sides of the face, resulting in a broader range of stress. The equivalent stress and shear stress at the mouth corners are significantly lower than those at the upper and lower lips.

To explore the underlying reasons, when the tracheal catheter is fixed at the corner of the mouth, it makes contact with the corner, the upper lip, and the lower lip. The pressure, shear force, and friction generated by this contact are dispersed across the three contact surfaces of the mouth and the upper and lower lips. The contact surface between the tracheal tube and the upper and lower lips serves as the primary stress point, leading to greater stress values at the upper and lower lips compared to the corners of the mouth, with the lower lip experiencing the highest stress. The results of the finite element analysis indicate that the stress at the corners of the mouth is lower, followed by that at the upper lip (Figures 1-4).

Figure 1. Mimics21.0 software was used to reconstruct the patient's head, face and oral tissues in 3D with an interval of 0.25 mm, and the contour range of the skin mucosa and muscle tissue was constructed through the thresholds of different tissues.
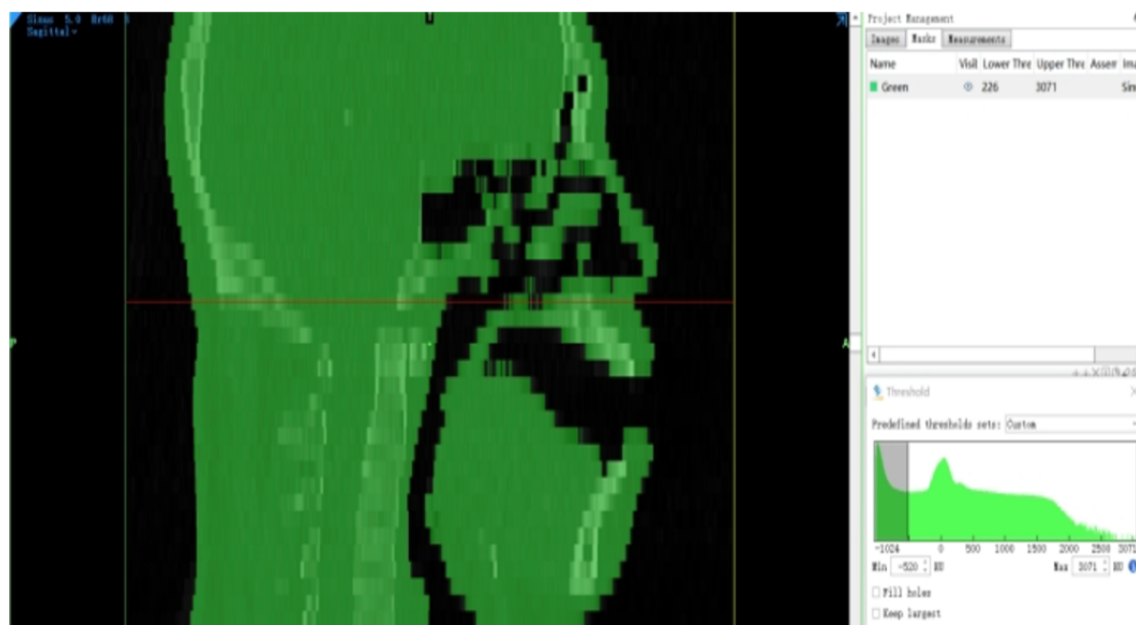
Figure 2. The probe contour line is used to redraw the contour line of the model, so that the surface pieces are more extensible and the concave and convex surfaces are reduced. The structural patch trims the model patch again to make the patch smoother and smoother, which is consistent with the characteristics of the skin tissue. Construct grids, and optimize and adjust all patch nodes and elements. Finally, the fitting surface constructs a model that is similar to the actual oral and facial features of the human body.
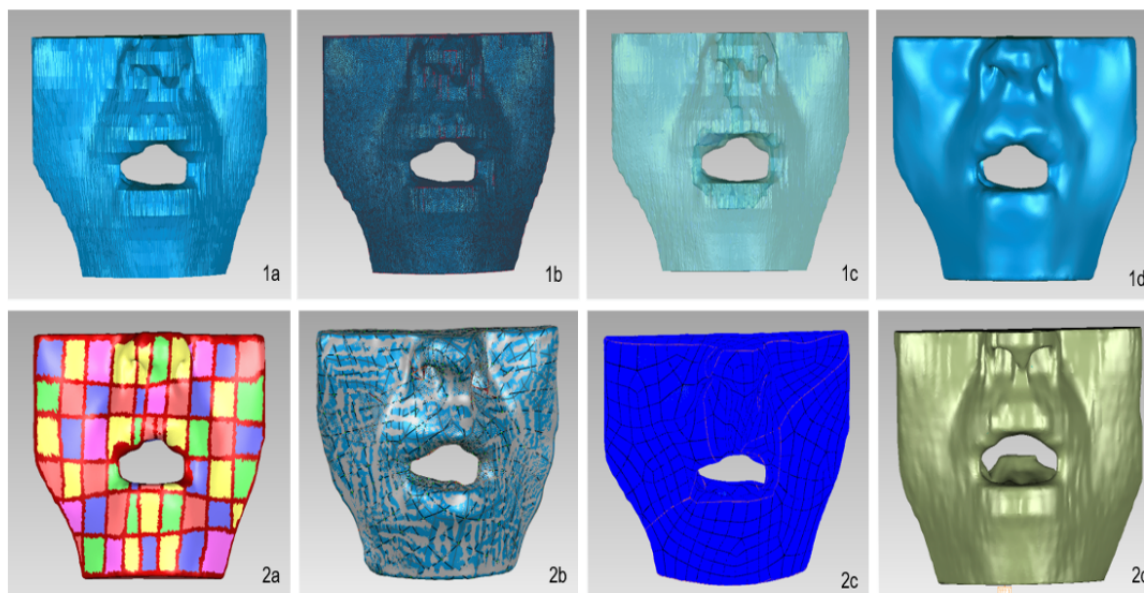
Figure 3. (1a-1d) Equivalent stress nephogram of two tissues at 4 fixed locations: upper lip, lower lip, left mouth corner, and right mouth corner. (2a-2d) Equivalent stress nephogram of the muscle tissue of the upper lip, lower lip, left mouth corner, and right mouth corner.
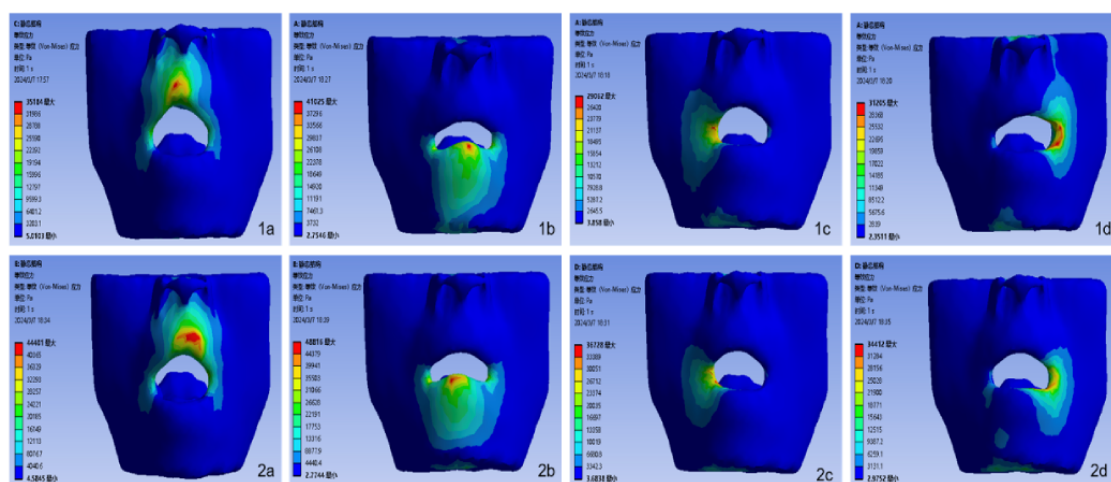
Figure 4. Shear stress nephogram of two tissues at 4 fixed locations. (1a-1d) Shear stress nephogram of the mucosa tissue of the upper lip, lower lip, left mouth corner, and right mouth corner. (2a-2d) Shear stress nephogram of the muscle tissue of the upper lip, lower lip, left mouth corner, and right mouth corner.
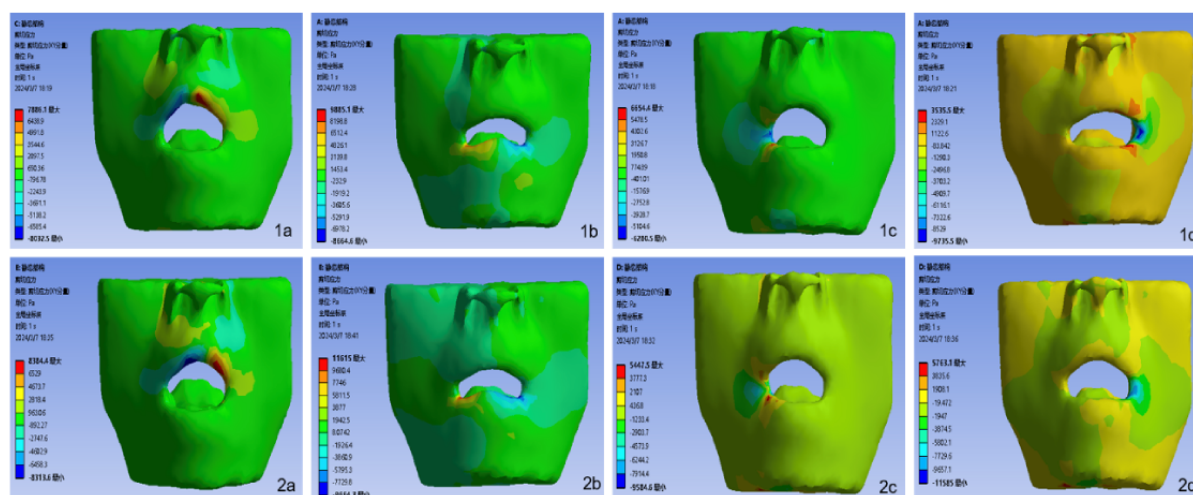
Discussion

The results of this study showed that when the tracheal tube was in contact with the lower lip, the equivalent stress and shear stress values of muscle tissue and mucosal tissue were the largest, followed by the upper lip, and the left and right mouth angles were lower than those of the upper and lower lip. Finite element analysis modeling is a powerful bioengineering technique employed to assess tissue loading, encompassing the interactions between tissues, objects, and medical devices. This numerical method effectively addresses mechanical problems [10]. It enables rapid and accurate stress-strain analysis of the structure, shape, load, and mechanical properties of materials in any given model [11]. Moreover, finite element analysis objectively and accurately reflects the distribution of stress, strain, and deformation, and has gained widespread application in oral biomechanics research in recent years [12].

The tracheal catheter is a critical instrument for mechanical ventilator-assisted therapy in patients in the ICU; however, the catheter itself and improper fixation methods may lead to OMPI [6]. From a biomechanical perspective, the OMPI associated with tracheal catheters primarily results from vertical pressure, shear forces, and friction [13]. Continuous mechanical loading on soft tissues is the main contributor to stress injuries, typically occurring at bony prominences or in areas contacting medical devices. When skin or deep tissue deformation persists for a certain duration owing to the pressure from medical devices, pressure injuries may develop [14]. In this study, the mechanical load originated from the force exerted by the tracheal catheter on the oral soft tissue. Contact between the tracheal catheter and the oral mucosal tissue resulted in continuous pressure, leading to tissue deformation in the mucosa. Research indicates that tracheal catheters and their fixation devices are stiffer than oral soft tissues. When the mechanical properties of these instruments do not align with those of the soft tissues, deformation occurs in the latter, concentrating mechanical stress and strain at the points of direct contact, which then gradually extends to the surrounding areas [15,16].

Continuous vertical pressure on soft tissues is a significant factor in the occurrence of stress injuries. The incidence of OMPI correlates with the intensity and duration of pressure; the greater the pressure and the longer its application, the higher the risk of developing OMPI is [17]. Furthermore, when the tracheal tube is improperly fitted and fixed too tightly, the pressure and shear force exerted will increase [14]. Shear forces applied to deep skin tissues can obstruct capillaries, leading to localized ischemia and hypoxia, which may result in deep tissue necrosis. Consequently, damage from shear forces is often undetected in the early stages and is more challenging to heal than damage from typical wounds [13]. Friction arises from the movement between the oral mucosal tissue and the surface of the tracheal tube; while it does not directly cause OMPI, it can compromise the epidermal cuticle, leading to the shedding of the mucosal surface layer and heightened sensitivity to pressure injuries. Once the compromised oral mucosal tissue is subjected to stimuli from saliva and other secretions, the risk of pressure injury escalates. Additionally, friction raises the temperature of the local mucosal tissue, disrupts the local microenvironment, alters pH levels, and increases tissue oxygen consumption, further exacerbating tissue ischemia and heightening the risk of OMPI [16].

The magnitude of the internal mechanical load required to cause tissue damage depends on the duration of the applied force and the specific biomechanical tolerance of the stressed tissue, which is influenced by factors such as age, shape, health status, and the functional capacity of the body systems, including tissue repair ability [18]. Both high loads applied for short durations and low loads sustained over extended periods can lead to tissue damage [18-20]. Continuous loading is one of the primary contributors to this damage; it refers to loads applied over prolonged periods (ranging from a few minutes to several hours or even days), also known as quasi-static mechanical loading. Research indicates that when soft tissues come into contact with the support surfaces of medical devices, pressure and shear forces are generated between the soft tissues and these surfaces [21]. This interaction results in distortion and deformation of the soft tissues under pressure, affecting both the skin and deeper tissues (including fat, connective tissue, and muscle), leading to stress and strain within the tissues [21]. Excessive internal stress in the tissues can disrupt intracellular material transport by damaging cellular structures (such as the cytoskeleton or plasma membrane) or by hindering the transport process itself (for example, by reducing blood perfusion, impairing lymphatic function, and affecting material transport in the interstitial space), which can ultimately result in cell death and trigger an inflammatory response. Concurrently, the emergence of endothelial cell spacing increases vascular permeability, leading to inflammatory edema, which further exacerbates the mechanical load on cells and tissues due to elevated tissue pressure, thus contributing to the development of pressure injuries [22-24].

According to the results of finite element analysis, the stress experienced by the lower lip is the highest, followed by the upper lip, with levels significantly exceeding those at the corners of the mouth. Therefore, in clinical practice, when fixing a tracheal catheter, it is advisable to select the mouth corner to maximize the contact surface area between the catheter and this region. Placing the tracheal catheter in the middle of the mouth minimizes the contact time between the catheter and the oral mucosa. Additionally, regular changes in the fixation position can help redistribute pressure, thereby reducing pressure, shear forces, and friction on the oral mucosa, ultimately lowering the risk of OMPI.

This study analyzed alterations in the stress experienced by oral soft tissue under pressure at various fixation positions of the tracheal catheter within the mouth, from a biomechanical perspective. It provides a theoretical foundation for preventing OMPI in patients with tracheal catheters in the ICU. While this study effectively simulates the biomechanical effects of contact between oral soft tissue and the tracheal catheter, it does not fully replicate the actual forces experienced by oral soft tissue in real-life situations, as the area of contact between the tracheal catheter and the oral soft tissue cannot be completely simulated. Additionally, the study included only one young adult male, which limits the generalizability of the findings. Therefore, it is essential to include participants of varying genders and ages to enhance the scientific validity of the research. Furthermore, improvements in the identification rate and curvature of the 3D grid of the model should be pursued to generate higher-quality 3D models, thereby enhancing data accuracy.

Data Availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

References

1. Maofan Y, Huilan Z, Keyu C, et al. Research progress on oral mucosal pressure injuries in patients with oral tube intubation in the ICU. Journal of Nursing 2023;38(2):21-24. [doi: 10.3870/j.issn.1001-4152.2023.02.021]
2. Qian L, Lizhu W, Yirong Z, et al. Research progress on oral mucosal pressure injuries in patients with oral tube intubation in the ICU. Chinese Journal of Acute and Critical Care 2023;4(5):473-477. [doi: 10.3761/j.issn.2096-7446.2023.05.019]
3. Li T, Min L, Qian Y. Research progress on nursing care for oral mucosal pressure injuries in patients with oral catheterization. Nurs Res 2023;37(20):3682-3686. [doi: 10.12102/j.issn.1009-6493.2023.20.013]

4.   Choi BK, Kim MS, Kim SH. Risk prediction models for the development of oral-mucosal pressure injuries in intubated patients in intensive care units: A prospective observational study. J Tissue Viability 2020 Nov;29(4):252-257. [doi: 10.1016/j.jtv.2020.06.002]

5.   Tian-kuang L, Yu-Juan L, Huan L, et al. Research progress on the characteristics and nursing care of mucosal pressure injuries in different parts of ICU patients. Journal of Nursing 2022;29(8):35-39. [doi: 10.16460/j.issn1008-9969.2022.08.-035]

6.   Jia L, Deng Y, Xu Y, et al. Development and validation of a nomogram for oral mucosal membrane pressure injuries in ICU patients: A prospective cohort study. J Clin Nurs 2024 Oct;33(10):4112-4123. [doi: 10.1111/jocn.17296] [Medline: 38797947]

7.   Zhang K, Chen Y, Feng C, et al. Machine learning based finite element analysis for personalized prediction of pressure injury risk in patients with spinal cord injury. Comput Methods Programs Biomed 2025 Apr;261:108648. [doi: 10.1016/j.cmpb.2025.108648] [Medline: 39922124]

8.   Keenan BE, Evans SL, Oomens CWJ. A review of foot finite element modelling for pressure ulcer prevention in bedrest: Current perspectives and future recommendations. J Tissue Viability 2022 Feb;31(1):73-83. [doi: 10.1016/j.jtv.2021.06.004]

9.   Amrani G, Gefen A. Which endotracheal tube location minimises the device-related pressure ulcer risk: The centre or a corner of the mouth? Int Wound J 2020 Apr;17(2):268-276. [doi: 10.1111/iwj.13267] [Medline: 31724822]

10. Welch-Phillips A, Gibbons D, Ahern DP, Butler JS. What Is Finite Element Analysis? Clin Spine Surg 2020 Oct;33(8):323-324. [doi: 10.1097/BSD.0000000000001050] [Medline: 32675684]

11. Wang CX, Rong QG, Zhu N, Ma T, Zhang Y, Lin Y. Finite element analysis of stress in oral mucosa and titanium mesh interface. BMC Oral Health 2023 Jan 17;23(1):25. [doi: 10.1186/s12903-022-02703-3] [Medline: 36650512]

12. Guo R, Lam XY, Zhang L, Li W, Lin Y. Biomechanical analysis of miniscrew-assisted molar distalization with clear aligners: a three-dimensional finite element study. Eur J Orthod 2024 Jan 1;46(1):cjad077. [doi: 10.1093/ejo/cjad077] [Medline: 38134411]

13. Zhijun R, Xinhua X, Anqi C, et al. New progress in the prevention of stress injuries induced by mechanical factors. Nurs Res 2017;31(10):1167-1170. [doi: 10.3969/j.issn.1009-6493.2017.10.005]

14. Na W, Yuan-Ting L, Yin-shi X, et al. Summary of evidence for the prevention of medical device-related stress injuries in ICU patients. Chinese Journal of Practical Nursing 2022;38(13):992-997. [doi: 10.3760/cma.j.cn211501-20210710-01863]

15. Gefen A. The aetiology of medical device-related pressure ulcers and how to prevent them. Br J Nurs 2021 Aug 12;30(15):S24-S30. [doi: 10.12968/bjon.2021.30.15.S24] [Medline: 34379465]

16. Mak AFT, Zhang M, Tam EWC. Biomechanics of pressure ulcer in body tissues interacting with external forces during locomotion. Annu Rev Biomed Eng 2010 Aug 15;12:29-53. [doi: 10.1146/annurev-bioeng-070909-105223] [Medline: 20415590]

17. Lustig A, Margi R, Orlov A, Orlova D, Azaria L, Gefen A. The mechanobiology theory of the development of medical device-related pressure ulcers revealed through a cell-scale computational modeling framework. Biomech Model Mechanobiol 2021 Jun;20(3):851-860. [doi: 10.1007/s10237-021-01432-w] [Medline: 33606118]

18. Grigatti A, Gefen A. The biomechanical efficacy of a hydrogel-based dressing in preventing facial medical device-related pressure ulcers. Int Wound J 2022 Aug;19(5):1051-1063. [doi: 10.1111/iwj.13701] [Medline: 34623741]

19. Bogie KM, Zhang GQ, Roggenkamp SK, et al. Individualized Clinical Practice Guidelines for Pressure Injury Management: Development of an Integrated Multi-Modal Biomedical Information Resource. JMIR Res Protoc 2018 Sep 6;7(9):e10871. [doi: 10.2196/10871] [Medline: 30190252]

20. Morrow MM, Hughes LC, Collins DM, Vos-Draper TL. Clinical Remote Monitoring of Individuals With Spinal Cord Injury at Risk for Pressure Injury Recurrence Using mHealth: Protocol for a Pilot, Pragmatic, Hybrid Implementation Trial. JMIR Res Protoc 2024 Apr 10;13:e51849. [doi: 10.2196/51849] [Medline: 38598267]

21. Gawlitta D, Li W, Oomens CWJ, Baaijens FPT, Bader DL, Bouten CVC. The relative contributions of compression and hypoxia to development of muscle tissue damage: an in vitro study. Ann Biomed Eng 2007 Feb;35(2):273-284. [doi: 10.1007/s10439-006-9222-5] [Medline: 17136445]

22. Caulk AW, Chatterjee M, Barr SJ, Contini EM. Mechanobiological considerations in colorectal stapling: Implications for technology development. Surg Open Sci 2023 Jun;13:54-65. [doi: 10.1016/j.sopen.2023.04.004] [Medline: 37159635]

23. Pan Y, Yang D, Zhou M, et al. Advance in topical biomaterials and mechanisms for the intervention of pressure injury. iScience 2023 Jun 16;26(6):106956. [doi: 10.1016/j.isci.2023.106956] [Medline: 37378311]

24. Peko Cohen L, Ovadia-Blechman Z, Hoffer O, Gefen A. Dressings cut to shape alleviate facial tissue loads while using an oxygen mask. Int Wound J 2019 Jun;16(3):813-826. [doi: 10.1111/iwj.13101] [Medline: 30838792]

Abbreviations

ICU: intensivec are unit
OMPI: oral mucosal pressure injury

XSL•FO
RenderX

Protein-Protein Interactions in Papillary and Nonpapillary Urothelial Carcinoma Architectures: Comparative Study

Charissa Chou[1*], BS; Yiğit Baykara[1,2,3*], MD; Sean Hacking[1,2,4], MD; Ali Amin[1,2], MD; Liang Cheng[1,2], MD; Alper Uzun[2,5,6,7], MS, PhD; Ece Dilber Gamsiz Uzun[1,2,6,8,9], MS, PhD

[1]Department of Pathology and Laboratory Medicine, Rhode Island Hospital, 593 Eddy Street, Providence, RI, United States

[2]Department of Pathology and Laboratory Medicine, Alpert Medical School, Brown University, Providence, RI, United States

[3]Department of Pathology, University of Arizona, Tucson, AZ, United States

[4]Department of Pathology, Grossman School of Medicine, New York University, New York, NY, United States

[5]Brown Center for Clinical Cancer Informatics and Data Science, Providence, RI, United States

[6]Legorreta Cancer Center, Brown University, Providence, RI, United States

[7]Department of Pediatrics, Brown University, Providence, RI, United States

[8]Brown Center for Clinical Cancer Informatics and Data Science (CCIDS), Providence, RI, United States

[9]Center for Computational Molecular Biology (CCMB), Brown University, Providence, RI, United States

[*]these authors contributed equally

Corresponding Author:
Ece Dilber Gamsiz Uzun, MS, PhD
Department of Pathology and Laboratory Medicine, Rhode Island Hospital, 593 Eddy Street, Providence, RI, United States

Abstract

Background: Bladder cancer is a disease characterized by complex perturbations in gene networks and is heterogeneous in terms of histology, mutations, and prognosis. Advances in high-throughput sequencing technologies, genome-wide association studies, and bioinformatics methods have revealed greater insights into the pathogenesis of complex diseases. Network biology–based approaches have been used to identify complex protein-protein interactions (PPIs) that can lead to potential drug targets. There is a need to better understand PPIs specific to urothelial carcinoma.

Objective: This study aimed to elucidate PPIs specific to papillary and nonpapillary urothelial carcinoma and identify the most connected or "hub" proteins, as these are potential drug targets.

Methods: A novel PPI analysis tool, Proteinarium, was used to analyze RNA sequencing data from 132 patients with papillary and 270 patients with nonpapillary urothelial carcinoma from the TCGA Cell 2017 dataset and 39 patients with papillary and 88 patients with nonpapillary urothelial carcinoma from the TCGA Nature 2014 dataset. Hub proteins were identified in distinct PPI networks specific to papillary and nonpapillary urothelial carcinoma. Statistical significance of clusters was assessed using the Fisher exact test ($P<.001$), and network separation was quantified using the interactome-based separation score.

Results: RPS27A, UBA52, and VAMP8 were the most connected or "hub" proteins identified in the network specific to the papillary urothelial carcinoma. In the network specific to the nonpapillary carcinoma, GNB1, RHOA, UBC, and FPR2 were found to be the hub proteins. Notably, GNB1 and FPR2 were among the proteins that have existing drugs targeting them.

Conclusions: We identified distinct PPI networks and the hub proteins specific to papillary and nonpapillary urothelial carcinomas. However, these findings are limited by the use of transcriptomic data and require experimental validation to confirm the functional relevance of the identified targets.

Introduction

According to the National Cancer Institute Surveillance, Epidemiology, and End Results program report, the estimated number of deaths from bladder cancer in the United States in 2025 was 17,420. Urothelial carcinoma contributes to more than 90% of bladder cancers and is more prevalent in men than in women [1]. Urothelial carcinomas can be divided into papillary and nonpapillary based on their architecture. Papillary urothelial carcinomas grow in slender projections with fibrovascular cores. Although most do not penetrate the deeper layers of the bladder (noninvasive papillary carcinoma), others can become invasive. Nonpapillary urothelial carcinomas do

not grow toward the hollow center of the bladder and can be invasive or noninvasive (flat carcinoma in situ) [2]. Invasive nonpapillary urothelial carcinomas can aggressively infiltrate the bladder wall, leading to metastasis in lymph nodes and invade adjacent surrounding anatomical structures, such as the prostatic stroma, seminal vesicles, uterus, vagina, pelvic wall, and the abdominal wall. Previous studies have used gene expression profiling and whole-genome comparative genomic hybridization to provide evidence that papillary and nonpapillary urothelial carcinomas have distinct molecular origins and characteristics [3].

Network biology is an emerging paradigm that aims to understand the complex interactions between molecules at the cellular level [4]. Experimental studies describing biological networks demonstrate that they are not random and are characterized by well-known organizing principles. With the advancement of high-throughput screening methods, protein-protein interaction (PPI) networks have been discovered [5]. In addition, large-scale PPI maps showed that proteins involved in specific phenotypes often interact physically. Moreover, it was also shown that similar phenotypes share neighboring network environments not only in model organisms but also in humans [6-9]. Network biology can be used to identify molecular biomarkers for cancer identification and progression. At the proteomic level, PPI network analysis is used for the discovery of novel biomarkers and disease-related functional modules. Several studies based on transcriptomic or microarray data for bladder cancer have identified PPI networks based on differentially expressed genes or coexpressed genes in clusters [10-13]. Most of the time, differentially expressed genes between patients and healthy participants or among patients with different tumor stages are used as an input for building PPI networks. Despite the advances in PPI analysis algorithms, none of the current network analysis tools are designed to identify clusters of patients based on their PPI networks by using the data from individual patient samples. In order to address this, we developed a novel PPI analysis and visualization tool, Proteinarium [14]. Proteinarium is a multisample PPI tool that identifies clusters of samples with shared networks [14]. The Proteinarium tool builds PPI networks from genomic data for individual samples, measures network similarity between samples using the Jaccard distance, clusters them based on these similarities, and identifies shared PPI networks associated with a disease phenotype. In this study, we aimed to identify distinct PPI networks as well as hub proteins specific to papillary and nonpapillary urothelial carcinomas using the Proteinarium tool, leading to potential drug targets for those 2 architectures.

Methods

Datasets

We studied 2 datasets, including clinical and genomic data from patients with urothelial carcinoma. The datasets including The Cancer Genome Atlas (TCGA) Cell 2017 [15] and TCGA Nature 2014 [16] were obtained from the cBioPortal for Cancer Genomics [17,18]. Only patients with mRNA z-score data were included in the analysis. Patient groups were identified based on urothelial carcinoma architecture (papillary or nonpapillary), sex, patient status (alive, recurrent, or deceased), and cancer stage (stage I, II, III, and IV) using custom-made R scripts (R version 4.0.4). The TCGA Cell 2017 dataset [15] included 133 patients with papillary urothelial carcinoma and 273 patients with nonpapillary urothelial carcinoma. Of these patients, 132 with papillary and 270 with nonpapillary carcinomas had RNA-seq data and were included in the PPI network analysis. In the TCGA Nature 2014 dataset [16], there were 39 patients with papillary urothelial carcinoma and 88 patients with nonpapillary urothelial carcinoma with RNA-seq data. Histologic variants of urothelial carcinoma were not included in either dataset. For both studies, the data were downloaded as gene expression z scores.

In the phenotypic data analysis, sex, age, follow-up time, overall survival, disease-free time, pathological distant metastasis stage, survival status, disease-free status, histologic grade, pathological primary tumor stage, pathological regional lymph node stage, and stage grouping data were included in the analysis of the TCGA Cell 2017 dataset. For the TCGA Nature 2014 dataset, patient data such as sex, age, survival status, overall survival, disease-free status, and disease-free time were included. A total of 346 and 129 patients with urothelial carcinoma were analyzed in the TCGA Cell 2017 and TCGA Nature 2014 datasets, respectively. Patients without phenotypic data were excluded from the analysis. Thus, there were differences in the number of patients in the PPI network analysis and phenotypic data analysis.

Proteinarium Analysis and Network Construction

We used the Proteinarium tool to identify shared PPI networks in patients with papillary and nonpapillary urothelial carcinoma using the RNA-seq data from the TCGA Cell 2017 and TCGA Nature 2014 datasets. For each patient, we ranked the genes based on their z scores to identify the upregulated genes. To determine the number of seed genes used to build the PPI networks of each patient, we applied cophenetic correlation scores to the dendrogram output from the Proteinarium tool comparison of patients with papillary and nonpapillary urothelial carcinomas using the top 50, 100, 125, 150, 200, 250, and 300 seed genes. For papillary versus nonpapillary comparisons in both the TCGA Cell 2017 and TCGA Nature 2014 datasets, the cophenetic correlation coefficients remained stable after increasing the number of seed genes to 125, indicating that the overall dendrogram structure was not substantially altered by the larger number of seed genes (Figures S1-S3 in Multimedia Appendix 1). Therefore, we used the top 125 most upregulated seed genes for each patient as inputs for the Proteinarium tool. We performed the same cophenetic correlation score analysis for the downregulated genes in the TCGA Cell 2017 dataset (Figure S4 in Multimedia Appendix 1). Downregulated genes were chosen by ranking the genes from lowest to highest z score, then choosing the top N number of seed genes.

As a first step, the Proteinarium tool converts gene names to their protein names and uses PPI information from the STRING Database [19]. It builds the network graphs based on the seed genes for each patient. It calculates the similarity between each pair of graphs by using the Jaccard distance and records it in a

matrix. Proteinarium uses this similarity matrix as an input for clustering graphs. The unweighted pair group method with arithmetic mean was used for clustering. Patients were clustered based on their network similarities as an output. The significance of clusters was assessed using the Fisher exact test, based on the relative abundance of cases and controls within each cluster. Networks of the significant clusters can be visualized and analyzed. In the Proteinarium analysis, we used the default parameters of the tool for building the graphs, and the maximum path length parameter was set to 2. By applying this, we included only the seed proteins connected directly to each other or via a single intermediary protein referred to as an imputed protein.

The analysis was performed using the following configuration settings: maximum number of vertices to render: 50, meta cluster threshold: 0.8, number of bootstrapping rounds: 0, maximum path length: 2, and repulsion constant: 1.2. Proteinarium was executed using the following command: java -jar <path to Proteinarium.jar> <arguments>, where <arguments> includes user-defined input files and options consistent with the parameters described earlier. A reproducibility package is provided in Multimedia Appendix 2, which includes (1) per-patient seed gene lists, (2) the exact Proteinarium command and configuration file, (3) the final network edge lists, and (4) the README file. The Proteinarium repository is available on GitHub [14].

The primary aim of the Proteinarium analysis was to identify hub and unique proteins in the PPI networks specific to papillary and nonpapillary urothelial carcinoma. Hub proteins were further examined using a drug repurposing tool to identify potential therapeutics, as targeting these highly connected nodes may disrupt critical interactions within the network. Unique proteins were evaluated through enrichment analysis to determine biological pathways significantly associated with each network.

Network Annotation and Modularity

Networks of the significant clusters were annotated by g:Profiler, a web server for identifying enriched biological categories [20]. The unique proteins specific to the papillary and nonpapillary PPI networks were analyzed using g:Profiler to identify enriched gene ontology terms and the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. Only those with a false discovery rate less than 0.05 were included. GEPHI version 0.9.2, a network visualization tool, was used to visualize the networks and their network-specific modularity [21]. We used the modularity function that was based on the Blondel algorithm on GEPHI [22]. Modularity is defined as a measure of network structure that identifies the strength of division of a network into modules.

Separation Test

Separation testing is a computational method to determine the genetic similarity between 2 groups by comparing their PPI networks in the interactome. We performed separation tests to determine the distance between the networks specific to patients with papillary and nonpapillary urothelial carcinomas in the interactome using a Python script adapted from Menche et al [23]. The interactome is composed of 141,296 experimentally determined physical interactions between 13,460 proteins. s_{AB} is a measure of separation between the networks of 2 groups in the interactome. $s_{AB}>0$ indicates a topological separation of the networks specific to 2 groups, and those groups can be regarded as distinct molecular entities. $s_{AB}<0$ indicates that the networks specific to 2 groups overlap in the interactome.

Connectivity Map and Drug Repurposing Tool

We used the connectivity map (CMap) from the Broad Institute and applied the repurposing tool to explore potential drugs that may target the hub genes of networks A, B, and C for CMapap [24]. CMap-L1000 Build 03 was queried with the genes of interest from networks A, B, and C; compounds with $|\tau| >90$ were considered hits.

Statistical Analysis

The chi-square or Fisher exact test were used for categorical variables when appropriate. Paired t tests (2-sided) were used to compare the means of pairs of groups. All statistical analyses were carried out using JMP Pro (version 16.0; SAS Institute). P values <.05 were considered to be statistically significant.

Ethical Considerations

In this study, we used publicly available, deidentified datasets obtained from cBioPortal. Because all data were fully anonymized and accessible to the public, institutional review board approval was not required.

Results

Clinicopathologic Characteristics of the Study Cohorts

Of all the 346 patients with urothelial carcinoma in the TCGA Cell 2017 dataset [15], no significant differences were found between patients with papillary and nonpapillary urothelial carcinoma regarding sex, age, follow-up time, overall survival, disease-free time, and pathological distant metastasis stage. However, statistically significant differences were observed in survival status, disease-free status, histologic grade, pathological primary tumor stage, pathological regional lymph node stage, and stage grouping (Table 1). In 129 patients with urothelial carcinoma in the TCGA Nature 2014 dataset [16], no significant differences were found between the patients with papillary and nonpapillary carcinoma regarding age, overall survival, disease-free status, and disease-free time. However, statistically significant differences were observed in terms of sex and survival status (Table 2).

Table . Clinical parameters of urothelial carcinomas in the Cancer Genome Atlas Cell 2017 dataset.

Parameter	Patients with papillary carcinoma	Patients with nonpapillary carcinoma	P value
Sex, n/n (%)			.13
Female	27/133 (20.3%)	60/213 (28.2)	
Male	106/133 (79.7%)	153/213 (71.8)	
Age (y), mean (SEM)	66.9 (1.0)	68.4 (0.7)	.21
Follow-up time[a] (d), mean (SEM)	521.8 (69.8)	652.5 (72.5)	.21
Survival status, n/n (%)			<.001
Living	90/133 (67.7)	104/213 (48.8)	
Deceased	43/133 (32.3)	109/213 (51.2)	
Overall survival[b] (mo), mean (SEM)	24.4 (2.0)	27.1 (1.9)	.34
Disease-free status[c], n/n (%)			<.001
Disease-free	75/108 (69.4)	79/167 (47.3)	
Recurred or progressed	33/108 (30.6)	88/167 (52.7)	
Disease-free time[d] (mo), mean (SEM)	22.8 (2.2)	26.8 (2.3)	.23
Histologic grade[e], n/n (%)			<.001
Low-grade	18/131 (13.7)	1/213 (0.5)	
High-grade	113/131(86.3)	212/213 (99.5)	
Pathological primary tumor stage, n/n (%)[f]			<.001
T1 or T2	56/118 (47.5)	53/199 (26.6)	
T3	41/118 (34.7)	117/199 (58.8)	
T4	21/118 (17.8)	29/199 (14.6)	
Pathological regional lymph node; stage, n/n (%)			.03
N0	89/116 (76.7)	117/187 (62.6)	
N1	8/116 (6.9)	27/187 (14.4)	
N2	19/116 (16.4)	43/187 (23.0)	
Pathological distant metastasis stage, n/n (%)			>.99
M0	75/79 (94.9)	89/93 (95.7)	
M1	4/79 (5.1)	4/93 (4.3)	
Stage grouping[g], n/n (%)			<.001
Stage I[h] or II	61/129 (47.2)	53/213 (24.9)	
Stage III	34/129 (26.4)	84/213 (39.4)	
Stage IV	34/129 (26.4)	76/213 (35.7)	

[a]Follow-up time was not available for 26 patients with papillary urothelial carcinoma and 68 patients with nonpapillary urothelial carcinoma; t test was performed for the available data in patients with papillary urothelial carcinoma (n=107) and patients with nonpapillary urothelial carcinoma (n=145).

[b]Overall survival was not available for 3 patients in patients with papillary urothelial carcinoma; a t test was performed for the available data in patients with papillary urothelial carcinoma (n=130) and nonpapillary group (n=213).

[c]Disease-free status was not available for 25 patients in the papillary and 46 patients in the nonpapillary group; a chi-square test was performed for the available data in patients with papillary urothelial carcinoma (n=108) and patients with nonpapillary urothelial carcinoma (n=167).

[d]Disease-free time was not available for 26 patients with papillary and 46 patients with nonpapillary carcinoma; a t test was performed for the available data in patients with papillary carcinoma (n=107) and patients with nonpapillary carcinoma (n=167).

[e]Histologic grade was not available for 2 patients with papillary carcinoma; a chi-square test was performed for the available data in patients with papillary urothelial carcinoma (n=131) and patients with nonpapillary carcinoma (n=213).

[f]There were 1 T1 stage patient, 15 Tx patients, 12 Nx patients, and 54 Mx patients among patients with papillary urothelial carcinoma and 1 T1 patient,

14 Tx patients, 24 Nx patients, and 120 Mx patients among patients with nonpapillary carcinoma.

[g]Stage information was unavailable for 2 patients in the papillary group; a chi-square analysis was conducted using the available data (papillary: n=131, nonpapillary: n=213).

[h]There was only 1 patient with a T1 stage tumor among patients with papillary urothelial carcinoma and 1 patient with T1 stage tumor among patients with nonpapillary urothelial carcinoma.

Table . Clinical parameters of urothelial carcinomas in the Cancer Genome Atlas Nature 2014 dataset.

Parameter	Patients with papillary urothelial carcinoma (n=41)	Patients with nonpapillary urothelial carcinoma (n=88)	P value
Sex, n (%)			.03
Female	5 (12.2)	27 (30.7)	
Male	36 (87.8)	61 (69.3)	
Age (y), mean (SEM)	67.5 (1.8)	69.0 (1.0)	.45
Survival status			.006
Living, n (%)	33 (80.5)	49 (55.7)	
Deceased, n (%)	8 (19.5)	39 (44.3)	
Overall survival (mo), mean (SEM)	14.6 (3.3)	19 (2.7)	.33
Disease-free status[a], n (%)			.15
Disease-free	20 (62.5)	37 (46.2)	
Recurred or progressed	12 (37.5)	43 (53.8)	
Disease-free time[b] (mo), mean (SEM)	9.0 (1.8)	11.0 (1.5)	.49

[a]Disease-free status was not available for 9 patients with papillary and 8 patients with nonpapillary carcinoma; chi-square test was performed for the available data in patients with papillary urothelial carcinoma (n=32) and nonpapillary group (n=80).

[b]Disease-free time was not available for 30 patients in papillary and 54 patients in nonpapillary group; t test was performed for the available data in patients with papillary urothelial carcinoma (n=11) and nonpapillary group (n=34).

Proteinarium Identifies Distinct PPI Networks in Urothelial Carcinoma

We performed Proteinarium analysis using RNA-seq data from patients with papillary and nonpapillary urothelial carcinoma to compare their PPI networks. We analyzed the top 125 upregulated genes obtained from RNA-seq data from TCGA Cell 2017 [15] and TCGA Nature 2014 [16] datasets, as well as the top 125 downregulated genes from the TCGA Cell 2017 dataset.

In the analysis of the TCGA Cell 2017 dataset, 1 significant cluster (n=393; P<.001) was identified for further analysis, which included 89 (67.4%) of 132 patients with papillary urothelial carcinoma and 214 (79.3%) of 270 patients with nonpapillary urothelial carcinoma (Figure 1). In the network specific to papillary urothelial carcinoma (network A), RPS27A, UBA52, and VAMP8 were the most connected or hub proteins identified. In the network specific to nonpapillary carcinoma (network B), GNB1, RHOA, UBC, and FPR2 were found to be the hub proteins (Figure 1). In the analysis of the TCGA Nature 2014 dataset, 1 cluster of 6 patients with papillary urothelial carcinoma (network C) was identified as a significant cluster based on Fisher exact test (P<.001; Figure 2). In this PPI network, CCDC22, ANAPC4, and UBR4 were identified as hub proteins unique to papillary urothelial carcinoma. Six proteins were shared between the 2 papillary urothelial carcinoma consensus networks (networks A and C), including UBR4, CUL1, UBE2K, CDC5L, UBA52, and RPS27A.

Figure 1. Dendrogram and consensus networks specific to papillary and nonpapillary urothelial carcinoma from The Cancer Genome Atlas Cell 2017 dataset (upregulated genes). The dendrogram shows 402 patients who were clustered based on their network similarities. Patients with green branches are part of the significant cluster whose networks are next to the dendrogram. In the networks, larger circles represent more connected nodes. Proteins specific to the papillary and nonpapillary groups are shown in yellow and blue, respectively. Proteins in both groups are shown in green. Proteins in red were imputed from the network analysis.
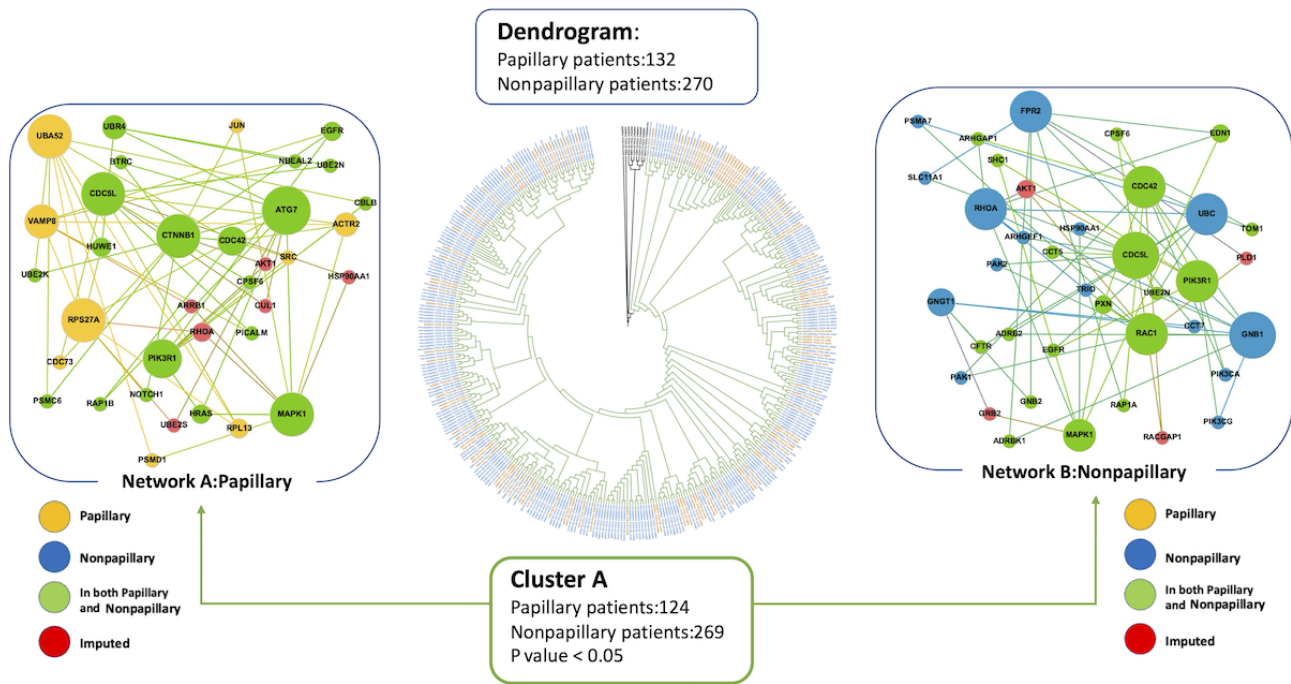
Figure 2. Dendrogram and consensus networks specific to papillary and nonpapillary urothelial carcinoma from The Cancer Genome Atlas Nature 2014 dataset (upregulated genes). The dendrogram shows 127 patients, who were clustered based on their network similarities. Patients with yellow branches are part of the significant cluster, whose network is shown to the left of the dendrogram. In the networks, larger circles represent more connected nodes. Proteins specific to the papillary and nonpapillary groups are shown in yellow and blue, respectively. Proteins shared by both groups are shown in green (none were identified in this dataset). Proteins in red were imputed from the network analysis.
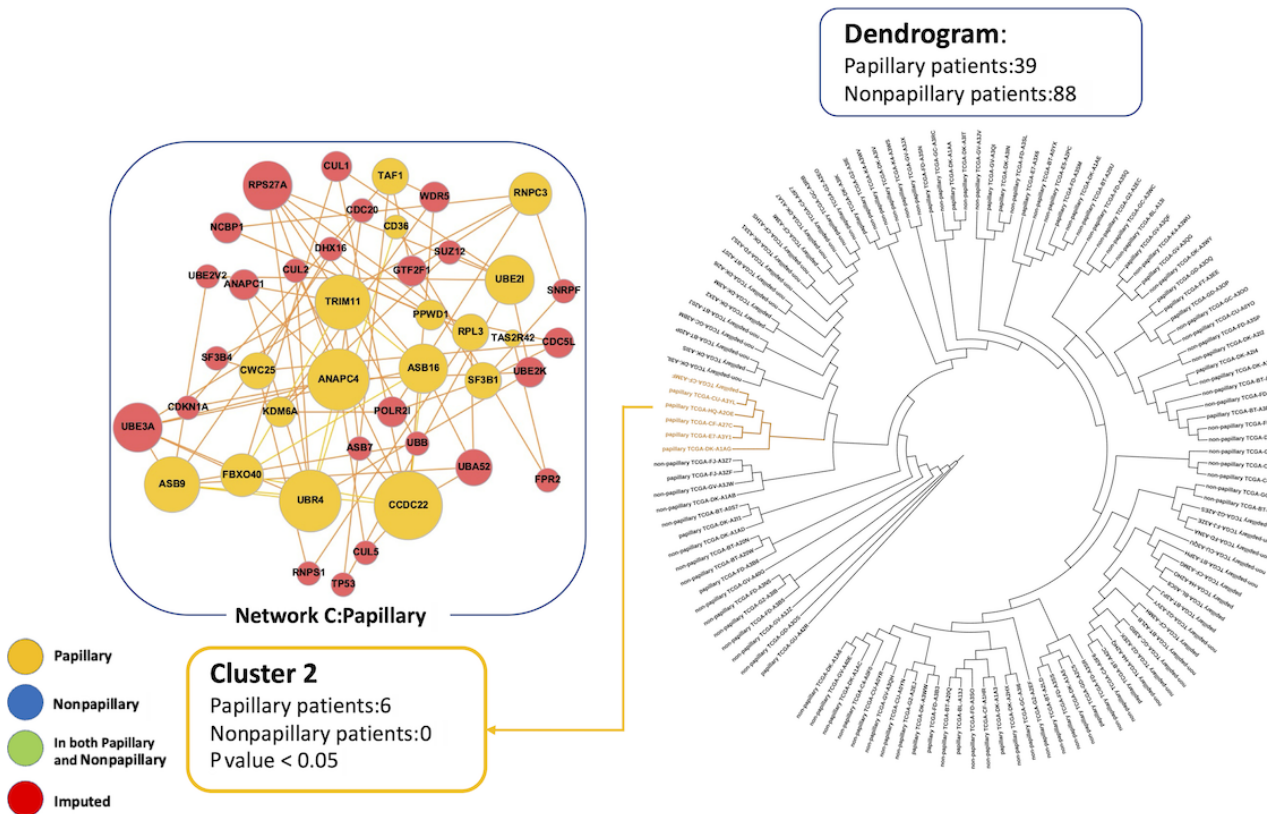
We calculated the separation score between the papillary- and nonpapillary-specific networks from the TCGA Cell 2017 dataset (networks A and B). Using only the seed genes unique to each network, the separation score was $s_{AB}=0.1722$. This

value indicated that the papillary and nonpapillary consensus networks contained protein modules that did not overlap in the interactome and can therefore be considered distinct. Per Menche et al [23], strongly segregated phenotypes, such as mismatch repair-deficient and mismatch repair-intact breast cancer, have s_{AB} of approximately 0.32. Thus, an s_{AB} equal to 0.1722 reflects a moderate yet biologically meaningful network separation. The separation score comparing networks A and C was calculated as $s_{AB}=-0.0661$. The negative s_{AB} value demonstrated that there was an overlap between the consensus PPI networks specific to papillary urothelial carcinoma using 2 independent datasets.

We analyzed PPI using the 125 most downregulated genes (Figure S4 in Multimedia Appendix 1). In the downregulated consensus networks, proteins unique to patients with papillary urothelial carcinoma included APP, RTP2, TP53, TAS2R13, SPTAN1, EPHB2, HIST1H2BH, MAPK9, and TAS2R42, whereas proteins unique to the PPI network specific to nonpapillary urothelial carcinoma included ANAPC13, LNX1, TAS2R46, REEP5, UBE2V2, UBB, EP300, and GPR31. Negative separation score ($s_{AB}=-1.4706$) indicated an overlap between these 2 PPI networks.

Given the distinct genetic differences between low-grade and high-grade urothelial carcinomas, we compared the PPI networks of patients with high-grade papillary carcinoma (n=114) and high-grade nonpapillary carcinoma (n=267; Figure S5 in Multimedia Appendix 1). Several proteins were shared between high-grade papillary and nonpapillary consensus networks, suggesting shared genetic features of high-grade urothelial carcinomas. The hub proteins shared between the high-grade papillary and nonpapillary networks included UBA52, CDC5L, CDC42, PIK3R1, and MAPK1. The proteins unique to the high-grade papillary network included SRC, ACTR2, WASL, CDC73, PSMD1, and JUN. The proteins unique to the network specific to high-grade nonpapillary urothelial carcinoma included GNB1, UBC, FPR2, GNGT1, RHOA, PIK3CA, PIK3CG, PXN, UBE2N, SLC11A1, CCT7, HSP90AA1, and DOCK4. The separation score comparing the high-grade papillary and nonpapillary urothelial carcinoma-specific PPI networks was −0.2739, demonstrating the overlap between these networks.

We annotated network modularity for patients with papillary and nonpapillary urothelial carcinoma in both datasets by using GEPHI, a network and graph visualization tool. In this context, modules represented clusters of proteins that were more densely connected to each other than to the rest of the network. We found 5 modules in which proteins were grouped by common functionality for each network in networks A and C, and 6 modules in network B. We used the default setting in GEPHI. gProfiler was used to annotate the modules for networks A, B, and C to identify significant Gene Ontology (GO) terms and KEGG pathways [20]. Notably, the papillary networks from both datasets (networks A and C) contained 1 module enriched in proteins involved in ubiquitin-related activities. The KEGG pathway "Ubiquitin-mediated proteolysis" and the molecular function GO terms "ubiquitin-like protein conjugating enzyme

activity" as well as "ubiquitin protein ligase binding" were common to the module found in networks A and C.

Pathway and Gene Set Enrichment Analysis

We used the g:GOSt functional profiling tool from g:Profiler to visualize significant GO terms and KEGG pathways for the genes in networks A, B, and C [20] to visualize significant GO terms and KEGG pathways for the genes in networks A, B, and C (Table S1, S2, and S3 in Multimedia Appendix 3). The proteins unique to nonpapillary urothelial carcinoma (network B) were enriched in the PI3K-Akt signaling pathway (KEGG pathway, q=o.0040), with the most significant molecular function GO term being phosphatidylinositol-4,5-bisphosphate-3-kinase activity. The GO terms and KEGG pathways related to the PI3K-Akt signaling pathway were unique to network B. Ubiquitin-mediated proteolysis was the most significant KEGG pathway for network C (q=2.742e-12), and ubiquitin-like protein ligase binding was the most significant molecular function term for network A (q=0.00110187).

Drug Repurposing

GNB1, the hub protein unique in the network specific to nonpapillary urothelial carcinoma, encodes the beta subunit of G proteins, which modulate transmembrane signaling, including that of the PI3K/AKT signaling pathway [25]. Using the CMap drug repurposing tool, we identified Y16 as a potential drug to target GNB1. Y16 is a rho-associated kinase inhibitor in preclinical trials. Other rho-associated kinase inhibitors, including Y-27632 and HA-1077, were shown to inhibit the proliferation and invasion of urothelial carcinoma cells and induce apoptosis, respectively [26,27]. FPR2, another hub protein unique in the networks specific to nonpapillary urothelial carcinoma, encodes the formyl peptide receptor 2 and is involved in the response to amyloid beta, regulation of the defense response, and positive regulation of monocyte chemotaxis [28]. TC-FPR2-43 is a formyl peptide receptor agonist in preclinical trials, which may be considered for use in treating nonpapillary urothelial carcinoma. We also identified drugs targeting other proteins unique to nonpapillary urothelial carcinoma, including PIK3CA and PAK1. PIK3CA, which encodes the p100α catalytic subunit of PI3K [29], is the target of several PI3K inhibitors used to treat breast cancer (alpelisib), follicular lymphoma (copanlisib), and chronic lymphocytic leukemia (idelalisib). PAK1, a member of the serine/threonine p2-activating kinase family, is targeted by several drugs in preclinical trials—serine/threonine kinase inhibitor FRAX486, p21 activated kinase inhibitors IPA-3 and NVS-PAK1-1, and rho-associated kinase inhibitor RKI-1447.

Discussion

Principal Findings

The genomic landscape distinguishing papillary and nonpapillary urothelial carcinoma architectures is not completely understood. Using a network biology approach, we sought to understand how the PPI networks differed between patients with papillary and nonpapillary urothelial carcinoma. Proteinarium, a multisample PPI analysis tool, was used to

analyze gene expression data from patients with papillary and nonpapillary urothelial carcinoma and identify the distinct PPI networks and hub proteins specific to these 2 architectures. Using the CMap drug repurposing tool, we identified known drugs that target the identified hub proteins. Our results showed that RPS27A, UBA52, and VAMP8 were the hub proteins in the network specific to papillary urothelial carcinoma (network A). This PPI network shared 6 proteins with network C, also specific to papillary urothelial carcinoma, obtained using the validation dataset. These 6 proteins—UBR4, CUL1, UBE2K, CDC5L, UBA52, and RPS27A—were enriched in gProfiler for ubiquitination processes, including ubiquitin-mediated proteolysis. Dysregulation of the ubiquitin-proteasome system, which regulates tumor suppressors and oncogenic proteins, is observed across cancer types [30]. In the PPI network specific to nonpapillary urothelial carcinoma, network B, we identified GNB1, RHOA, UBC, and FPR2 as hub proteins, which were involved in the PI3K/AKT signaling pathway. Additionally, we compared the PPI networks specific to high-grade papillary and nonpapillary urothelial carcinoma. Networks F and G (Figure S5 in Multimedia Appendix 1) exhibited an overlap. This finding is consistent with the shared genetic drivers that underlie high-grade tumors.

Comparison to Prior Work

Of the proteins shared by networks A and C, the 2 PPI networks specific to papillary urothelial carcinoma, CDC5L and CUL1, have been found to be upregulated in urothelial carcinoma in previous studies. It has been shown that their overexpression is significantly associated with poorer prognosis in patients diagnosed with urothelial carcinoma [31,32]. Knockdown of CDC5L inhibited the proliferation of urothelial carcinoma cells by inducing apoptosis and limiting bladder cancer cell migration, invasion, and epithelial-mesenchymal transition [32]. Currently, there are no drugs in development that target CDC5L or CUL1. RPS27A and UBA52, which are hub proteins unique to papillary urothelial carcinoma in network A, have been found to be overexpressed in colon cancer [33], prostate cancer [34], and leukemia [35]. RPS27A and UBA52 are both ribosomal fusion proteins composed of ubiquitin conjugated to the ribosomal proteins S27a and L40, respectively. RPS27A, an RNA-binding protein part of the ribosomal 40S subunit [36], plays a role in inhibiting apoptosis, regulating the progression of the cell cycle, and promoting proliferation [35].

UBA52 was identified as a hub protein in the network specific to papillary urothelial carcinoma from the TCGA Cell 2017 dataset, and it also emerged as a shared hub protein in both high-grade papillary and nonpapillary urothelial carcinoma-specific PPI networks. UBA52 encodes a ubiquitin-ribosomal fusion protein involved in essential processes such as protein synthesis and degradation, both of which are upregulated in rapidly proliferating cancers. UBA52 is required for embryonic development and regulates protein synthesis and the cell cycle by modulating the expression of cyclins D1 and D3 [37]. Other shared hub proteins, including MAPK1, CDC42, and PIK3R1, are also involved in key signaling and regulatory pathways that support tumor progression.

PIK3CA is a protein in the nonpapillary urothelial carcinoma-specific PPI network [29]. PIK3CA upregulation has been implicated in bladder cancer by promoting cell growth and regulating metastasis [38-40]. Specifically, knockdown of PIK3CA was found to substantially inhibit cell proliferation [40], whereas overexpression of PIK3CA promoted bladder cancer cell growth, migration, invasion, and metastasis [39]. Higher levels of PIK3CA expression were associated with worse prognosis in bladder cancer [39]. Subsequent activation of PI3K at the plasma membrane by dimerized receptor tyrosine kinases causes phosphorylation of phosphatidylinositol 4,5-bisphosphate to produce the secondary messenger phosphatidylinositol 3,4,5-trisphosphate. Several PI3K inhibitors have been developed to treat breast cancer (alpelisib), follicular lymphoma (copanlisib), and chronic lymphocytic leukemia (idelalisib). PI3K inhibition has been found to suppress the growth, migration, and colony formation of bladder cancer cells in vitro [41].

RHOA is a unique hub protein in the PPI network specific to the nonpapillary urothelial carcinoma. Overexpression of RHOA, a member of the rho family of small guanosine triphosphatases that regulates cell attachment, motility, and shape [29], has been associated with bladder cancer cell proliferation and metastasis [41-43]. Ras-induced RhOA and NF-kappaB activation were implicated in promoting the invasion and migration of bladder cancer cells [44,45]. Three of the 4 hub proteins of network B, the PPI network specific to nonpapillary urothelial carcinoma—GNB1, RHOA, and FPR2—have been found to be associated with worse prognosis in gastric cancer [46], bladder cancer [42], lung cancer [47], and cervical squamous cell carcinoma [48].

Limitations

There are several limitations in this study. First, given tumor-node-metastasis (TNM) staging data were not available for the TCGA Nature 2014 dataset, and 344 (99.4%) out of 346 patients in the TCGA Cell 2017 dataset were muscle invasive (T2, T3, and T4). Therefore, these findings may not apply to nonmuscle invasive disease (T1) and should not be generalized. Second, the data were based on z scores of RNA-seq data. Although RNA transcript abundance reflects relative protein abundance, it is not a direct measurement of the quantity of each specific protein. By using z scores rather than absolute RNA transcript counts, we were able to draw conclusions about how a gene is upregulated or downregulated relative to the rest of the samples.

Conclusions

This study identified 2 distinct PPI networks associated with the molecular architecture of papillary and nonpapillary urothelial carcinoma, suggesting that these architectures represent biologically distinct entities driven by different molecular mechanisms. The hub proteins within each network may serve as potential targets for papillary and nonpapillary urothelial carcinoma. Further studies are required to determine the efficacy of these identified drug targets.

Acknowledgments

Data Availability

All data collected and analyzed in this study are publicly available at cBioPortal [49,50].

Authors' Contributions

Conceptualization: AU, EDGU
Data Curation: CC, YB, SH
Formal Analysis: CC, YB, SH
Methodology: CC, YB, SH, AA, LC, AU, EDGU
Supervision: AU, EDGU
Writing - original draft: CC, YB, AU, EDGU
Writing - review and editing: CC, YB, SH, AA, LC, AU, EDGU

Conflicts of Interest

EDGU serves as the editor-in-chief of *JMIR Bioinformatics and Biotechnology*. SH and AU serve on the editorial board of *JMIR Bioinformatics and Biotechnology*. These authors have not been involved in the peer review of this study.

Multimedia Appendix 1
Details on protein-protein interaction network analysis and cophenetic scores.
[PDF File, 3912 KB - bioinform_v6i1e76736_app1.pdf]

Multimedia Appendix 2
Input data for the reproducibility of protein-protein interaction network analysis.
[ZIP File, 10 KB - bioinform_v6i1e76736_app2.zip]

Multimedia Appendix 3
Significant Gene Ontology terms and Kyoto Encyclopedia of Genes and Genomes pathways enriched in networks A, B, and C.
[XLSX File, 20 KB - bioinform_v6i1e76736_app3.xlsx]

References

1. Siegel RL, Kratzer TB, Giaquinto AN, Sung H, Jemal A. Cancer statistics, 2025. CA Cancer J Clin 2025;75(1):10-45. [doi: 10.3322/caac.21871] [Medline: 39817679]
2. Edge SB, Compton CC. The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. Ann Surg Oncol 2010 Jun;17(6):1471-1474. [doi: 10.1245/s10434-010-0985-4] [Medline: 20180029]
3. McConkey DJ, Lee S, Choi W, et al. Molecular genetics of bladder cancer: emerging mechanisms of tumor initiation and progression. Urol Oncol 2010;28(4):429-440. [doi: 10.1016/j.urolonc.2010.04.008] [Medline: 20610280]
4. Zhang B, Tian Y, Zhang Z. Network biology in medicine and beyond. Circ Cardiovasc Genet 2014 Aug;7(4):536-547. [doi: 10.1161/CIRCGENETICS.113.000123] [Medline: 25140061]
5. Vidal M, Cusick ME, Barabási AL. Interactome networks and human disease. Cell 2011 Mar 18;144(6):986-998. [doi: 10.1016/j.cell.2011.02.016] [Medline: 21414488]
6. Gavin AC, Aloy P, Grandi P, et al. Proteome survey reveals modularity of the yeast cell machinery. Nature 2006 Mar 30;440(7084):631-636. [doi: 10.1038/nature04532] [Medline: 16429126]
7. Li S, Armstrong CM, Bertin N, et al. A map of the interactome network of the metazoan C. elegans. Science 2004 Jan 23;303(5657):540-543. [doi: 10.1126/science.1091403] [Medline: 14704431]
8. Oti M, Snel B, Huynen MA, Brunner HG. Predicting disease genes using protein-protein interactions. J Med Genet 2006 Aug;43(8):691-698. [doi: 10.1136/jmg.2006.041376] [Medline: 16611749]
9. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res 2006 Jan 1;34(Database issue):D535-D539. [doi: 10.1093/nar/gkj109] [Medline: 16381927]

10. Bi D, Ning H, Liu S, Que X, Ding K. Gene expression patterns combined with network analysis identify hub genes associated with bladder cancer. Comput Biol Chem 2015 Jun;56:71-83. [doi: 10.1016/j.compbiolchem.2015.04.001] [Medline: 25889321]

11. Deng SP, Zhu L, Huang DS. Mining the bladder cancer-associated genes by an integrated strategy for the construction and analysis of differential co-expression networks. BMC Genomics 2015;16 Suppl 3(Suppl 3):S4. [doi: 10.1186/1471-2164-16-S3-S4] [Medline: 25707808]

12. Li S, Liu X, Liu T, et al. Identification of biomarkers correlated with the TNM staging and overall survival of patients with bladder cancer. Front Physiol 2017 Nov 28;8:947. [doi: 10.3389/fphys.2017.00947] [Medline: 29234286]

13. Yuan L, Shu B, Chen L, et al. Overexpression of COL3A1 confers a poor prognosis in human bladder cancer identified by co-expression analysis. Oncotarget 2017 Jul 28;8(41):70508-70520. [doi: 10.18632/oncotarget.19733] [Medline: 29050298]

14. Armanious D, Schuster J, Tollefson GA, et al. Proteinarium: multi-sample protein-protein interaction analysis and visualization tool. Genomics 2020 Nov;112(6):4288-4296. [doi: 10.1016/j.ygeno.2020.07.028] [Medline: 32702417]

15. Robertson AG, Kim J, Al-Ahmadie H, et al. Comprehensive molecular characterization of muscle-invasive bladder cancer. Cell 2017 Oct 19;171(3):540-556.e25. [doi: 10.1016/j.cell.2017.09.007] [Medline: 28988769]

16. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. Nature 2014 Mar 20;507(7492):315-322. [doi: 10.1038/nature12965] [Medline: 24476821]

17. Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov 2012 May;2(5):401-404. [doi: 10.1158/2159-8290.CD-12-0095] [Medline: 22588877]

18. Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 2013 Apr 2;6(269):pl1. [doi: 10.1126/scisignal.2004088] [Medline: 23550210]

19. Szklarczyk D, Gable AL, Nastou KC, et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res 2021 Jan 8;49(D1):D605-D612. [doi: 10.1093/nar/gkaa1074] [Medline: 33237311]

20. Raudvere U, Kolberg L, Kuzmin I, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res 2019 Jul 2;47(W1):W191-W198. [doi: 10.1093/nar/gkz369] [Medline: 31066453]

21. Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. Proc Int AAAI Conf Web Soc Media 2009;3(1):361-362. [doi: 10.1609/icwsm.v3i1.13937]

22. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008 Apr;2008(10). [doi: 10.1088/1742-5468/2008/10/P10008]

23. Menche J, Sharma A, Kitsak M, et al. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. Science 2015 Feb 20;347(6224):1257601. [doi: 10.1126/science.1257601] [Medline: 25700523]

24. Subramanian A, Narayan R, Corsello SM, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell 2017 Nov 30;171(6):1437-1452.e17. [doi: 10.1016/j.cell.2017.10.049] [Medline: 29195078]

25. O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 2016 Jan 4;44(D1):D733-d745. [doi: 10.1093/nar/gkv1189] [Medline: 26553804]

26. Abe H, Kamai T, Hayashi K, et al. The rho-kinase inhibitor HA-1077 suppresses proliferation/migration and induces apoptosis of urothelial cancer cells. BMC Cancer 2014 Jun 7;14:412. [doi: 10.1186/1471-2407-14-412] [Medline: 24908363]

27. Jiang L, Wen J, Luo W. Rho-associated kinase inhibitor, Y-27632, inhibits the invasion and proliferation of T24 and 5367 bladder cancer cells. Mol Med Rep 2015 Nov;12(5):7526-7530. [doi: 10.3892/mmr.2015.4404] [Medline: 26459851]

28. Alliance of Genome Resources Consortium. Alliance of Genome Resources Portal: unified model organism research platform. Nucleic Acids Res 2020 Jan 8;48(D1):D650-D658. [doi: 10.1093/nar/gkz813] [Medline: 31552413]

29. Sayers EW, Bolton EE, Brister JR, et al. Database resources of the national center for biotechnology information. Nucleic Acids Res 2022 Jan 7;50(D1):D20-D26. [doi: 10.1093/nar/gkab1112] [Medline: 34850941]

30. Morrow JK, Lin HK, Sun SC, Zhang S. Targeting ubiquitination for cancer therapies. Future Med Chem 2015;7(17):2333-2350. [doi: 10.4155/fmc.15.148] [Medline: 26630263]

31. Mao SY, Xiong DB, Huang TB, Zheng JH, Yao XD. Expression of CUL1 correlates with tumour-grade and recurrence in urothelial carcinoma. ANZ J Surg 2017 Jul;87(7-8):624-629. [doi: 10.1111/ans.13438] [Medline: 27312089]

32. Zhang Z, Mao W, Wang L, et al. Depletion of CDC5L inhibits bladder cancer tumorigenesis. J Cancer 2020 Jan 1;11(2):353-363. [doi: 10.7150/jca.32850] [Medline: 31897231]

33. Mu Q, Luo G, Wei J, et al. Apolipoprotein M promotes growth and inhibits apoptosis of colorectal cancer cells through upregulation of ribosomal protein S27a. EXCLI J 2021 Jan 21;20:145-159. [doi: 10.17179/excli2020-2867] [Medline: 33564284]

34. Zhang Y, Huang Z, Zhu Z, Liu J, Zheng X, Zhang Y. Network analysis of ChIP-Seq data reveals key genes in prostate cancer. Eur J Med Res 2014 Sep 3;19(1):47. [doi: 10.1186/s40001-014-0047-7] [Medline: 25183411]

35. Wang H, Yu J, Zhang L, et al. RPS27a promotes proliferation, regulates cell cycle progression and inhibits apoptosis of leukemia cells. Biochem Biophys Res Commun 2014 Apr;446(4):1204-1210. [doi: 10.1016/j.bbrc.2014.03.086]

XSL•FO

RenderX

36. Chen L, Wang H. Nicotine promotes human papillomavirus (HPV)-immortalized cervical epithelial cells (H8) proliferation by activating RPS27a-Mdm2-P53 pathway in vitro. Toxicol Sci 2019 Feb 1;167(2):408-418. [doi: 10.1093/toxsci/kfy246] [Medline: 30272249]

37. Kobayashi M, Oshima S, Maeyashiki C, et al. The ubiquitin hybrid gene UBA52 regulates ubiquitination of ribosome and sustains embryonic development. Sci Rep 2016 Nov 10;6:36780. [doi: 10.1038/srep36780] [Medline: 27829658]

38. Liu X, Cui J, Gong L, et al. The CUL4B-miR-372/373-PIK3CA-AKT axis regulates metastasis in bladder cancer. Oncogene 2020 Apr;39(17):3588-3603. [doi: 10.1038/s41388-020-1236-1] [Medline: 32127645]

39. Wang Z, Shang J, Li Z, et al. PIK3CA Is regulated by CUX1, promotes cell growth and metastasis in bladder cancer via activating epithelial-mesenchymal transition. Front Oncol 2020;10:536072. [doi: 10.3389/fonc.2020.536072] [Medline: 33344221]

40. Zhu S, Ma AH, Zhu Z, et al. Synergistic antitumor activity of pan-PI3K inhibition and immune checkpoint blockade in bladder cancer. J Immunother Cancer 2021 Nov;9(11):e002917. [doi: 10.1136/jitc-2021-002917] [Medline: 34725212]

41. Chang HR, Huang HP, Kao YL, et al. The suppressive effect of Rho kinase inhibitor, Y-27632, on oncogenic Ras/RhoA induced invasion/migration of human bladder cancer TSGH cells. Chem Biol Interact 2010 Jan 5;183(1):172-180. [doi: 10.1016/j.cbi.2009.10.018] [Medline: 19896475]

42. Kamai T, Tsujii T, Arai K, et al. Significant association of rho/ROCK pathway with invasion and metastasis of bladder cancer. Clin Cancer Res 2003 Jul;9(7):2632-2641. [Medline: 12855641]

43. Volanis D, Zaravinos A, Kadiyska T, Delakas D, Zoumpourlis V, Spandidos DA. Expression profile of rho kinases in urinary bladder cancer. J BUON 2011;16(3):511-521. [Medline: 22006759]

44. Parri M, Chiarugi P. Rac and Rho GTPases in cancer cell motility control. Cell Commun Signal 2010 Sep 7;8:23. [doi: 10.1186/1478-811X-8-23] [Medline: 20822528]

45. Zhu J, Li Y, Chen C, et al. NF-κB p65 overexpression promotes bladder cancer cell migration via FBW7-mediated degradation of RhoGDIα protein. Neoplasia 2017 Sep;19(9):672-683. [doi: 10.1016/j.neo.2017.06.002] [Medline: 28772241]

46. Hou XL, Ji CD, Tang J, et al. FPR2 promotes invasion and metastasis of gastric cancer cells and predicts the prognosis of patients. Sci Rep 2017 Jun 9;7(1):3153. [doi: 10.1038/s41598-017-03368-7] [Medline: 28600569]

47. Fan Z, Bai Y, Zhang Q, Qian P. CircRNA circ_POLA2 promotes lung cancer cell stemness via regulating the miR-326/GNB1 axis. Environ Toxicol 2020 Oct;35(10):1146-1156. [doi: 10.1002/tox.22980] [Medline: 32511866]

48. Cao Y, Li J, Jia Y, Zhang R, Shi H. CircRNA circ_POLA2 promotes cervical squamous cell carcinoma progression via regulating miR-326/GNB1. Front Oncol 2020;10:959. [doi: 10.3389/fonc.2020.00959] [Medline: 32766125]

49. Bladder cancer (TCGA, cell 2017). cBioPortal. URL: https://www.cbioportal.org/study/summary?id=blca_tcga_pub_2017 [accessed 2025-11-10]

50. Bladder urothelial carcinoma. cBioPortal. URL: https://www.cbioportal.org/study/summary?id=blca_tcga_pub [accessed 2025-11-10]

Abbreviations

CMap: connectivity map
GO: Gene Ontology
KEGG: Kyoto Encyclopedia of Genes and Genomes
PPI: protein-protein interaction
RNA-seq: RNA sequencing
TCGA: The Cancer Genome Atlas

Structural and Functional Impacts of SARS-CoV-2 Spike Protein Mutations: Insights From Predictive Modeling and Analytics

Edem K Netsey[1*], MPhil; Samuel M Naandam[2], PhD; Joseph Asante Jnr[3*], PhD; Kuukua E Abraham[4], MS; Aayire C Yadem[5], PhD; Gabriel Owusu[6], PhD; Jeffrey G Shaffer[7], PhD; Sudesh K Srivastav[7], PhD; Seydou Doumbia[8], PhD; Ellis Owusu-Dabo[9], PhD; Chris E Morkle[10], MEd; Desmond Yemeh[11], PharmD; Stephen Manortey[12], PhD; Ernest Yankson[2], PhD; Mamadou Sangare[13], PhD; Samuel Kakraba[7,11*], PhD

[1]Department of Mathematics and Information Communication Technology, Dambai College of Education, Dambai, Ghana

[10]Department of Mathematics, Suhum Senior High School, Suhum, Ghana

[11]Tulane Center for Aging, School of Medicine, Tulane University, New Orleans, LA, United States

[12]Department of Community Health, Ensign Global University, Kpong, Ghana

[13]Laboratory of Malaria and Vector Research (LMVR), National Institute of Allergy and Infectious Diseases (NIAID), Rockville, MD, United States

[2]Department of Mathematics, School of Physical Sciences, University of Cape Coast, Cape Coast, Ghana

[3]Department of Geriatrics, School of Medicine, University of Arkansas for Medical Sciences, Little Rock, AR, United States

[4]Department of Mathematics, Memphis Shelby County Schools, Memphis, TN, United States

[5]Research and Development, CytoAstra LLC, Little Rock, AR, United States

[6]Office of Research, Celia Scott Weatherhead School of Public Health and Tropical Medicine at Tulane University, New Orleans, LA, United States

[7]Department of Biostatistics and Data Science, Celia Scott Weatherhead School of Public Health and Tropical Medicine at Tulane University, 1440 Canal Street, New Orleans, LA, United States

[8]Department of Public Health, Faculty of Medicine, Malaria Research and Training Center, University of Sciences, Techniques and Technologies of Bamako, Bamako, Mali

[9]School of Public Health, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

[*]these authors contributed equally

Corresponding Author:
Samuel Kakraba, PhD
Department of Biostatistics and Data Science, Celia Scott Weatherhead School of Public Health and Tropical Medicine at Tulane University, 1440 Canal Street, New Orleans, LA, United States

---

Related Article:

This is a corrected version. See correction statement: https://bioinform.jmir.org/2025/1/e89673

Abstract

Background: The COVID-19 pandemic requires a deep understanding of SARS-CoV-2, particularly how mutations in the spike receptor-binding domain (RBD) chain E affect its structure and function. Current methods lack comprehensive analysis of these mutations at different structural levels.

Objective: This study aims to analyze the impact of specific COVID-19–associated point mutations (N501Y, L452R, N440K, K417N, and E484A) on the SARS-CoV-2 spike RBD structure and function using predictive modeling, including a graph-theoretic model, protein modeling techniques, and molecular dynamics simulations.

Methods: The study used a multitiered graph-theoretic framework to represent protein structure across 3 interconnected levels. This model incorporated 19 top-level vertices, connected to intermediate graphs based on 6-angstrom proximity within the protein's 3D structure. Graph-theoretic molecular descriptors or invariants were applied to weigh vertices and edges at all levels. The study also used Iterative Threading Assembly Refinement (I-TASSER) to model mutated sequences and molecular dynamics simulation tools to evaluate changes in protein folding and stability compared to the wildtype.

Results: A total of 3 distinct predictive modeling and analytical approaches successfully identified structural and functional changes in the SARS-CoV-2 spike RBD (chain E) resulting from point mutations. The novel graph-theoretic model detected notable structural changes, with N501Y and L452R showing the most pronounced effects on conformation and stability compared to the wildtype. K147N and E484A mutations demonstrated less significant impacts compared to the severe mutations, N501Y and L452R. Ab initio modeling and molecular simulation dynamics findings corroborated the results from graph-theoretic analysis.

The multilevel analytical approach provided a comprehensive visualization of mutation effects, deepening our understanding of their functional consequences.

Conclusions: This study advanced our understanding of SARS-CoV-2 spike RBD mutations and their implications. The multifaceted approach characterized the effects of various mutations, identifying N501Y and L452R as having the most substantial impact on RBD conformation and stability. The findings have important implications for vaccine development, therapeutic design, and variant monitoring. Our research underscores the power of combining multiple predictive analytical approaches in virology, contributing valuable knowledge to ongoing efforts against the COVID-19 pandemic and providing a framework for future studies on viral mutations and their impacts on protein structure and function.

Introduction

Background

The COVID-19 pandemic, caused by SARS-CoV-2, had a devastating global impact, with hundreds of millions of confirmed cases and millions of deaths worldwide [1-4]. The virus' spike protein shares significant structural and sequence similarities with the severe acute respiratory syndrome virus from 2003, including the use of the angiotensin-converting enzyme 2 (ACE2) receptor for cell entry. The SARS-CoV-2 spike protein's receptor-binding domain (RBD) has undergone various mutations, each with distinct impacts on viral behavior and infectivity [5]. The wildtype phenotype serves as a baseline, exhibiting no significant changes or enhanced impacts [1]. Protein folding, maintenance, mutation, aggregation, neurodegenerative diseases, and COVID-19 are interconnected through complex biological processes [6-8]. Except for intrinsically disordered proteins, most proteins must fold correctly to function properly, but mutations can disrupt this process, leading to misfolding and aggregation. Cells have protective mechanisms to maintain protein homeostasis, but when these fail, aggregates can form and contribute to neurodegenerative diseases like Alzheimer disease (AD) and Parkinson disease [6-10]. COVID-19 has been linked to AD, Parkinson disease, and other neurodegenerative diseases through various mechanisms [11-13]. The SARS-CoV-2 spike protein contains aggregation-prone regions that can form amyloid aggregates, potentially contributing to neurological complications [14-16]. In addition, the virus' 3CL protease has been shown to induce tau protein aggregation, a hallmark of neurodegenerative diseases. Genetic studies have revealed a causal association between COVID-19 hospitalization and increased risk of AD [17]. The virus may accelerate neurodegeneration through inflammation, microvascular injury, and prion-like spread of misfolded protein. These aggregates were observed in vitro using various experimental techniques, including fluorescence spectroscopy and electron microscopy [18-20].

Specific mutations have emerged that alter the virus' characteristics in different ways. Mutations such as K417N, associated with the Beta, Gamma, and Omicron variants, demonstrate moderate severity due to their role in immune evasion [21-23]. These mutations reduce antibody neutralization, potentially compromising the effectiveness of vaccines and natural immunity [24]. Similarly, N440K and S477N mutations, linked to localized outbreaks, show mild impacts by enhancing binding to the ACE2 receptor, which increases infectivity without causing major antigenic alterations [25,26]. More concerning are mutations like T478K and L452R, found in Delta and Omicron variants, which display moderate to severe consequences. L452R, in particular, is associated with immune evasion and increased transmission due to its resistance to neutralizing antibodies [27]. The E484 mutations (A, K, Q), present in Beta, Gamma, and other variants, exhibit moderate to severe immune escape capabilities, significantly reducing neutralization by vaccine-induced or convalescent sera [28]. Of note is the profound N501Y mutation, observed in Alpha, Beta, and Omicron variants. This mutation is classified as very severe due to its dual role in enhancing binding affinity to the ACE2 receptor and aiding in immune escape [29]. N501Y has been a pivotal factor in increasing transmissibility and contributing to widespread outbreaks [30]. These diverse mutations in the SARS-CoV-2 RBD demonstrate varying degrees of severity and impact, ranging from mild infectivity increases to severe immune evasion and transmissibility. The evolving nature of these mutations underscores the critical importance of continuous genomic surveillance and adaptive vaccine strategies in combating the COVID-19 pandemic [31-33]. As SARS-CoV-2 has spread, it has evolved into various strains, with the D614G mutation becoming nearly ubiquitous. Several "variants of concern" (VOCs) have emerged, characterized by mutations that may affect viral behavior and immune evasion [34-38]. These VOCs, identified in different regions, include B.1.1.7 (UK), P.1 and P.2 (Brazil), B.1.351 (South Africa), B.1.617 (India), and B.1.526 (US), among others [39]. The emergence of these variants has significant implications for treatment strategies and vaccine efficacy. Some variants, such as B.1.351, have shown resistance to neutralization by convalescent plasma and certain monoclonal antibody treatments [40]. In addition, in vitro studies suggest that sera from vaccinated individuals may have reduced neutralizing capacity against variants with specific mutations, like E484K and N501Y [23,41,42]. While VOCs are defined by specific mutation patterns, the interplay between these mutations in affecting viral behavior is not fully understood [43-46]. Analysis of a large dataset of SARS-CoV-2 spike sequences revealed hundreds of amino acid variants, with a significant number occurring in the

RBD [47,48]. Machine learning algorithms have revolutionized data analysis by uncovering hidden patterns in datasets and addressing critical questions across various disciplines, including disease diagnostics, among others [49-52]. Unsupervised learning techniques like clustering and dimensionality reduction reveal intrinsic structures in data [53], enabling applications such as customer segmentation, anomaly detection, and in the field of bioinformatics [54]. To better understand the emergence and spread of new variants, novel bioinformatics approaches are being developed to identify spatially and temporally correlated mutations [55,56]. The evolving nature of SARS-CoV-2 suggests that future vaccine design may need to be tailored to address the specific strain ensembles prevalent in different regions [57-59]. This approach could enhance vaccine efficacy against locally dominant variants and potentially provide broader protection against emerging strains.

Previous Work

Recent advancements in graph-theoretic modeling have provided valuable insights into the complex relationship between point mutations and disease phenotypes [60,61]. Researchers have developed various graph-based approaches to model and analyze the intricate connections between genes, mutations, and phenotypes. These methods leverage the ability of graphs to represent complex biological relationships and interactions [62-75]. This approach has been particularly effective in elucidating the structural and functional impacts of genetic variations in proteins associated with hereditary disorders. One application of graph theory in mutation analysis involves the use of protein-protein interaction networks to identify discriminative subnetworks associated with specific diseases [9,76-84], allowing researchers to uncover patterns of mutations that may collectively contribute to phenotypes, going beyond the limitations of single-gene analyses.

Additionally, other studies exemplify the power of this methodology in understanding the molecular basis of point mutation-associated diseases like cystic fibrosis and sickle cell disease (SCD). Previously, we developed a hierarchical graph-theoretic model to investigate the effects of point mutations on the NBD2 domain of the CFTR protein, which is implicated in cystic fibrosis [60,85]. By constructing a multilevel graph (nested graph) representation of interacting amino acid residues, they developed a nested graph capable of quantifying both local and global structural changes resulting from virtual point mutations. This innovative method enabled the differentiation of mild mutations (such as Y1219G and G1271E) from severe ones (like N1303K) when compared to the wildtype, demonstrating the sensitivity and relevance of their graph-theoretic methods and resulting molecular descriptors (graph invariants) in analyzing complex biological networks. Building on this framework, Netsey et al [61] applied similar graph-theoretic techniques to explore the impact of point mutations on the hemoglobin protein (1A3N) in SCD. Using author-adopted and molecular descriptors from our previous works [60,85], the authors successfully captured the structural effects of various mutations, including E6V, V23I, and K82N associated with SCD. Their analysis not only distinguished mild mutations from the wildtype but also highlighted the significant devastation caused by the severe E6V mutation, further

validating the usability of graph-theoretic molecular nested graphs in understanding many disease mechanisms. These studies collectively demonstrate the power of graph-theoretic modeling in bridging the gap between genetic mutations and their phenotypic manifestations in complex diseases. In recent years, the fields of ab initio modeling and molecular dynamics (MD) simulations have revolutionized our approach to studying intricate biological systems and networks [86-89]. These advanced computational methods enable researchers to create detailed models of biological entities, ranging from individual proteins to entire cellular structures, with unprecedented atomic-level precision [90]. By combining these techniques with systems biology principles, researchers can now explore the complex relationships between genetic alterations and disease manifestations. These sophisticated MD simulations provide a unique window into the molecular consequences of mutations, allowing scientists to track cascading effects from the smallest cellular components up to organism-wide changes [86-89]. This multitiered modeling strategy offers valuable insights into disease mechanisms, illuminating how specific genetic changes can influence protein behavior, disrupt cellular functions, and ultimately result in diverse clinical phenotypes. The fusion of structural data with network analysis significantly enhances our ability to predict and understand the connections between genetic makeup and disease outcomes. This integrated approach provides a more holistic view of the intricate relationships between genotypes and phenotypes in complex disorders, paving the way for more targeted and effective therapeutic interventions [91]. By providing a computational framework for quantifying structural changes at multiple levels, this approach offers a promising avenue for predicting disease severity and potentially informing therapeutic strategies.

Goal of This Study

In this study, we hypothesized that specific point mutations in the SARS-CoV-2 spike protein significantly alter its structural conformation, dynamic behavior, and stability, thereby influencing its binding affinity, immune escape potential, and overall viral fitness. To test this hypothesis, we employed a multifaceted computational approach combining graph-theoretic modeling with ab initio protein structure prediction and MD simulations to create a comprehensive analysis framework. Using the Iterative Threading Assembly Refinement (I-TASSER) [92] platform, we generated detailed 3D models of mutated spike proteins, which allowed us to visualize how point mutations affect protein structure, including potential changes in binding sites and overall conformation. We then conducted MD simulations to study the dynamic behavior of these mutated proteins, revealing how mutations impact protein movement and stability compared to the wildtype. By integrating these computational methods with our graph-based approach, we developed a holistic understanding of mutation effects, bridging the gap between genetic changes and their structural consequences. This comprehensive approach provided insights into both sequence-level alterations and their macromolecular impacts, enhancing our understanding of SARS-CoV-2's evolution and potential behavioral changes. Through the use of these complementary methods, we were able to construct a more

complete picture of how spike protein mutations influence the virus's structure and function.

Methods

Graph-Theoretic Model of SARS-CoV-2 Spike RBD (Chain E)

This work extends our previous graph-theoretic modeling approaches [60,85] by focusing specifically on the chain E of the SARS-CoV-2 spike protein and expanding combinatorial descriptors to include edge-weight assignments, among others. In this study, we focused exclusively on the spike protein, and

human ACE2 (hACE2) was not included as part of the analysis. The study incorporates molecular indices from earlier research [60,85] and introduces additional graph invariants. As a result, the molecular descriptors for subdomain graphs, the chain E of the SARS-CoV-2 spike protein, and the examined mutations are tailored to chain E of the SARS-CoV-2 spike RBD protein.

Subsequence Partition of SARS-CoV-2 Spike Protein Chain E

Figure 1 shows the tertiary crystal structure of SARS-CoV-2 spike RBD bound with ACE2 [93], which was retrieved from the Protein Data Bank [94] (PDB; 6M0J) and visualized using UCSF Chimera [95].

Figure 1. Overall structure of the SARS-CoV-2 receptor-binding domain (RBD) in complex with human angiotensin-converting enzyme 2 (ACE2). The crystal structure of the SARS-CoV-2 spike protein RBD bound to human ACE2 receptor is shown in two orientations related by a 180° rotation. ACE2 (green) forms the primary attachment point for viral entry, with its N-terminal helix directly accommodating the concave surface of the viral RBD. The SARS-CoV-2 RBD consists of 2 major regions: the core structure (cyan) containing a twisted 5-stranded antiparallel β sheet (β1-β7, labeled in right panel), and the receptor-binding motif (RBM; red) that directly interfaces with ACE2. A total of 4 critical disulfide bonds are highlighted as yellow sticks and indicated by arrows: 3 within the core (C336–C361, C379–C432, and C391–C525) stabilizing the β sheet structure, and one (C480–C488) connecting loops at the distal end of the RBM. The N-terminal helix of ACE2, which serves as the primary interaction surface with the viral RBD, is specifically labeled. This binding interface is the initial critical point of contact during viral infection and serves as a major determinant of host range and transmissibility. The structure was determined by x-ray crystallography at 2.45 Å resolution (Protein Data Bank [PDB] 6M0J). This figure was adapted from Lan et al [93].
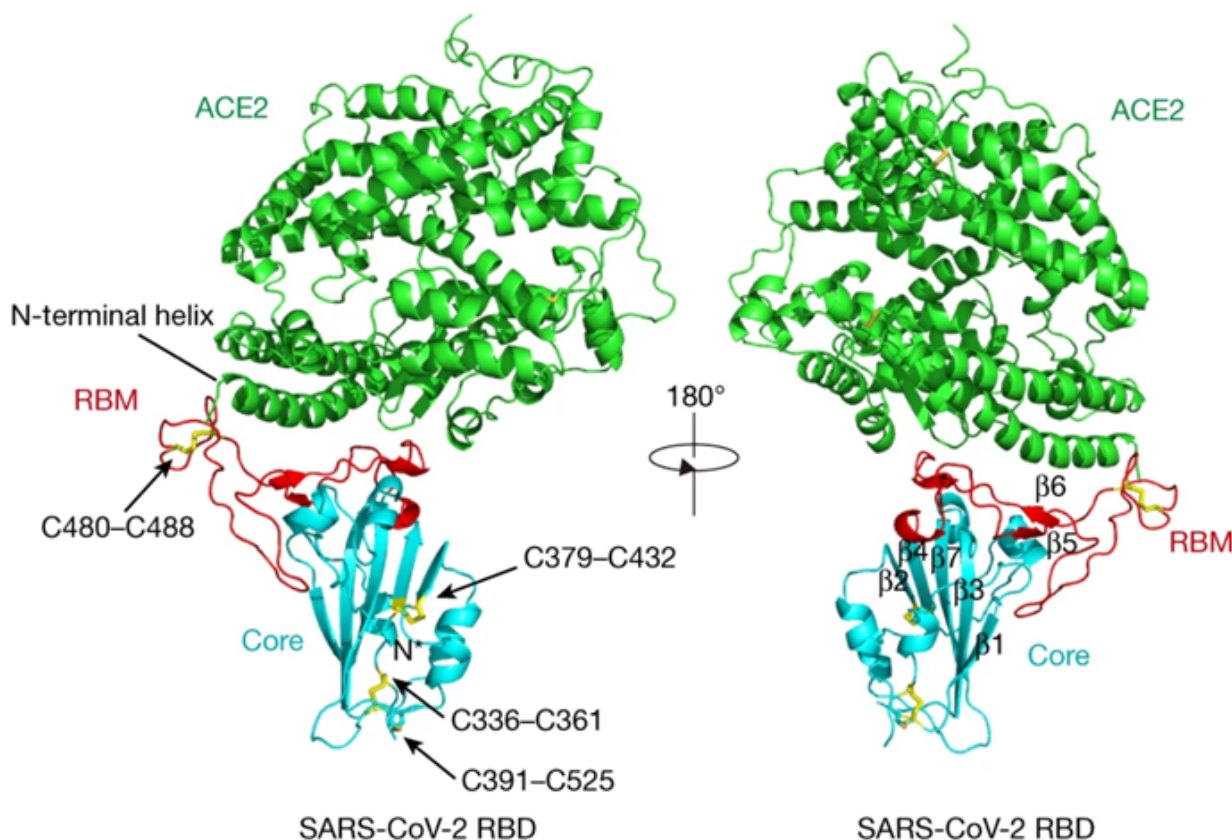
To provide a more comprehensive understanding of the SARS-CoV-2 Spike RBD structure, we generated a network graph for only the SARS-CoV-2 spike protein, without the ACE2, using Cytoscape (Cytoscape Consortium) [96]. Figure 2 illustrates this network visualization, offering valuable insights

into the RBD's intricate architecture and its potential functional significance. This graphical representation elucidates the complex relationships between various structural elements within the RBD, enhancing our understanding of its overall organization and possible mechanistic implications.

Figure 2. Network graph of SARS-CoV-2 spike receptor-binding domain (RBD). This figure illustrates a network graph representation of the SARS-CoV-2 spike protein's RBD, derived from the crystal structure 6M0J obtained from the Protein Data Bank. Each vertex in the graph represents an individual amino acid residue of the RBD. Edges connecting these vertices were established using a 6-angstrom proximity threshold, where two residues are connected if any of their atoms are within 6Å of each other. This representation provides a comprehensive view of the RBD structure, from individual amino acids to their network interactions.
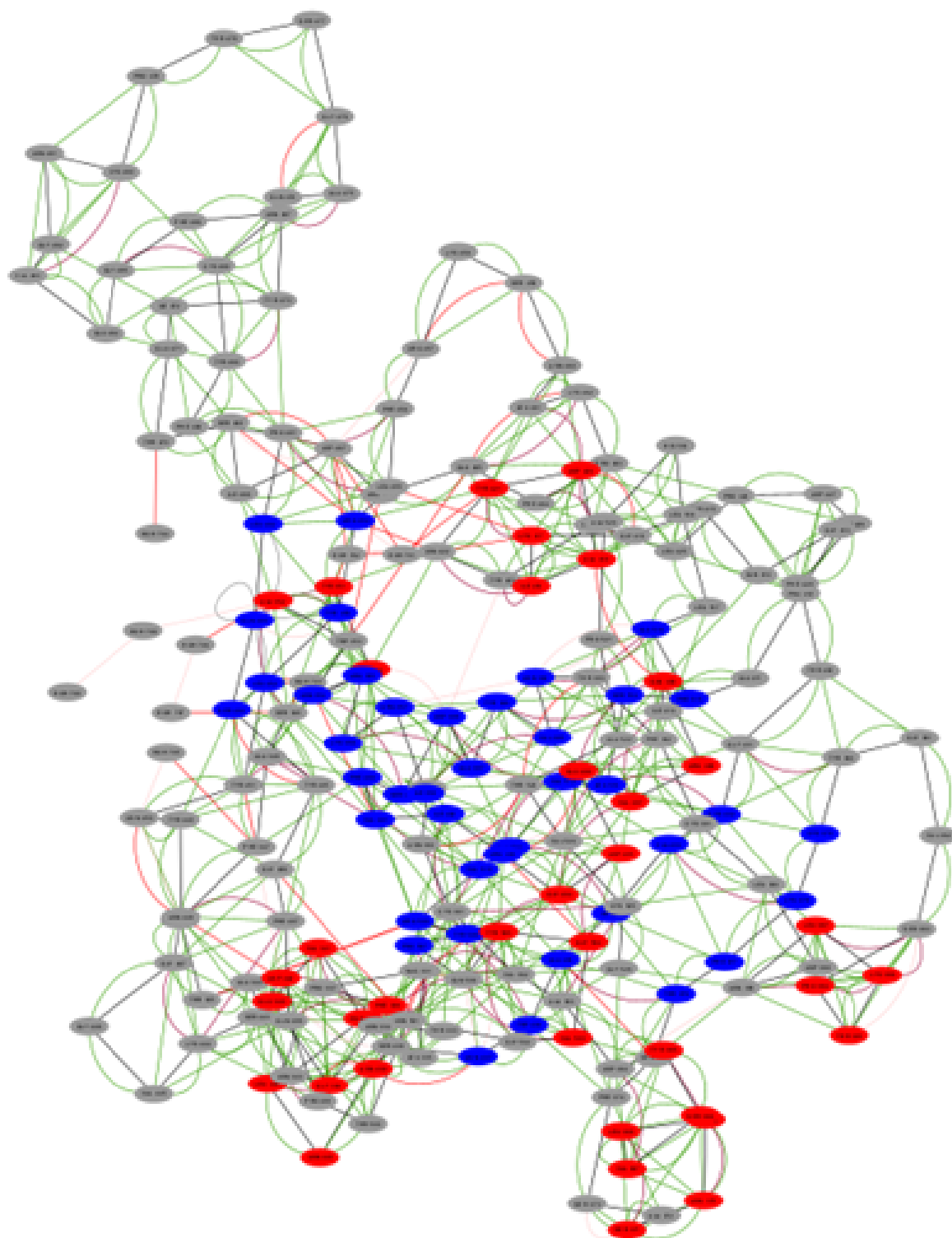
The spike protein RBD, corresponding to chain E of the structure [93], was meticulously divided into 19 subsequences (G1 to G19), while preserving the integrity of crucial biological information within the protein's secondary structures. This partitioning approach, following our previously published method [60,85], enables a more detailed analysis of the RBD's structure, which in turn informs the graph-theoretic modeling of the subsequence. During the partitioning for analyzing the SARS-CoV-2 RBD (chain E), we followed 3 key principles to maintain structural integrity and facilitate detailed analysis. We preserved binding sites and secondary structures, isolated different structural elements into separate subsequences, and

limited each subsequence to 13 amino acid residues. To account for protein complexity, loop regions were given more flexibility, allowing for the inclusion of turns, 3/10-helices, and short alpha helices. This careful approach provided a comprehensive view of the RBD's structure, enabling more in-depth analysis and modeling. The core structure of the SARS-CoV-2 RBD consists of a twisted 5-stranded antiparallel β sheet (β1, β2, β3, β4, and β7), interconnected by short helices and loops. This detailed

structural breakdown sets the foundation for further study of the RBD's composition and function [93,97]. Table 1 provides a comprehensive overview of the subsequence partition, including subsequence identifiers (G1 to G19), corresponding amino acid residues, secondary structure classification, corresponding subdomains, and reason for the partitions. This detailed structural breakdown facilitates in-depth analysis of the SARS-CoV-2 spike RBD (chain E).

Table . Subsequence partition of SARS-CoV-2 spike receptor-binding domain (chain E).

Subdomain graph	Subsequence	Amino acid sequence	Structural or functional regions
G1	TNLCP	333 - 337	Coil and turn
G2	FGEVFNA	338 - 344	Alpha helix, turn, genome variant site (339), and mutagenesis (343)
G3	TRFASVYA	345 - 352	Alpha helix, coil, and genome variant site (346)
G4	WNRKRISNCV	353 - 362	Beta sheet and coil
G5	ADYSVLYNSAS	363 - 373	Alpha helix, coil, and genome variant site (371, 373)
G6	FSTFKCYG	374 - 381	Beta sheet, coil, and genome variant site (375)
G7	VSPTKLNDLCF	382 - 392	Coil and alpha helix
G8	TNVYADSFVIR	393 - 403	Beta sheet and coil
G9	GDEVRQIAPG	404 - 413	Alpha helix and coil
G10	QTGKIADYNYKLPDD	414 - 428	Alpha helix, coil, and genome variant site (417)
G11	FTGCVIAWNS	429 - 438	Beta sheet and coil
G12	NNLDSKVGGNY	439 - 449	Alpha helix, coil, turn, and genome variant site (440, 446)
G13	NYLYRLFRK	450 - 458	Beta sheet, coil, genome variant site (452, 453), and mutagenesis (452, 453)
G14	SNLKPFERDISTEIY	459 - 473	Coil and turns
G15	QAGSTPCNGVEGFN	474 - 487	Coil, turns, genome variant site (477, 478, 484), and mutagenesis (475, 483)
G16	CYFPLQSYGF	488 - 497	Beta sheet, coil, genome variant site (490, 493, 496), and mutagenesis (490, 493)
G17	QPTNGVGYQ	498 - 506	Alpha helix, coil, genome variant site (501, 505), and mutagenesis (501)
G18	PYRVVVLSFELLHA	507 - 520	Beta sheet, coil, and mutagenesis (519)
G19	PATVCG	521 - 526	Coil

Subdomain Graphs (Corresponding to Subsequence Partitions) of SARS-CoV-2 Spike Protein Chain E

After partitioning into subsequences, as shown in Table 1, we applied a sophisticated multistep graph-theoretic modeling approach. Using I-TASSER [92,98], a state-of-the-art protein structure prediction tool, we generated ab initio models for each subsequence, using a proximity threshold of 6 angstroms and

determining end points based on each amino acid residue's center of mass. These structural predictions were visualized using Cytoscape [96], a powerful network visualization software, resulting in 19 comprehensive subdomain graphs depicted in Figures 3-6 for all subdomain graphs. Each subdomain graph, corresponding to a previously classified subsequence in Table 1, offers a detailed visual representation of the structural and interaction patterns within each RBD subdomain.

XSL•FO
RenderX

Figure 3. Nested interaction graphs (G1-G6) derived from ab initio modeling of receptor-binding domain subsequences using Iterative Threading Assembly Refinement (I-TASSER). Graphs represent local structural context of amino acids, with nodes as residues and edges indicating spatial proximity (≤6Å between residue centers of mass). (A) Subdomain graph G1 corresponding to subsequence S1. (B) Subdomain graph G2 corresponding to subsequence S2. (C) Subdomain graph G3 corresponding to subsequence S3. (D) Subdomain graph G4 corresponding to subsequence S4. (E) Subdomain graph G5 corresponding to subsequence S5. (F) Subdomain graph G6 corresponding to subsequence S6.
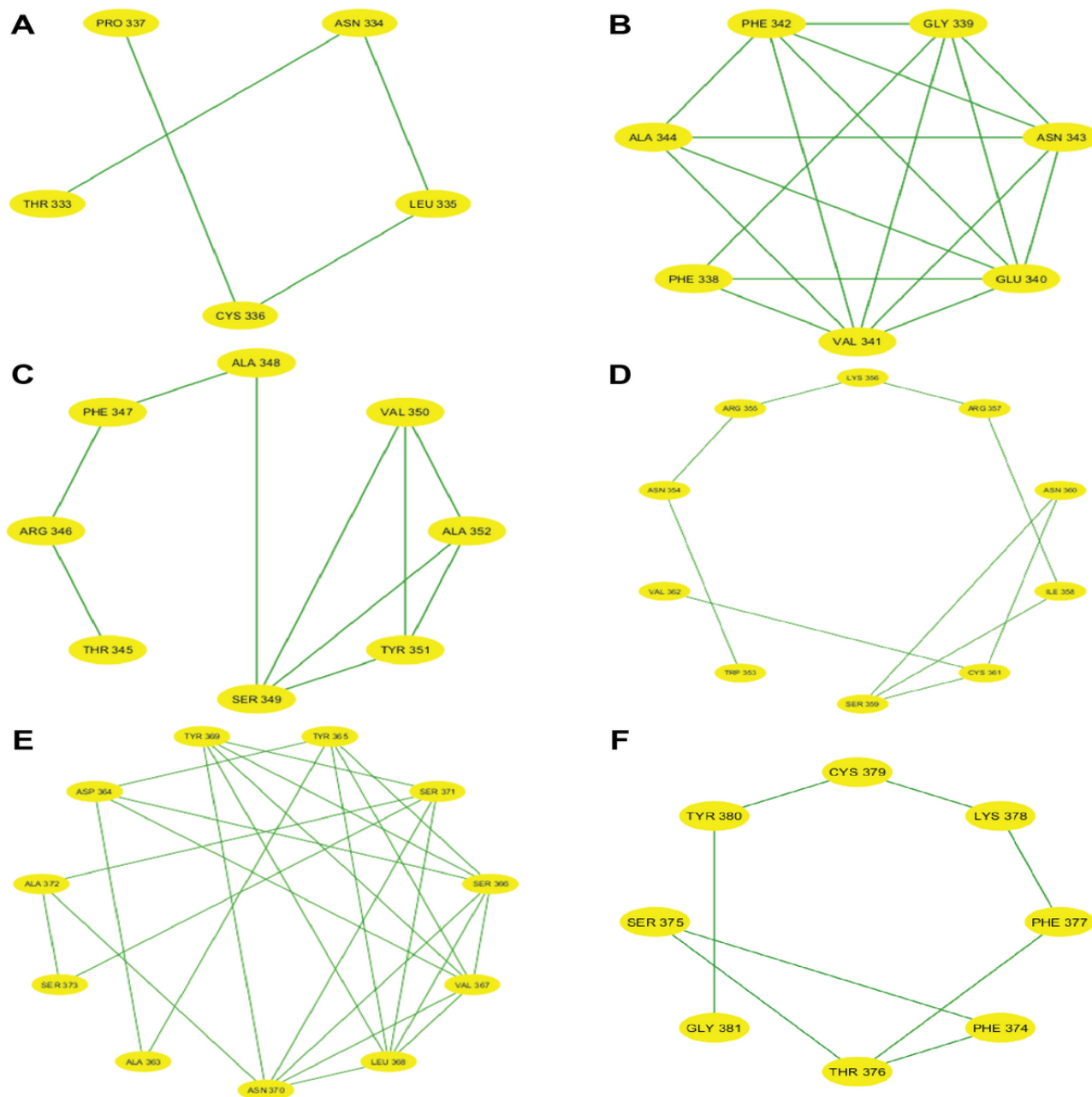
Figure 4. Nested interaction graphs (G7-G12) derived from ab initio modeling of receptor-binding domain subsequences using Iterative Threading Assembly Refinement (I-TASSER). Graphs represent local structural context of amino acids, with nodes as residues and edges indicating spatial proximity (≤6Å between residue centers of mass). (A) Subdomain graph G7 corresponding to subsequence S7. (B) Subdomain graph G8 corresponding to subsequence S8. (C) Subdomain graph G9 corresponding to subsequence S9. (D) Subdomain graph G10 corresponding to subsequence S10. (E) Subdomain graph G11 corresponding to subsequence S11. (F) Subdomain graph G12 corresponding to subsequence S12.
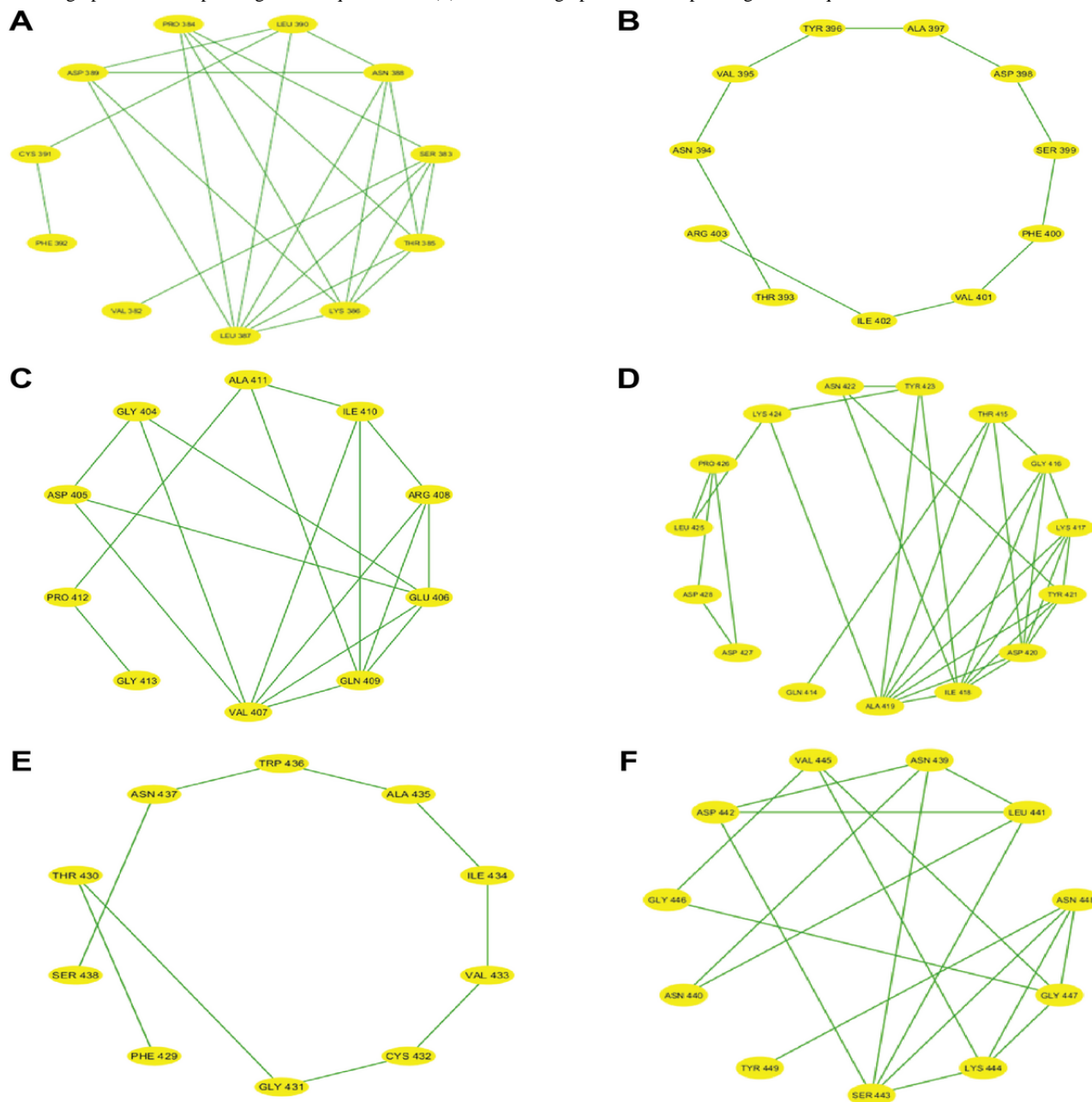
Figure 5. Nested interaction graphs (G13-G18) derived from ab initio modeling of receptor-binding domain subsequences using Iterative Threading Assembly Refinement (I-TASSER). Graphs represent local structural context of amino acids, with nodes as residues and edges indicating spatial proximity (≤6Å between residue centers of mass). (A) Subdomain graph G13 corresponding to subsequence S13. (B) Subdomain graph G14 corresponding to subsequence S14. (C) Subdomain graph G15 corresponding to subsequence S15. (D) Subdomain graph G16 corresponding to subsequence S16. (E) Subdomain graph G17 corresponding to subsequence S17. (F) Subdomain graph G18 corresponding to subsequence S18.
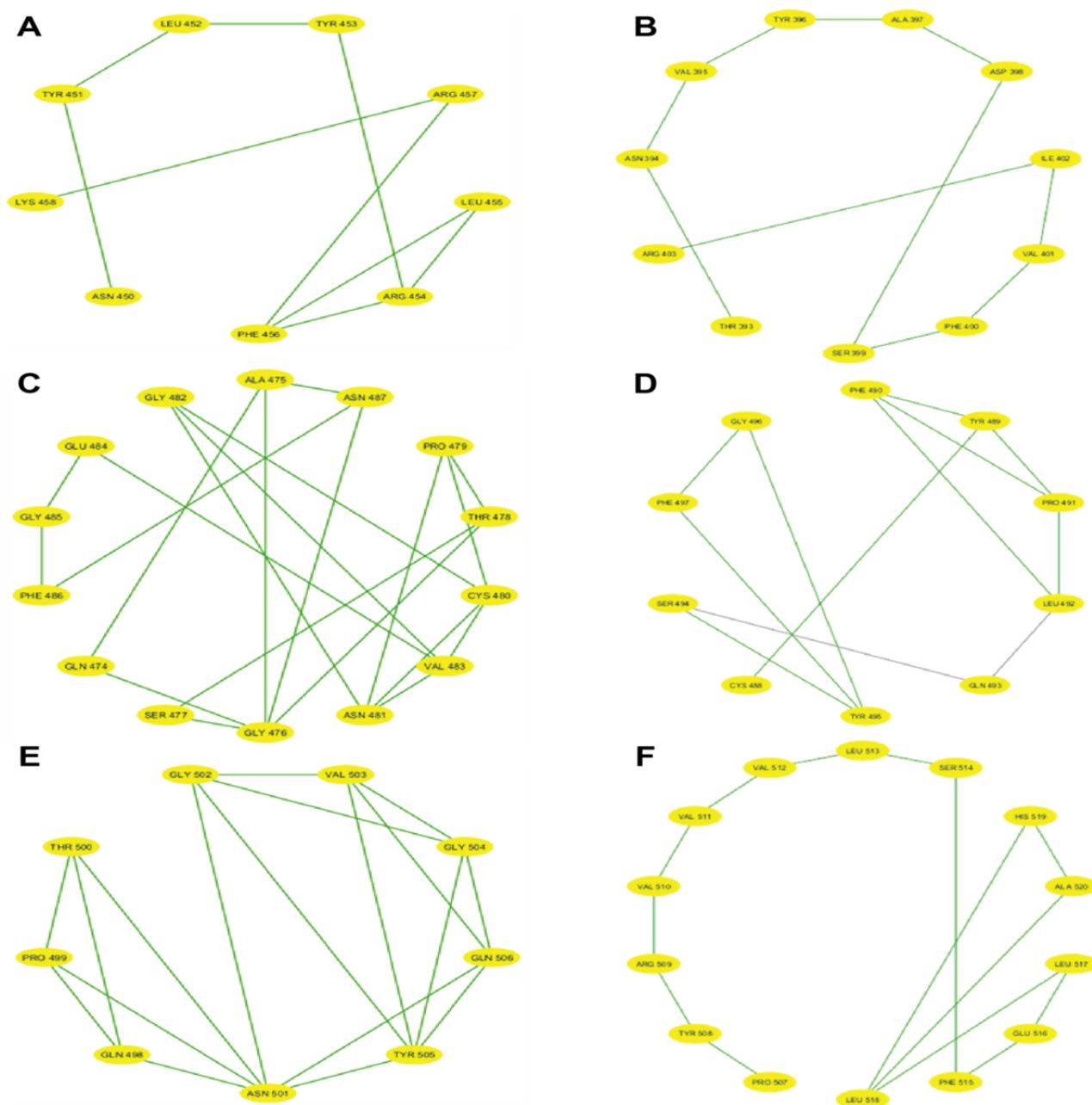
Figure 6. Nested interaction graphs (G19) derived from ab initio modeling of receptor-binding domain subsequences using Iterative Threading Assembly Refinement (I-TASSER). Graphs represent local structural context of amino acids, with nodes as residues and edges indicating spatial proximity (≤6Å between residue centers of mass). Subdomain graph G19 corresponding to subsequence S19.
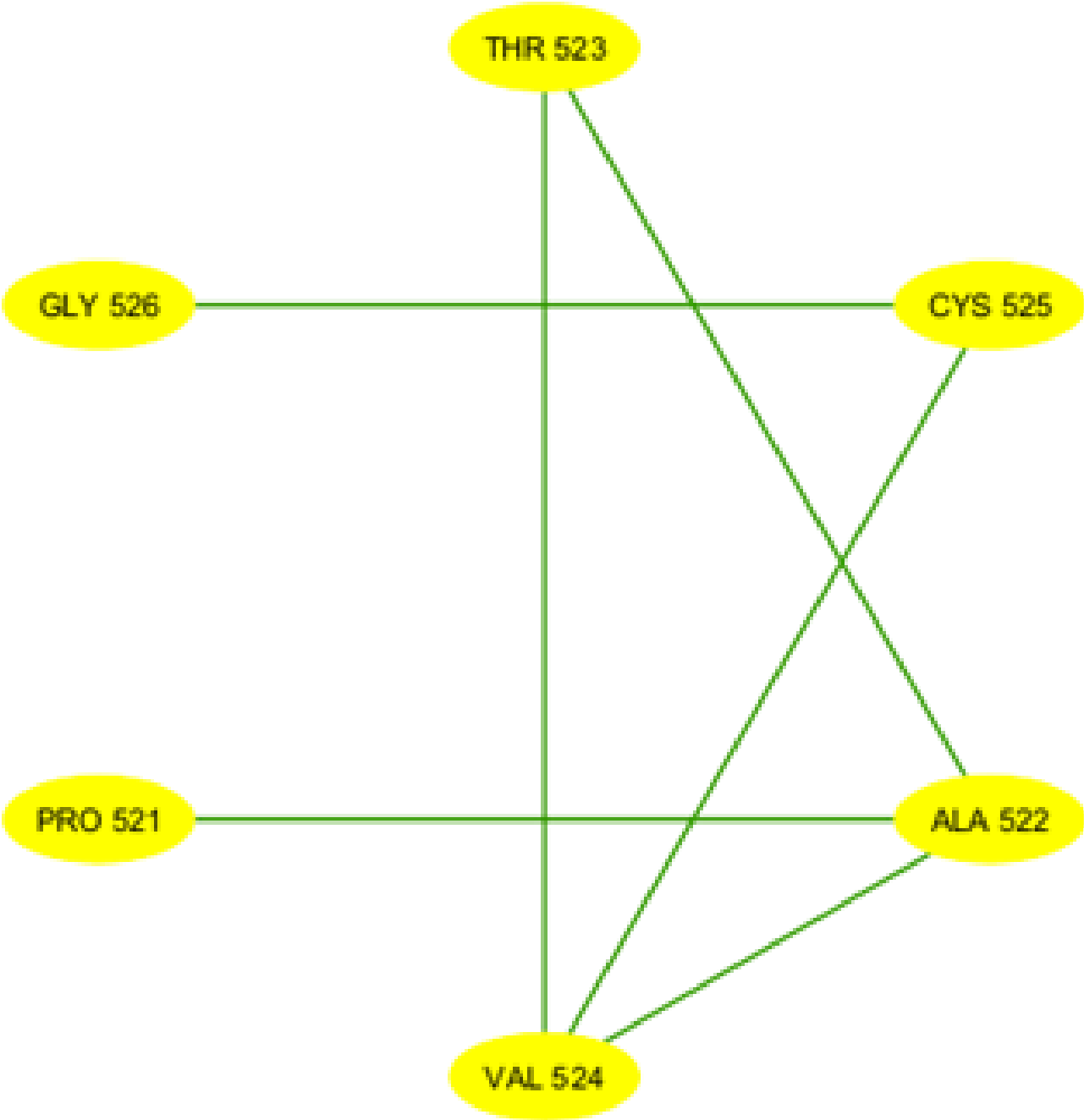
Figure 7 illustrates the subdomain graphs for subsequences G13 and G17 of the SARS-CoV-2 spike protein's RBD in chain E. These specific subdomains are of particular interest, as they contain the locations of severe mutations N501Y and L452R, respectively. The graphical representation provides a detailed view of the structural context surrounding these critical mutation sites, offering insights into how these changes might affect the protein's function and interactions.

This methodical approach not only enables a more nuanced understanding of the RBD's structure but also lays the groundwork for further analysis of how mutations might affect these interactions and, consequently, the entire protein.

Figure 7. Subdomain graphs of SARS-CoV-2 spike protein receptor-binding domain (RBD; chain E) where N501Y and L452R mutations occur. (A) Subdomain graph for G13 containing N501Y mutation. (B) Subdomain graph for G17 containing L452R mutation. Ab initio models were generated for each subsequence, employing a proximity threshold of 6 angstroms and determining end points based on each amino acid residue's center of mass. These graphs illustrate the local structural context surrounding the mutation sites, with nodes representing amino acid residues and edges indicating spatial proximity. The central nodes (N501Y in A and L452R in B) highlight the locations of the severe mutations, while surrounding nodes depict neighboring residues that may influence or be affected by these mutations. This representation aids in visualizing potential structural and functional impacts of these critical mutations on the spike protein's RBD.
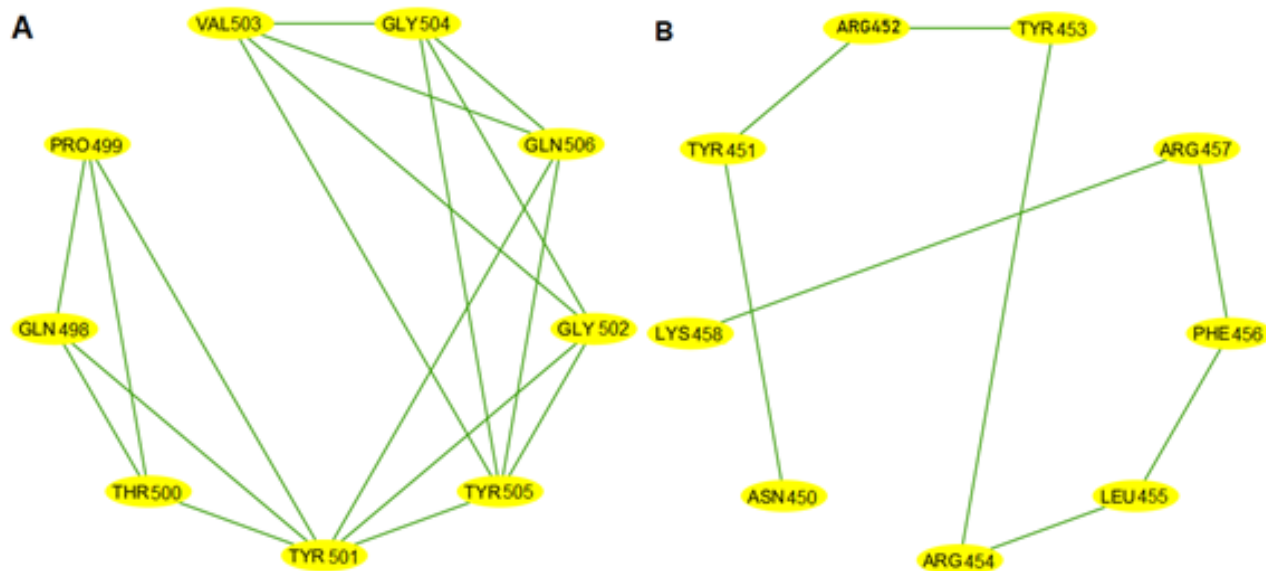
Graph-Theoretic Model of SARS-CoV-2 Spike RBD (Chain E)

Next, we modeled SARS-CoV-2 spike RBD corresponding to chain E of the structure protein to consist of 3 distinct levels, with each level offering a unique perspective on the protein's structure and interactions. At the foundation, we have the lowest level, comprising 20 vertex-weighted amino acids representing the essential building blocks of the protein. Moving up, the middle level features 19 distinct vertex-weighted subgraphs (subdomain graphs), each corresponding to a specific subsequence of SARS-CoV-2 spike RBD corresponding to chain E (Table 1 and Figures 3-7) of the structure protein. In these subgraphs or subdomain graphs, individual amino acids are represented as vertices, with weights assigned based on graph invariants derived from the lower-level graphs. The top level of our model presents a more condensed view, where each subdomain graph from the middle level is consolidated into a single weighted vertex. The weights of these vertices are determined by molecular descriptors calculated from their

respective subdomain graphs. To establish connections between vertices in the SARS-CoV-2 RBD corresponding to chain E nested domain graph, we used a proximity threshold of 6 angstroms between adjacent residues. A 6-angstrom proximity threshold is often used to define connections in the SARS-CoV-2 spike RBD graph because it captures meaningful atomic interactions, such as hydrogen bonds and van der Waals forces, essential for protein structure and function. This threshold ensures accurate representation of residue connectivity, aiding in understanding how the RBD stabilizes its structure and interacts with the ACE2 receptor [99]. In addition, it accommodates the dynamic flexibility of the RBD, reflecting both stable and transient interactions critical for receptor binding and regulation [100]. This balance of specificity and inclusivity makes it effective for modeling structural and functional relationships. Figure 8 provides a visual representation of our graph-theoretic model for the SARS-CoV-2 spike RBD corresponding to chain E, illustrating the hierarchical structure and interactions captured by this approach.

Figure 8. Hierarchical graph model of SARS-CoV-2 spike receptor-binding domain (RBD) chain E. This figure depicts a three-level hierarchical graph model of the SARS-CoV-2 spike RBD (chain E), illustrating the protein's structure and interactions at different scales. The lower level shows 20 vertex-weighted amino acids, the middle level presents 19 vertex-weighted subdomain graphs, and the top level displays a condensed view with single weighted vertices representing each subdomain. Edges between vertices are established using a 6-angstrom proximity threshold, providing a comprehensive view from individual amino acids to subdomain interactions.
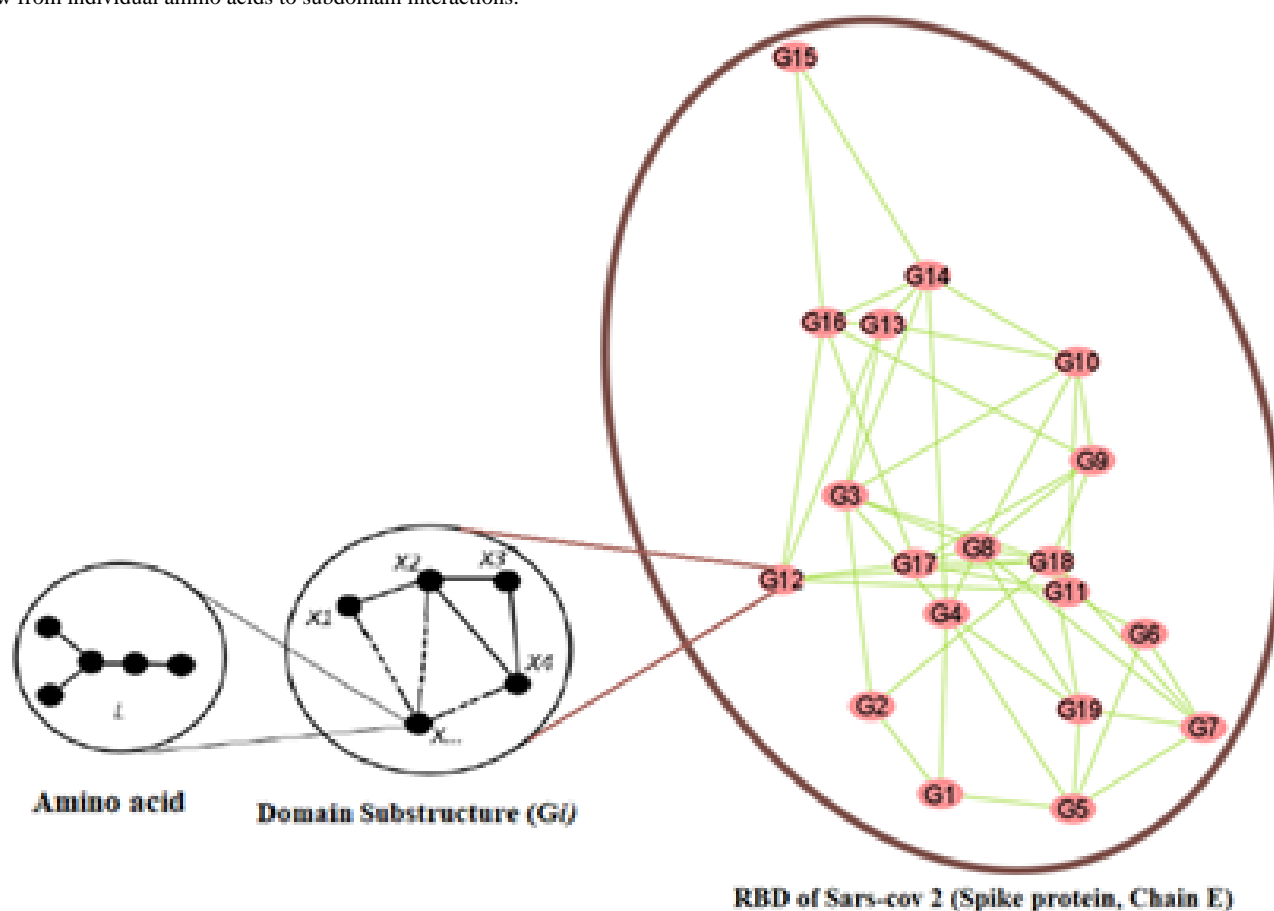
Building on our previous work [60,85], we used Cytoscape [96] to analyze the structural properties of subdomain graphs (referred to as mid-level graphs) by calculating a specific parameter: the change in molar mass between adjacent vertices normalized by the average degree of these graphs. This parameter, termed ΔMd, quantifies the variation in molar mass between two connected vertices, R_i and R_j, in the mid-level graphs, adjusted by the average connectivity (degree) of the graph. The results of this analysis are presented in Table 2 and are derived using equation 1, as established in our previous work [101]. Specifically, ΔMd represents the change in molar mass per average degree along the edge connecting vertices R_i and R_j. It is calculated using the following equation:

$$(1) \Delta Md = |R_i - R_j d^-|g/deg$$

Table . Molar mass as weighted degrees of subdomain.

Subdomain	Molar mass (g/mol)
G1	618.7
G2	890.9
G3	1040.1
G4	1437.6
G5	1369.4
G6	1078.3
G7	1416.6
G8	1464.5
G9	1203.3
G10	1993.2
G11	1259.4
G12	1360.4
G13	1416.7
G14	2064.2
G15	1614.7
G16	1386.7
G17	1107.2
G18	1877.1
G19	636.7

Here, $|R_i - R_j|$ is the absolute difference in molar mass between adjacent residues (vertices Ri and Rj), and $\bar{d}$ is the average degree of the top-level graph, reflecting the typical number of connections per vertex [61]. This normalization by the average degree accounts for the structural density of the graph, providing a more comparable measure across different subdomains. The parameter ΔMd is critical to our study as it offers insight into the structural and chemical heterogeneity within the midlevel graphs. By measuring the molar mass variation between adjacent vertices relative to the graph's connectivity, ΔMd helps us understand how mass distribution correlates with the network topology of the subdomains. This is particularly relevant for identifying regions of significant chemical or structural divergence, which may influence the functional properties of the system under study. For instance, a high ΔMd value could indicate a sharp transition in molecular composition across connected residues, potentially pointing to functionally important boundaries or interfaces within the system. Incorporating this analysis, therefore, supports our broader objective of mapping structural features to functional outcomes, as detailed in subsequent sections of our research.

Virtual Mutations in SARS-CoV-2 Spike RBD (Chain E)

To evaluate the impact of single point mutations on the entire SARS-CoV-2 spike RBD corresponding to chain E of the structural protein, we selected 5 prevalent mutations associated with mild or severe COVID-19 from existing literature. Table 3 provides a comprehensive view of how each mutation influences the SARS-CoV-2 spike RBD corresponding to chain E of the structural protein.

Table . Mutation phenotypes of SARS-CoV-2 spike receptor-binding domain [102-106].

Mutation	Strain	Phenotype	Impact or clinical manifestation
Wildtype	• Wuhan-Hu-1	• Typical or normal form of a species as it occurs in nature; baseline for comparison with mutated or altered forms	• No impact/baseline
K417N	• Beta (B.1.351) • Gamma (P.1) • Omicron (BA.1) • Omicron (BA.2)	• Moderate	• Immune evasion, associated with reduced neutralization by antibodies
N440K	• Omicron (BA.1) • Omicron (BA.2)	• Mild	• Potential immune escape, enhances binding to ACE2[a] receptor but not a major antigenic change
L452R	• Delta (B.1.617.2 and AY lineages) • Kappa (B.1.617.1)	• Severe	• Immune evasion and transmission, linked to resistance to neutralizing antibodies and higher infectivity
E484A	• Omicron	• Moderate	• Immune evasion, reduces neutralization by convalescent and vaccine-induced sera
N501Y	• Alpha • Beta • Omicron • Mu	• Very severe	• Increased binding affinity and immune escape, enhances ACE2 binding and is associated with immune evasion

[a]ACE2: angiotensin-converting enzyme 2.

Subsequently, the changes in molar mass between adjacent vertices per average degree (as shown in Table 2), derived from the weighted network interaction data, were assigned as vertex weights for the subdomain graphs within the top-level graph, G. This step enabled the generation of molecular descriptors (graph invariants) based on the weighted network interaction data in Cystoscape [92], representing the edge-interaction weights of the top-level graph. The resulting molecular database of graph invariants represented the wildtype graph (no mutation). For each of the 5 mutations, a virtual mutation process was performed. This involved identifying the specific amino acid in the relevant subdomain graph, mutating it via substitution or deletion, and creating a "mutant-specific vertex-weighted graph" for the affected subdomain by submitting the mutated FASTA sequence to I-TASSER [90,91] for ab initio modeling (see Multimedia Appendix 1 for mutant-specific vertex-weighted graphs). New graph-theoretic molecular descriptors were then computed for each mutated subdomain graph and applied to the top-level graph. Subsequently, the same molecular descriptors previously computed for the top-level graph of the wildtype were recalculated to capture the mutation's impact. This approach allowed us to observe both local (subdomain) and global (entire SARS-CoV-2 spike RBD, corresponding to chain E of the structural protein) effects of each point mutation on the SARS-CoV-2 spike protein. Table 4 presents a comprehensive set of molecular descriptors for the wildtype SARS-CoV-2 spike RBD and its various mutations. These graph invariants offer quantitative insights into how mild and severe mutations affect the RBD's structural and functional characteristics. Through analysis of these metrics across various variants, we elucidated the distinct effects of each mutation on the spike protein's characteristics and functional implications.

Table . Edge (interaction) weights of mutation phenotypes for SARS-CoV-2 spike receptor-binding domain (chain E).

Subdomain interaction	Wildtype	K417N	N440K	L452R	N501Y	E484A
G14 (meta) G15	0.08990	0.08990	0.08990	0.08990	0.08990	0.10150
G13 (meta) G14	0.12950	0.12950	0.12950	0.12090	0.12950	0.12950
G12 (meta) G13	0.01126	0.01126	0.00844	0.01986	0.01126	0.01126
G11 (meta) G12	0.02020	0.02020	0.02302	0.02020	0.02020	0.02020
G7 (meta) G18	0.09210	0.09210	0.09210	0.09210	0.09210	0.09210
G2 (meta) G18	0.19724	0.19724	0.19724	0.19724	0.19724	0.19724
G11 (meta) G18	0.12354	0.12354	0.12354	0.12354	0.12354	0.12354
G12 (meta) G18	0.10334	0.10334	0.10052	0.10334	0.10334	0.10334
G9 (meta) G18	0.13476	0.13476	0.13476	0.13476	0.13476	0.13476
G4 (meta) G14	0.12532	0.12532	0.12532	0.12532	0.12532	0.12532
G3 (meta) G13	0.07532	0.07532	0.07532	0.08392	0.07532	0.07532
G3 (meta) G18	0.16740	0.16740	0.16740	0.16740	0.16740	0.16740
G3 (meta) G4	0.07950	0.07950	0.07950	0.07950	0.07950	0.07950
G2 (meta) G3	0.02984	0.02984	0.02984	0.02984	0.02984	0.02984
G3 (meta) G14	0.20482	0.20482	0.20482	0.20482	0.20482	0.20482
G11 (meta) G17	0.03044	0.03044	0.03044	0.03044	0.02062	0.03044
G12 (meta) G17	0.05064	0.05064	0.05346	0.05064	0.04082	0.05064
G9 (meta) G17	0.01922	0.01922	0.01922	0.01922	0.00940	0.01922
G17 (meta) G18	0.15398	0.15398	0.15398	0.15398	0.14416	0.15398
G8 (meta) G9	0.05224	0.05224	0.05224	0.05224	0.05224	0.05224
G8 (meta) G18	0.08252	0.08252	0.08252	0.08252	0.08252	0.08252
G3 (meta) G8	0.08488	0.08488	0.08488	0.08488	0.08488	0.08488
G4 (meta) G8	0.00538	0.00538	0.00538	0.00538	0.00538	0.00538
G8 (meta) G17	0.07146	0.07146	0.07146	0.07146	0.06164	0.07146
G8 (meta) G7	0.00958	0.00958	0.00958	0.00958	0.00958	0.00958
G6 (meta) G7	0.06766	0.06766	0.06766	0.06766	0.06766	0.06766
G6 (meta) G11	0.03622	0.03622	0.03622	0.03622	0.03622	0.03622
G8 (meta) G19	0.16556	0.16556	0.16556	0.16556	0.16556	0.16556
G7 (meta) G19	0.15598	0.15598	0.15598	0.15598	0.15598	0.15598
G4 (meta) G19	0.16018	0.16018	0.16018	0.16018	0.16018	0.16018
G19 (meta) G18	0.24808	0.24808	0.24808	0.24808	0.24808	0.24808
G1 (meta) G2	0.05444	0.05444	0.05444	0.05444	0.05444	0.05444
G1 (meta) G4	0.16378	0.16378	0.16378	0.16378	0.16378	0.16378
G5 (meta) G19	0.14654	0.14654	0.14654	0.14654	0.14654	0.14654
G5 (meta) G4	0.01364	0.01364	0.01364	0.01364	0.01364	0.01364
G1 (meta) G5	0.15014	0.15014	0.15014	0.15014	0.15014	0.15014
G5 (meta) G6	0.05822	0.05822	0.05822	0.05822	0.05822	0.05822
G5 (meta) G7	0.00944	0.00944	0.00944	0.00944	0.00944	0.00944
G16 (meta) G15	0.04560	0.04560	0.04560	0.04560	0.04560	0.03400
G12 (meta) G16	0.00526	0.00526	0.00244	0.00526	0.00526	0.00526
G13 (meta) G16	0.00600	0.00600	0.00600	0.01460	0.00600	0.00600

Subdomain interaction	Wildtype	K417N	N440K	L452R	N501Y	E484A
G16 (meta) G17	0.05590	0.05590	0.05590	0.05590	0.04608	0.05590
G14 (meta) G16	0.13550	0.13550	0.13550	0.13550	0.13550	0.13550
G9 (meta) G16	0.03668	0.03668	0.03668	0.03668	0.03668	0.03668
G8 (meta) G10	0.10574	0.10292	0.10574	0.10574	0.10574	0.10574
G9 (meta) G10	0.15798	0.15516	0.15798	0.15798	0.15798	0.15798
G10 (meta) G14	0.01420	0.01702	0.01420	0.01420	0.01420	0.01420
G10 (meta) G11	0.14676	0.14394	0.14676	0.14676	0.14676	0.14676
G3 (meta) G10	0.19062	0.18780	0.19062	0.19062	0.19062	0.19062
G10 (meta) G13	0.11530	0.11248	0.11530	0.10670	0.11530	0.11530

Unsupervised Machine Learning Analysis of Point Mutations Associated With SARS-CoV-2 Spike RBD (Chain E)

To analyze the impact of point mutations associated with SARS-CoV-2 spike RBD (chain E), we used our previous novel approach combining graph theory and machine learning [59,84,85]. Our method used graph-theoretic molecular-weighted invariants or descriptors, as detailed in Table 4, for the wildtype SARS-CoV-2 spike RBD (chain E) and mild and severe mutations associated with SCD. We applied an unsupervised machine learning technique, specifically hierarchical clustering, to visualize the variations between each SARS-CoV-2 spike RBD (chain E) point mutation and the wildtype SARS-CoV-2 spike RBD (chain E). Hierarchical clustering provided insights into the structure and relationships within the datasets. The analysis was conducted using Python statistical software (Python Software Foundation), using the single linkage function and Euclidean distance without setting a predefined number of clusters or distance threshold to generate a dendrogram for the SARS-CoV-2 spike RBD mutation phenotypes. We opted for the single-linkage function to minimize potential biases in clustering the SARS-CoV-2 spike RBD (chain E)–related point mutations [100,107]. This approach allowed us to create a visual representation of how virtual SARS-CoV-2 spike RBD (chain E)–related point mutations affected the entire SARS-CoV-2 spike RBD (chain E) structure in comparison to the wildtype. Through this approach, we were able to distinguish how the different virtual mutations differed from the wildtype, providing an insight into how the point mutations impacted the entire SARS-CoV-2 spike RBD.

Ab Initio Modeling of Mutated Protein Sequences Using I-TASSER

To further elucidate the structural and functional implications of spike protein mutations, we implemented an ab initio modeling approach using the I-TASSER platform [90]. This state-of-the-art predictive modeling tool was used to generate high-resolution, 3D structural models of the spike protein variants [90,91,101]. The I-TASSER algorithm uses a hierarchical approach, combining threading, fragment assembly, and atomic-level structure refinement to predict protein structure and function. We input the mutated spike protein sequences into the I-TASSER server [92], which then produced detailed structural models. These models were analyzed to identify potential alterations in protein folding, binding site configurations, and overall conformational changes resulting from the point mutations. The visual representations derived from this process provided crucial insights into the molecular-level effects of the mutations, complementing our graph-theoretic modeling approach. By integrating these computational methodologies, we were able to establish a comprehensive framework for understanding the relationship between sequence-level mutations and their macromolecular consequences, offering a multifaceted view of the spike protein's structural and functional adaptations.

MD Simulations of Wildtype and Mutated Proteins in Water

To complement our findings from graph-theoretic modeling, we investigated whether MD simulations could effectively replicate the effects of point mutations leading to diverse phenotypes among the COVID-19 mutations analyzed in this study. MD simulations have emerged as a critical tool for elucidating the structural and functional implications of protein mutations, specifically in the context of the SARS-CoV-2 spike protein. In this work, we focused exclusively on the spike protein, both in its wildtype form and mutated variants, without the inclusion of the hACE2 receptor or other interacting proteins. This approach allowed us to isolate the intrinsic dynamic behavior and stability of the spike protein under varying mutational conditions. The dynamic insights gained from these simulations have proven instrumental in designing stabilized S2 immunogens for SARS-CoV-2, which demonstrate enhanced protein expression, superior thermostability, and preserved immunogenicity against sarbecoviruses.

The MD simulations were conducted using the WebGRO for Macromolecular Simulations server [108], which operates on the GROMACS simulation package [102]. This platform was selected for its intuitive interface and capability to perform fully solvated MD simulations. The wildtype spike protein structures were obtained from the PDB, while mutated variants were generated using I-TASSER [90]. Both wildtype and mutated structures were prepared in orthorhombic simulation boxes, solvated with simple point charge (SPC) water, counterions, and 0.15 M NaCl to mimic physiological conditions. This setup was designed to closely replicate the biological environment,

ensuring accurate observations of protein behavior in a dynamic state.

Before the production runs, energy minimization was carried out using the steepest descent integrator to eliminate steric clashes or unfavorable contacts within the system. This step is essential to position the system at a local energy minimum, a prerequisite for stable MD simulations. Subsequently, the system was equilibrated at 300 K and 1.1023 bar, reflecting the temperature and pressure conditions of the human body [103,109]. Position restraints on protein atoms were applied during the initial equilibration phase to stabilize the system before full MD production runs were initiated.

To ensure the robustness and reliability of our findings, triplicate simulations were performed at 3 distinct timescales: 50 nanoseconds, 100 nanoseconds, and 200 nanoseconds. Each simulation incorporated varied random sampling seed inputs to account for stochastic variations in the system's dynamics. This methodology enabled the capture of both short-term fluctuations and long-term stability trends, providing a comprehensive view of the dynamic behavior of the spike protein over extended periods.

Protein interactions were modeled using the Optimized Potentials for Liquid Simulations–All Atom (OPLS-AA) force field, recognized for its precision in simulating protein-water interactions. The SPC/extended (SPC/E) water model was chosen for its ability to accurately represent the properties of water, particularly its dielectric constant and density, which are critical for realistic solvation dynamics. This combination of force field and water model ensured a high-fidelity representation of the protein's environment during simulations.

Postsimulation analyses were conducted to evaluate the stability and conformational changes of the spike protein trajectories. Root mean square deviation (RMSD) was calculated to assess trajectory stability and deviations in atomic positions over time. RMSD plots provided visual insights into how much the protein structure deviated from its initial configuration, shedding light on conformational stability. Statistical metrics, including minimum and maximum RMSD values, SDs, and SEs, were computed to compare the stability between wildtype and mutated variants. In addition, the Kolmogorov-Smirnov test was applied to determine whether mutations significantly altered the distribution of protein conformations relative to the wildtype. Visualization and detailed analysis of trajectories were performed using VMD [104,110] and BIOVIA Discovery Studio (Dassault Systèmes BIOVIA, Discovery Studio Modeling Environment, Release 2017; Dassault Systèmes), enabling an in-depth examination of conformational changes and protein-water interactions. The simulations revealed dynamic insights into how mutations influence the stability of the spike protein, particularly with respect to conformational flexibility and hydration patterns around active sites, as observed through bundled SPC water models. These differences in hydration could potentially impact enzymatic activity or antigenicity.

The methodologies used in these MD simulations, from system setup to detailed postsimulation analysis, highlight the value of computational approaches in complementing experimental data. By focusing solely on the spike protein and excluding interactions with hACE2, we obtained a clearer understanding of the intrinsic effects of mutations on protein structure and dynamics. These insights are vital for structure-based vaccine design, as understanding the dynamic behavior of the spike protein can guide the development of vaccines targeting stable, immunogenic conformations. Ultimately, the MD simulations adopted in this study provide a powerful lens through which the effects of mutations on protein function can be observed in a dynamic, physiological context, contributing to a deeper comprehension of how mutations in the SARS-CoV-2 spike protein may influence its behavior and interaction with the host immune system.

Results

Graph-Theoretic Modeling Reveals Impact of Point Mutations on SARS-CoV-2 Spike RBD

We developed a 3-level weighted hierarchical graph-theoretic model of the SARS-CoV-2 spike RBD corresponding to chain E (see Figure 8). The model consisted of a foundation level with 20 vertex-weighted amino acids, a middle level with 19 vertex-weighted subdomain graphs, and a top level where subdomain graphs are consolidated into single weighted vertices. Key features of the model included vertex connections established using a 6-angstrom proximity threshold, subdomain graph vertex weights based on molar mass changes between adjacent vertices per average degree, and a virtual mutation process applied for 5 specific mutations (K417N, N440K, L452R, N501Y, and E484A). Mutant-specific vertex-weighted graphs were created using I-TASSER [90] for ab initio modeling, and new graph-theoretic molecular descriptors were computed for mutated subdomains and applied to the top-level graph. This approach enabled observation of both local (subdomain) and global (entire RBD) effects of each point mutation on the SARS-CoV-2 spike protein, facilitating understanding of how point mutations lead to different COVID-19 phenotypes. Hierarchical clustering analysis of SARS-CoV-2 spike protein variants was performed in Python using interaction data imported from a CSV file (see Multimedia Appendix 1) via the pandas library [111]. The data matrix was transposed to organize protein variants as rows, and clustering was conducted with the scipy.cluster.hierarchy.linkage function [112], applying the Euclidean distance metric and single linkage method to determine pairwise similarities. Variant labels included their respective mutated phenotypes for clarity. The dendrogram was visualized using matplotlib with bold axis labels and tick marks, and x-axis labels were rotated for readability. The figure was rendered at 300 dpi.

Figure 9 displays the resulting dendrogram, constructed from graph-theoretic descriptors (see Table 4 and the Methods section), providing a visual summary of how each virtual point mutation affects the SARS-CoV-2 spike RBD (chain E) relative to the wildtype. This clustering diagram highlights the structural and functional relationships among the mutated variants and the original wildtype RBD.

Dendrogram analysis (see Figure 8) revealed varying degrees of divergence between the wildtype and the analyzed mutations based on Euclidean distance. The severe N501Y mutation,

characterized by the substitution of asparagine with tyrosine at position 501, showed the greatest divergence from the wildtype, with an approximate Euclidean distance of 0.023. The L425R mutation, involving the replacement of leucine with a positively charged arginine at position 425, also exhibited a significant difference, with a Euclidean distance of approximately 0.019. Similarly, the E484A mutation, where glutamic acid is substituted with alanine at position 484, demonstrated a

divergence of approximately 0.017 Euclidean distance from the wildtype. In contrast, the K147N mutation, which replaces lysine with asparagine at position 147 within the N-terminal domain, was closer to the wildtype but still distinct, with an approximate Euclidean distance of 0.007. These results underscore the structural and functional variability introduced by these mutations relative to the wildtype.

Figure 9. Clustering analysis of COVID-19 mutation-associated point mutations in SARS-CoV-2 receptor-binding domain. The hierarchical clustering dendrogram of SARS-CoV-2 spike protein variants based on interaction data (Table 4). Clustering was performed using the Euclidean distance metric and single linkage method. Each tip on the x-axis is labeled with the protein variant and its associated mutated phenotype: N501Y (Alpha/Beta/Gamma/Omicron), L452R (Delta/Epsilon), E484A (Omicron), T478K (Delta/Omicron), K417N (Beta/Gamma), wildtype (Reference), and N440K (Immune escape). The y-axis represents the Euclidean distance, indicating the degree of dissimilarity between variants. Both axis labels and tick marks are displayed in bold for clarity. The dendrogram reveals that N501Y and L452R form a distinct cluster, reflecting greater divergence from the other variants, while wildtype, K417N, and N440K group together, indicating higher similarity in their interaction profiles. This clustering highlights the structural and functional relationships among spike protein variants and their phenotypic classifications, providing insights into how specific mutations may influence SARS-CoV-2 spike protein network architecture. The clustering analysis was performed using Python with the scipy.cluster.hierarchy.linkage function [111], applying the Euclidean distance metric and single linkage method to determine pairwise similarities.
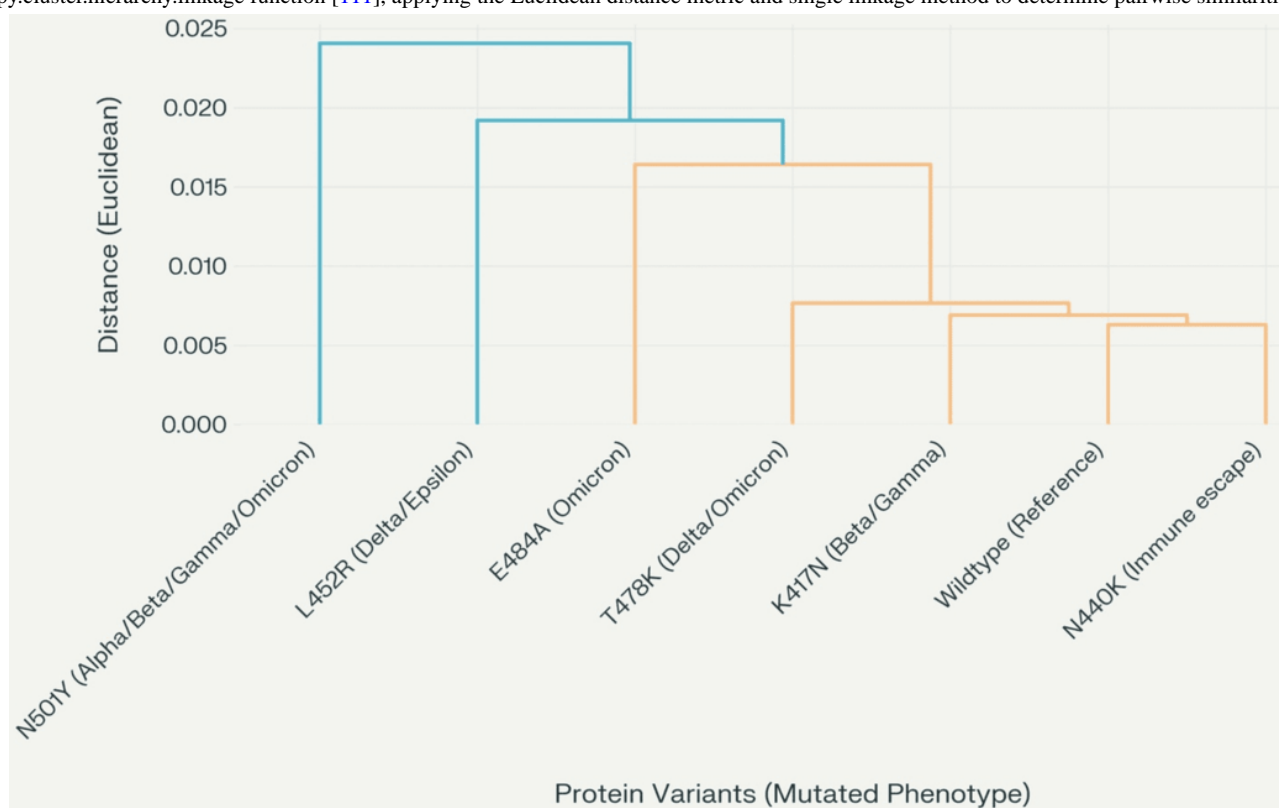
Ab Initio Models Reveal Structural Alterations in Mutated SARS-CoV-2 Spike RBD

To gain deeper insights into the impact of COVID-19–specific mutations on the SARS-CoV-2 spike RBD (chain E) protein conformation, ab initio models were generated using I-TASSER [92] and visualized in Cystoscape [96]. These models provide a comparative view of selected point mutations (N440K, E484A, N501Y, K417N, and L452R) against the wildtype structure. Figures 10-12 illustrate the structural changes induced by each mutation, revealing that even a seemingly mild mutation like E484A can significantly affect secondary structures and global protein folding. The models demonstrate alterations in protein folding patterns, changes in structural integrity, and modifications to local and global conformations. These visualizations highlight the profound impact of mutations on

the SARS-CoV-2 spike RBD, emphasizing how small changes can lead to substantial structural rearrangements. Such alterations may influence the virus's infectivity, immune evasion, and interaction with the ACE2 receptor. Notably, the N501Y mutation shows significant local and regional changes in the ab initio model, affecting protein folding and structural integrity. These modifications are more pronounced when compared to the wildtype structure, indicating that the N501Y mutation substantially impacts the RBD's conformation as shown by the ab initio modeling results in Figure 11F. The E484A mutation, despite appearing mild, demonstrates detrimental effects on secondary structures and global protein folding (see Figure 10A). Other mutations (N440K, K417N, and L452R) exhibit varying degrees of impact on the RBD structure (see Figures 10D-12), potentially influencing the spike protein's function and stability. These structural insights provide

a foundation for understanding the molecular mechanisms behind the enhanced transmissibility and potential immune evasion of SARS-CoV-2 variants, contributing to our knowledge of the virus' evolution and informing future therapeutic strategies.

Figure 10. Local and global effects of mutations E484A and L452R on the SARS-CoV-2 spike receptor-binding domain (RBD; chain E) compared to wildtype. (A) Mutated subdomain graph G15 illustrating the mild E484A mutation. (B) Ab initio model of wildtype SARS-CoV-2 spike RBD (chain E). (C) Ab initio model of SARS-CoV-2 spike RBD (chain E) following the mild E484A mutation. (D) Mutated subdomain graph G13 depicting the severe L452R mutation. (E) Ab initio model of wildtype SARS-CoV-2 spike RBD (chain E). (F) Ab initio model of SARS-CoV-2 spike RBD (chain E) after the severe L452R mutation, illustrating significant alterations in protein folding and structural integrity.
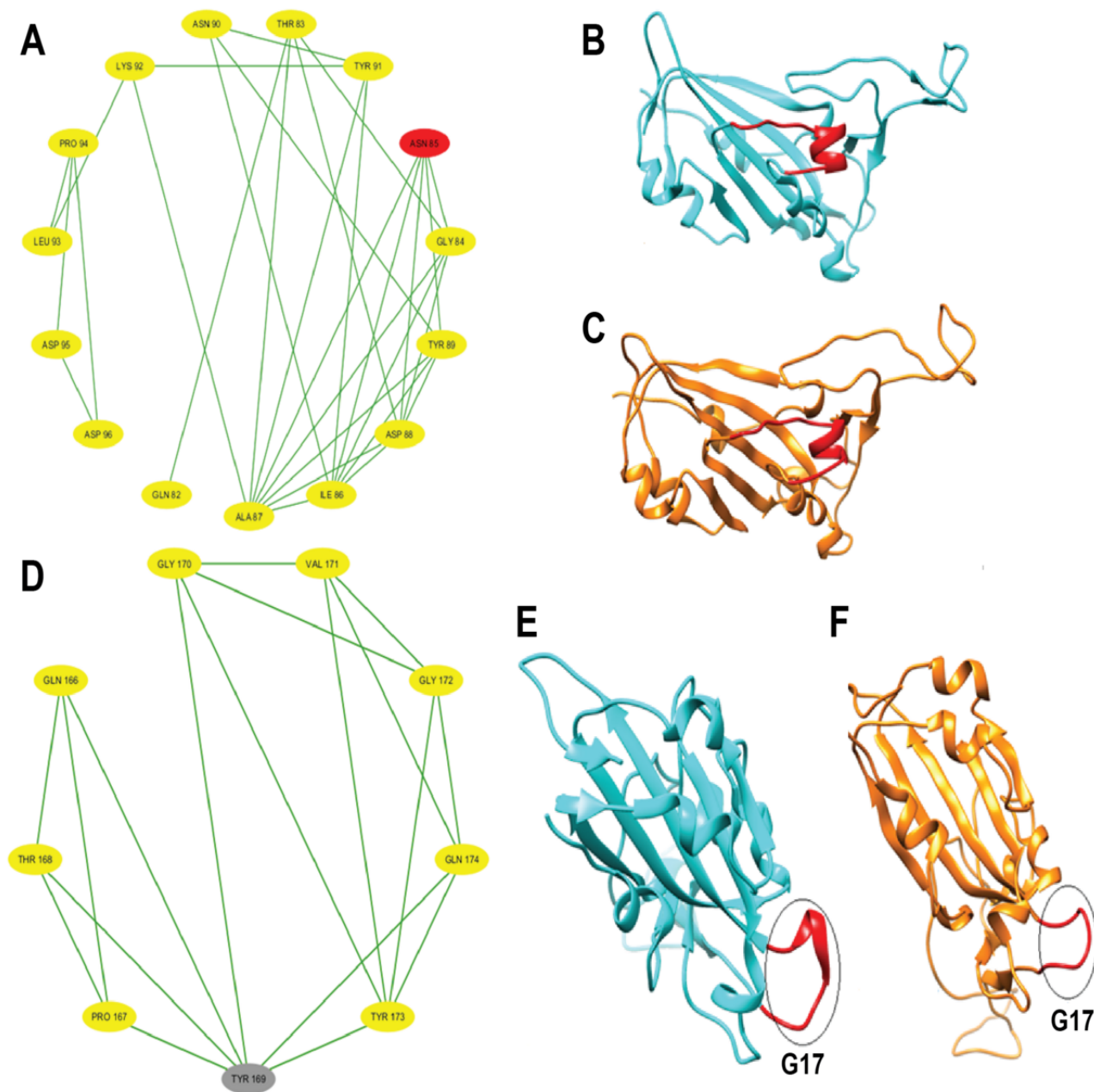
Figure 11. Local and global effects of mutations K417N and N501Y on the SARS-CoV-2 spike receptor-binding domain (RBD; chain E) compared to wildtype. (A) Mutated subdomain graph G10 illustrating the mild K417N mutation. (B) Ab initio model of wildtype SARS-CoV-2 spike RBD (chain E). (C) Ab initio model of SARS-CoV-2 spike RBD (chain E) following the mild K417N mutation. (D) Mutated subdomain graph G17 depicting the severe N501Y mutation. (E) Ab initio model of wildtype SARS-CoV-2 spike RBD (chain E). (F) Ab initio model of SARS-CoV-2 spike RBD (chain E) after the N501Y mutation, illustrating significant alterations in protein folding and structural integrity.
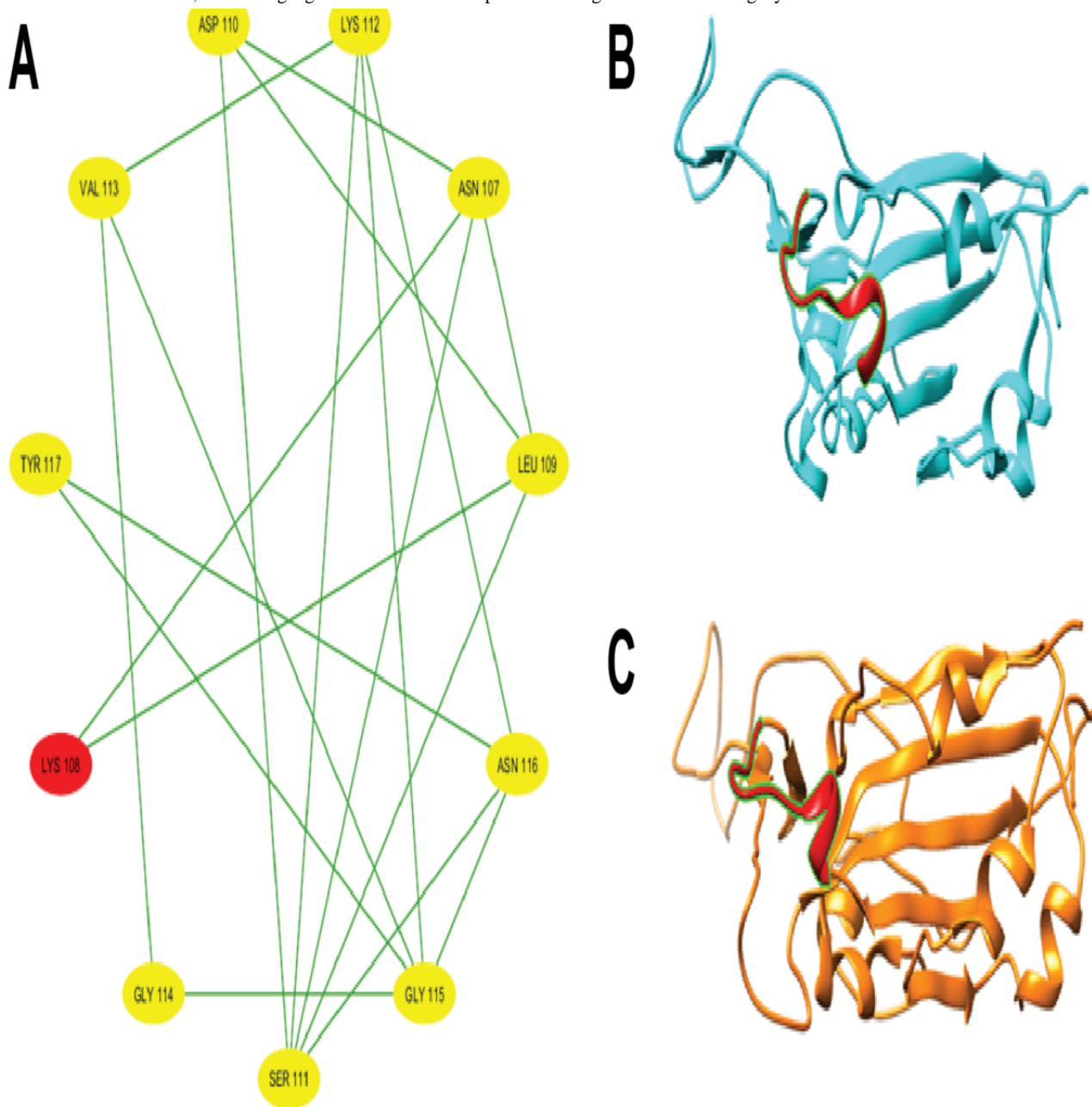
Figure 12. Local and global effects of mutation N440K on the SARS-CoV-2 spike receptor-binding domain (RBD; chain E) compared to wildtype. (A) Mutated s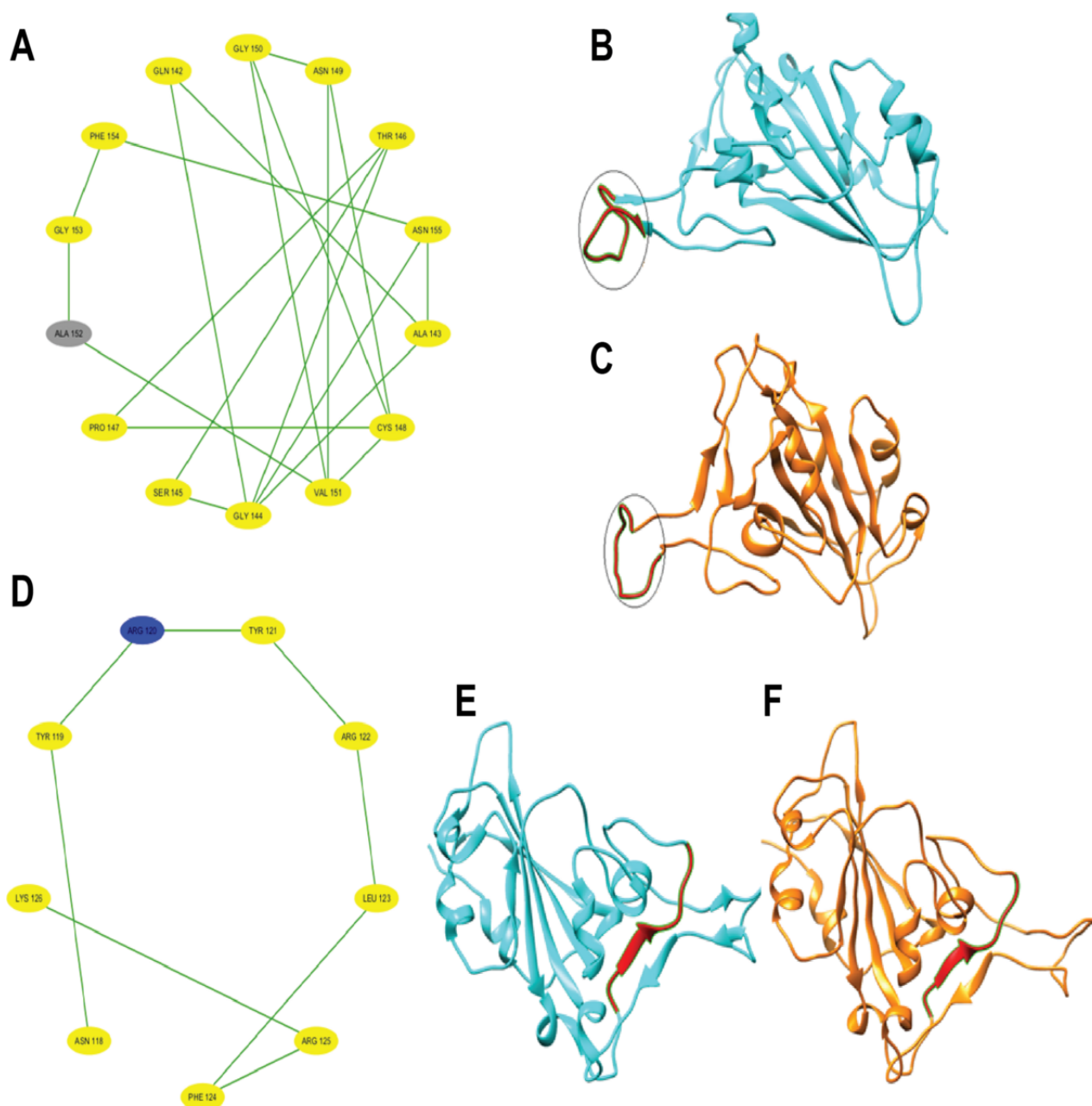ubdomain graph G12 illustrating the mild N440K mutation. (B) Ab initio model of wildtype SARS-CoV-2 spike RBD (chain E). (C) Ab initio model of SARS-CoV-2 spike RBD (chain E) following the mild N440K mutation.

MD Simulation Analysis Unveils Differential Structural Stability Between Wildtype and Mutated SARS-CoV-2 Spike Proteins

To complement our graph-theoretic and ab initio modeling findings, we conducted MD simulations to investigate the effects of point mutations on COVID-19 variants. These simulations provided insights into the binding interactions and stability of wildtype and mutated SARS-CoV-2 spike proteins in dynamic states. Our approach aligns with recent studies that have used MD simulations to examine the conformational behavior of SARS-CoV-2 spike protein variants.

We utilized the WebGRO server [108], based on GROMACS [109], to prepare modeled structures in orthorhombic simulation boxes. The proteins were solvated with SPC water, counterions, and 0.15 M NaCl to mimic physiological conditions. To maintain consistency, only the spike protein and the mutated spike proteins were used in the protein-in-water simulation studies, mirroring the approach used in the graph-theoretic and ab initio modeling. Energy minimization was performed using the steepest descent integrator, followed by equilibration at 300 K and 1.1023 bar. For robustness, we conducted triplicate 50-, 100-, and 200-nanosecond simulations with varied random sampling seed inputs. Both wildtype protein structures from PDB and mutated variants generated via I-TASSER [92,98] were simulated using the OPLS-AA force field and SPC/E water models.

Postsimulation analyses focused on RMSD plots to assess trajectory stability and atomic position deviations. Visualization and analysis of trajectories were performed using VMD and

BIOVIA Discovery Studio. This comprehensive approach enabled detailed comparisons of protein-water interactions and stability across variants, providing valuable insights into the effects of mutations on protein behavior in dynamic environments. As illustrated in Figure 13, RMSD analysis revealed differential structural stability between wildtype and mutated SARS-CoV-2 spike proteins.

Figure 13. Molecular dynamics simulations reveal timescale-dependent differential structural stability of wildtype and mutated SARS-CoV-2 spike proteins. Root mean square deviation (RMSD) analysis was performed on wildtype (black) and mutated spike proteins—E484A (green), K417N (lavender), L452R (pink), N440K (olive), and N501Y (indigo)—across short (50 ns), intermediate (100 ns), and long (200 ns) timescales. Each panel displays the RMSD trajectory for wildtype and a specific mutant at a given timescale: E484A (green) in panels (A) 50 ns (RMSD range 4.00×10^{-7} to 0.4477065; mean 0.2435, standard error of the mean [SEM] 0.000727), (B) 100 ns (RMSD range 4.00×10^{-7} to 0.4236228; mean 0.3108, SEM 0.000896), and (C) 200 ns (RMSD range 4.00×10^{-7} to 0.4477065; mean 0.2435, SEM 0.000727); K417N (lavender) in panels (D) 50 ns (RMSD range 5.00×10^{-7} to 0.3620131; mean 0.2492, SEM 0.000624), (E) 100 ns (RMSD range 4.00×10^{-7} to 0.4897364; mean 0.3362, SEM 0.000897), and (F) 200 ns (RMSD range 5.00×10^{-7} to 0.3620131; mean 0.2492, SEM 0.000624); L452R (pink) in panels (G) 50 ns (mean 0.3224, SEM 0.000984), (H) 100 ns (RMSD range 5.00×10^{-7} to 0.561134; mean 0.4151, SEM 0.000957), and (I) 200 ns (mean 0.3224, SEM 0.000984); N440K (olive) in panels (J) 50 ns (RMSD range 5.00×10^{-7} to 0.3220472; mean 0.2230, SEM 0.00048), (K) 100 ns (RMSD range 1.10×10^{-6} to 0.3744204; mean 0.2123, SEM 0.000862), and (L) 200 ns (RMSD range 4.00×10^{-7} to 0.367109; mean 0.2796, SEM 0.000353); N501Y (indigo) in panels (M) 50 ns (RMSD range 4.00×10^{-7} to 0.320323; mean 0.2279, SEM 0.000559), (N) 100 ns (RMSD range 3.00×10^{-7} to 0.3636647; mean 0.2529, SEM 0.000484), and (O) 200 ns (RMSD range 4.00×10^{-7} to 0.320323; mean 0.2279, SEM 0.000559); and wildtype (black) with RMSD values of 50 ns (RMSD range 3.00×10^{-7} to 0.284101; mean 0.1838, SEM 0.000381), 100 ns (RMSD range 3.00×10^{-7} to 0.3198064; mean 0.2351, SEM 0.000495), and 200 ns (RMSD range 3.00×10^{-7} to 0.284101; mean 0.1838, SEM 0.000381). The 2-sample Kolmogorov-Smirnov tests confirmed significant conformational deviations for all mutations compared to wildtype ($P < 2.2 \times 10^{-16}$; $\alpha = .05$), except for N440K ($P = 1.43 \times 10^{-232}$) and N501Y ($P = 1.18 \times 10^{-64}$) at the 100-ns timescale. L452R consistently exhibited the highest RMSD means among all mutants, especially at the intermediate timescale, while N440K showed the lowest RMSD mean among mutants at 100 ns. For N440K at 200 ns, the RMSD ranged from 4.00×10^{-7} to 0.367109, with a mean of 0.2796 (SEM 0.000353) and a Kolmogorov-Smirnov test P value of 2.20×10^{-16}, indicating significant structural deviation from wildtype. These results highlight mutation and timescale-dependent destabilization of the spike protein structure.
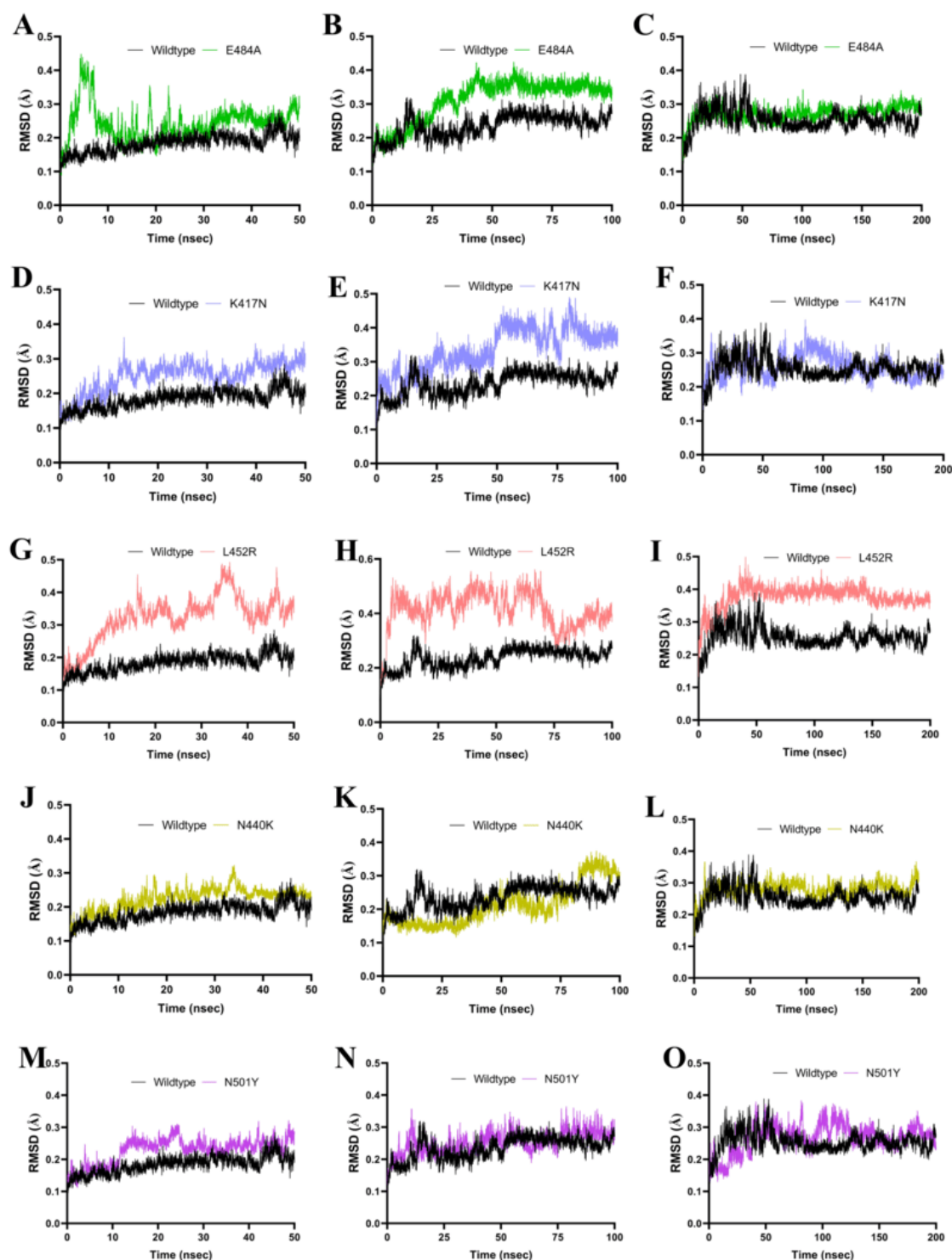
In Figure 13, RMSD analysis shows timescale-dependent structural stability for wildtype and mutated SARS-CoV-2 spike proteins. The wildtype (black) consistently exhibited stability at 50 nanoseconds (range 3×10^{-7} to 0.2841; mean 0.1838, standard error of the mean [SEM] 0.000381), 100 nanoseconds (range 3×10^{-7} to 0.3198; mean 0.2351, SEM 0.000495), and 200 nanoseconds (range 3×10^{-7} to 0.2841; mean 0.1838, SEM 0.000381). Mutants showed increased instability: E484A (green) had 50 nanoseconds (range 4×10^{-7} to 0.4477; mean 0.2435,

SEM 0.000727), 100 nanoseconds (range 4×10^{-7} to 0.4236; mean 0.3108, SEM 0.000896), and 200 nanoseconds (range 4×10^{-7} to 0.4477; mean 0.2435, SEM 0.000727); K417N (lavender) had 50 nanoseconds (range 5×10^{-7} to 0.3620; mean 0.2492, SEM 0.000624), 100 nanoseconds (range 4×10^{-7} to 0.4897; mean 0.3362, SEM 0.000897), and 200 nanoseconds (range 5×10^{-7} to 0.3620; mean 0.2492, SEM 0.000624); L452R (pink) showed a mean of 0.3224 (SEM 0.000984) at 50 nanoseconds and 200 nanoseconds, and at 100 nanoseconds

(range $5×10^{-7}$ to 0.5611; mean 0.4151, SEM 0.000957); N440K (olive) had 50 nanoseconds (range $5×10^{-7}$ to 0.3220; mean 0.2230, SEM 0.000480), 100 nanoseconds (range $1.10×10^{-6}$ to 0.3744; mean 0.2123, SEM 0.000862), and 200 nanoseconds (range $4×10^{-7}$ to 0.3671; mean 0.2796, SEM 0.000353, SD 0.02494), with the Kolmogorov-Smirnov test confirming significant deviation from wildtype at 200 nanoseconds ($P=2.20×10^{-16}$); N501Y (indigo) had 50 nanoseconds (range $4×10^{-7}$ to 0.3203; mean 0.2279, SEM 0.000559), 100 nanoseconds (range $3×10^{-7}$ to 0.3637; mean 0.2529, SEM

0.000484), and 200 nanoseconds (range $4×10^{-7}$ to 0.3203; mean 0.2279, SEM 0.000559). Statistical tests confirmed significant conformational deviation for most mutations compared to wildtype ($P<2.2×10^{-16}$; $α=.05$), except for N440K ($P=1.43×10^{-23}$) and N501Y ($P=1.18×10^{-64}$) at 100 ns. L452R exhibited the highest mean RMSD especially at 100 nanoseconds, while N440K showed the lowest among mutants at 100 nanoseconds, highlighting mutation- and timescale-dependent destabilization. Figure 14 presents corresponding RMSD trajectories.

Figure 14. Overlaid root mean square deviation (RMSD) plots illustrate the conformational stability of wildtype (black) and mutated SARS-CoV-2 spike proteins—E484A (green), K417N (lavender), L452R (pink), N440K (olive), and N501Y (indigo)—across molecular dynamics simulation timescales of 50 ns, 100 ns, and 200 ns. At 50 ns, mean (standard error of the mean [SEM]) values for WT, E484A, K417N, L452R, N440K, and N501Y were 0.1838 (SEM 0.000381), 0.2435 (SEM 0.000727), 0.2492 (SEM 0.000624), 0.3224 (SEM 0.000984), 0.2230 (SEM 0.000480), and 0.2279 (SEM 0.000559) Å, respectively. At 100 ns, values were mean 0.2351 (SEM 0.000495) for WT, mean 0.3108 (SEM 0.000896) for E484A, mean 0.3362 (SEM 0.000897) for K417N, mean 0.4151 (SEM 0.000957) for L452R, mean 0.2123 (SEM 0.000862) for N440K, and mean 0.2529 (SEM 0.000484) for N501Y. At 200 ns, WT, E484A, K417N, L452R, N440K, and N501Y had mean (SEM) values of 0.1838 (SEM 0.000381), 0.2435 (SEM 0.000727), 0.2492 (SEM 0.000624), 0.3224 (SEM 0.000984), 0.2796 (SEM 0.000353, SD 0.02494), and 0.2279 (SEM 0.000559) Å, respectively. L452R demonstrates the highest RMSD means across all timepoints, indicating pronounced instability, while N440K shows the lowest mean among mutants at 100 ns but increased deviation at 200 ns. These overlaid RMSD trajectories highlight mutation- and timescale-dependent conformational changes in the spike protein structure.
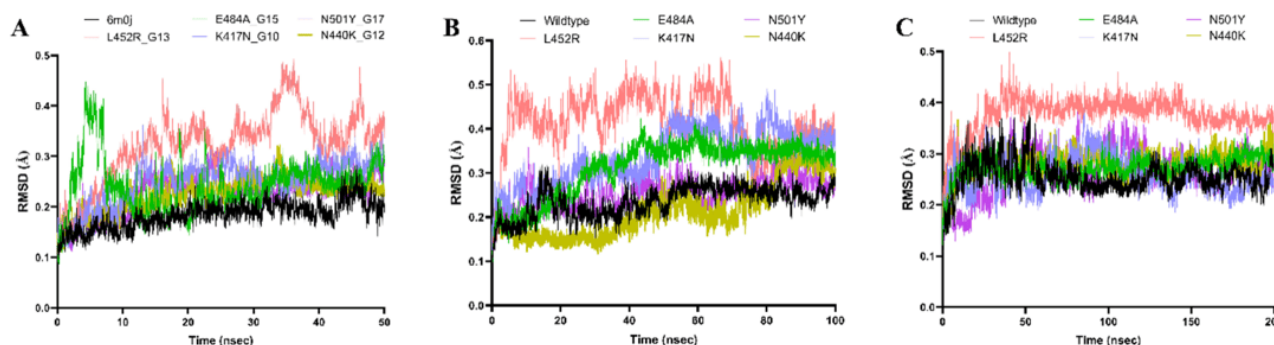

In Figure 13, RMSD analysis shows timescale-dependent structural stability for wildtype and mutated SARS-CoV-2 spike proteins. The wildtype (black) consistently exhibited stability at 50 nanoseconds (range $3×10^{-7}$ to 0.2841; mean 0.1838, SEM 0.000381), 100 nanoseconds (range $3×10^{-7}$ to 0.3198; mean 0.2351, SEM 0.000495), and 200 nanoseconds (range $3×10^{-7}$ to 0.2841; mean 0.1838, SEM 0.000381). Mutants showed increased instability: E484A (green) had 50 nanoseconds (range $4×10^{-7}$ to 0.4477; mean 0.2435, SEM 0.000727), 100 nanoseconds (range $4×10^{-7}$ to 0.4236; mean 0.3108, SEM 0.000896), and 200 nanoseconds (range $4×10^{-7}$ to 0.4477; mean 0.2435, SEM 0.000727); K417N (lavender) had 50 nanoseconds (range $5×10^{-7}$ to 0.3620; mean 0.2492, SEM 0.000624), 100 nanoseconds (range $4×10^{-7}$ to 0.4897; mean 0.3362, SEM 0.000897), and 200 nanoseconds (range $5×10^{-7}$ to 0.3620; mean 0.2492, SEM 0.000624); L452R (pink) showed a mean of 0.3224 (SEM 0.000984) at 50 nanoseconds and 200 nanoseconds, and at 100 ns (range $5×10^{-7}$ to 0.5611; mean 0.4151, SEM 0.000957); N440K (olive) had 50 nanoseconds (range $5×10^{-7}$ to 0.3220; mean 0.2230, SEM 0.000480), 100 nanoseconds (range $1.10×10^{-6}$ to 0.3744; mean 0.2123, SEM 0.000862), and 200 nanoseconds (range $4×10^{-7}$ to 0.3671; mean 0.2796, SEM 0.000353, SD 0.02494), with the Kolmogorov-Smirnov test

confirming significant deviation from wildtype at 200 nanoseconds ($P=2.20×10^{-16}$); N501Y (indigo) had 50 nanoseconds (range $4×10^{-7}$ to 0.3203; mean 0.2279, SEM 0.000559), 100 nanoseconds (range $3×10^{-7}$ to 0.3637; mean 0.2529, SEM 0.000484), and 200 nanoseconds (range $4×10^{-7}$ to 0.3203; mean 0.2279, SEM 0.000559). Statistical tests confirmed significant conformational deviation for most mutations compared to wildtype ($P<2.2×10^{-16}$; $α=.05$), except for N440K ($P=1.43×10^{-2}$) and N501Y ($P=1.18×10^{-64}$) at 100 nanoseconds. L452R exhibited the highest mean RMSD especially at 100 nanoseconds, while N440K showed the lowest among mutants at 100 nanoseconds, highlighting mutation- and timescale-dependent destabilization; Figure 14 presents corresponding RMSD trajectories.

Discussion

Principal Findings

The dendrogram analysis (see Figure 9) highlights significant structural and functional differences between several mutations and the wildtype SARS-CoV-2 spike protein, particularly the N501Y, L452R, E484A, and K147N mutations. These differences, quantified by Euclidean distances from the wildtype, reflect the profound alterations induced by these mutations. The

N501Y mutation (asparagine to tyrosine at position 501) and L452R mutation (leucine to arginine at position 452) exhibit substantial deviations from the wildtype, with Euclidean distances of approximately 0.023 and 0.019, respectively. These mutations significantly impact the spike RBD, enhancing infectivity and immune evasion. As illustrated in Figure 8, these mutations are distinctly clustered within the dendrogram, highlighting their significant divergence from the wildtype. The notable Euclidean distances (N501: 0.023; L452R: 0.019) indicate substantial structural alterations, which correspond to their clinical relevance in enhancing viral transmission and facilitating immune escape. Specifically, the N501Y mutation increases binding affinity to the hACE2 receptor, enhancing viral transmissibility and infectivity. It has been linked to VOCs, such as Alpha, Beta, and Gamma, and enables infection across a broader range of hosts, including mice. This mutation represents a critical adaptive change in SARS-CoV-2 evolution, underscoring its role in the pandemic [28,100]. Similarly, the L452R mutation introduces a positively charged arginine in place of leucine within a hydrophobic region of the RBD. This disrupts local hydrophobic interactions and destabilizes the protein structure, contributing to immune evasion and enhanced infectivity. The dendrogram clustering (see Figure 8) underscores its significant impact, consistent with clinical observations of resistance to neutralizing antibodies and increased viral transmission [101,102].

The E484A mutation (glutamic acid to alanine at position 484) and the K147N mutation (lysine to asparagine at position 147) also exhibit deviations from the wildtype, with Euclidean distances of approximately 0.017 and 0.0075, respectively. Although these mutations show less pronounced structural alterations compared to N501Y and L452R, as indicated by their Euclidean distances in Figure 9, they still represent significant changes in protein structure and function, as illustrated in Figure 9. Research has demonstrated that the E484A mutation impairs antibody recognition, enhancing immune evasion [103]. Similarly, the K147N mutation, located in the N-terminal domain, reduces neutralization by antibodies [104,105]. Although their effects are milder compared to N501Y and L452R, these findings highlight that even seemingly minor mutations can induce important structural changes with functional consequences.

Our study used I-TASSER [90] models and Cytoscape visualizations to investigate structural changes in mutated RBDs of the SARS-CoV-2 spike protein. The analysis revealed significant alterations in the mutated RBDs compared to the wildtype, both at the local mutation sites and in the overall RBD conformation. These findings suggest potential long-range effects on protein dynamics. Our visualizations highlighted localized changes in protein folding near mutation sites and potential impacts on the overall stability and flexibility of the RBD. These structural changes provide a molecular basis for understanding the observed impacts on binding affinity and dynamics of the spike protein. A prime example of these effects is the N501Y mutation in the SARS-CoV-2 spike RBD, known to enhance ACE2 receptor binding. Our study showed that this mutation induces significant local and regional changes in the ab initio model of chain E. Specifically, N501Y leads to

substantial alterations in protein folding and notable changes in structural integrity. These modifications are more pronounced when compared to the wildtype structure, indicating that the N501Y mutation substantially impacts the RBD's conformation. Such structural changes suggest that this mutation may have considerable implications for the RBD's function, potentially affecting viral behavior and interactions with host cells.

In addition, our research, as demonstrated in Figure 10, further illustrates that even small local perturbations can significantly affect the overall structure of the spike protein's RBD. Both L452R and N501Y mutations disrupt protein folding and secondary structures of the SARS-CoV-2 spike protein RBD, as evident from our results in Figures 10-12. Consistent with research findings, the L452R mutation, at a local level, introduces a positively charged residue that alters hydrophobic interactions and stability. On a global scale, it enhances spike protein stability, promotes viral fusion with host membranes, and strengthens ACE2 receptor binding [105]. Similarly, the N501Y mutation is known to form additional hydrogen bonds and π-π interactions with ACE2 locally while globally shifting the spike protein into an "open" prefusion conformation that facilitates receptor engagement. Collectively, these mutations increase ACE2-binding affinity, enhance viral infectivity and transmissibility, and contribute to immune evasion by reducing neutralizing antibody recognition.

Conclusively, our ab initio models offer a valuable structural framework for interpreting experimental data on these mutations. They also provide hypotheses for future investigations into their functional consequences, such as altered receptor binding or antibody recognition. This research contributes significantly to our understanding of how SARS-CoV-2 mutations affect the virus' structure and function, potentially informing future strategies for treatment and prevention.

To gain further insights into the structural stability and dynamic behavior of both wildtype and mutated spike proteins, we conducted 3 sets of MD simulation studies at 50, 100, and 200 nanoseconds. Our results from MD simulations presented in Figure 13 provide a foundational understanding of short-, intermediate-, and long-term dynamics and structural changes for both wildtype and mutated spike proteins. All simulations offered valuable additional insights into conformational changes and stability patterns, demonstrating the high levels of instability of the mutated spike proteins when compared to the wildtype, thus providing a more robust independent confirmation of our graph-theoretic and ab initio model promising predictive analytic tools for complex networks and systems biology. These MD studies complemented our static structural analyses, offering a dynamic perspective on the effects of mutations on the spike protein's behavior over time. Figure 13 displays the overlaid RMSD plots comparing the structural stability of wildtype (black) and mutated SARS-CoV-2 spike proteins (E484A: green; L452R: pink; K417N: lavender) across MD simulation timescales of 50 nanoseconds, 100 nanoseconds, and 200 nanoseconds. For each panel, mean RMSD values are indicated: at 50 ns, wildtype (0.1838, SEM 0.000381), E484A (0.2435, SEM 0.000727), L452R (0.3224, SEM 0.000984), K417N (0.2492, SEM 0.000624); at 100 ns, wildtype (0.2351, SEM 0.000495), E484A (0.3108, SEM 0.000896), L452R (0.4151,

SEM 0.000957), K417N (0.3362, SEM 0.000897); at 200 ns, wildtype (0.1838, SEM 0.000381), E484A (0.2435, SEM 0.000727), L452R (0.3224, SEM 0.000984), K417N (0.2492, SEM 0.000624). The plot demonstrates the minimal fluctuation and high stability of the wildtype protein across all timescales, while mutants—most notably L452R—show significantly higher RMSD means, reflecting greater conformational variability and instability, especially at intermediate and longer durations.

To further validate our initial findings, we conducted extended MD simulations up to 100 nanoseconds and 200 ns, which revealed significant differences in structural stability for most mutant structures compared to the wildtype across these timescales, highlighting the mutation- and timescale-dependent destabilization of the spike protein structure. E484A, K417N, and L452R mutations consistently displayed higher RMSD values, indicating increased instability compared to the wildtype, with L452R demonstrating the highest instability among severe mutations. In contrast, the N440K mutation showed lower RMSD means, suggesting milder effects on protein stability. These findings indicate that mutations can substantially impact the structural dynamics of the spike protein, potentially affecting its function and interactions with host receptors. The observed changes in structural stability across different mutations correlate with clinical observations, suggesting a mechanistic link between altered protein dynamics and enhanced viral properties [105,106,109,110]. Specifically, these structural changes may contribute to the increased infectivity, immune evasion capabilities, and resistance to therapeutic antibodies observed in variants carrying these mutations. This alignment between MD simulations and clinical data underscores the importance of studying protein structural changes in understanding and predicting the behavior of SARS-CoV-2 variants [105,106,109,110].

The overlaid RMSD plot for the N440K (olive) and N501Y (indigo) variants compares their conformational stability with the wildtype spike protein across 50-nanosecond, 100-nanosecond, and 200-nanosecond simulations. For N440K, mean values were 0.2230 (SEM 0.000480) at 50 nanoseconds, 0.2123 (SEM 0.000862) at 100 nanoseconds (the lowest mean among all mutants at this timescale), and 0.2796 (SEM 0.000353) at 200 nanoseconds (SD 0.02494), with increasing mean and range at longer duration signifying greater fluctuation; at 200 nanoseconds, N440K displayed significant deviation from wildtype (Kolmogorov-Smirnov P=2.20×10^{-16}). The N501Y mutation yielded mean RMSDs of 0.2279 (SEM 0.000559) at both 50 nanoseconds and 200 nanoseconds, and 0.2529 (SEM 0.000484) at 100 nanoseconds, indicating persistent but moderate structural alteration relative to wildtype. Statistical analyses confirmed significant conformational deviations for both mutations across timescales (P<2.2×10^{-16} at α=.05), except for N440K (P=1.43×10^{-23}) and N501Y (P=1.18×10^{-64}) at 100 nanoseconds, where deviations were less pronounced, as represented in Figures 13 and 14.

To further validate our initial findings, we conducted extended MD simulations up to 100 nanoseconds and 200 nanoseconds, which revealed significant differences in structural stability for most mutant structures compared to the wildtype across these timescales, highlighting the mutation- and timescale-dependent destabilization of the spike protein structure. E484A, K417N, and L452R mutations consistently displayed higher RMSD values, indicating increased instability compared to the wildtype, with L452R demonstrating the highest instability among severe mutations. In contrast, the N440K mutation showed lower RMSD means, suggesting milder effects on protein stability. These findings indicate that mutations can substantially impact the structural dynamics of the spike protein, potentially affecting its function and interactions with host receptors. The observed changes in structural stability across different mutations correlate with clinical observations, suggesting a mechanistic link between altered protein dynamics and enhanced viral properties [105,106,109,110]. Specifically, these structural changes may contribute to the increased infectivity, immune evasion capabilities, and resistance to therapeutic antibodies observed in variants carrying these mutations. This alignment between MD simulations and clinical data underscores the importance of studying protein structural changes in understanding and predicting the behavior of SARS-CoV-2 variants [105,106,109,110].

Comparison With Previous Work

A few studies have applied graph-theoretic models to investigate the effects of point mutations on protein structures and their associated disease phenotypes [84,85,107]. This research on SARS-CoV-2 spike protein mutations, particularly N501Y, L452R, E484A, and K147N, provides a detailed computational molecular analysis of the structural and functional alterations induced by these mutations. Specifically, it highlights the impact of mutations on the RBD of the spike protein, quantified using Euclidean distances and MD simulations. For instance, the N501Y and L452R mutations significantly disrupt protein folding, enhance ACE2 binding affinity, and contribute to immune evasion [21,28,29,41]. These findings align with previous studies like Knisley et al's [113] graph-theoretic analysis of cystic fibrosis mutations in NBD1 of CFTR proteins, which quantified local and global structural changes caused by mutations. However, unlike previous studies that relied solely on graph-theoretic metrics to model structural perturbations in CFTR proteins [22,84,85,107-109], this research integrated MD simulations to compute RMSD values for mutated spike proteins in comparison to the wildtype. This approach reveals varying degrees of instability among the analyzed mutations, providing insights into mutation-induced conformational dynamics absent in earlier works. Similarly, while Kakraba and Knisley [60,85] focused on CFTR mutations in NBD2, and Netsey et al [61] examined point mutations like Glu6Val in sickle cell hemoglobin [85], this study uniquely addresses structural changes in SARS-CoV-2 spike protein RBD mutations. By combining quantitative approaches with molecular-level dynamics, this research offers additional strengths over previous methodologies that primarily used graph-theoretic models to analyze mutation effects [84,85,107]. Together, these complementary studies underscore the diverse applications of computational methods in unraveling mutation-driven phenomena across biological systems. While previous research has addressed distinct biological contexts, such as cystic fibrosis and SCD, this study emphasizes the importance of integrating

multifaceted approaches, such as ab initio modeling and MD simulations with graph-theoretic modeling, to achieve a deeper understanding of mutation-induced structural and functional changes.

Limitations and Future Directions

Despite these valuable insights from graph-theoretic analysis and MD simulations, certain limitations remain with this study. Specifically, our graph-based model did not fully capture the impact of mutations like N440K—associated with increased infectivity and immune evasion—as it may require more comprehensive molecular descriptors [107,108]. To address this concern, future research should aim to refine these models for better representation of such mutations. In addition, other mutations like S477N and T478K might be included in future studies. Also, future studies can use our graph-theoretic modeling approach to predict point mutations with potentially devastating consequences by analyzing changes in vertex weights and combinatorial descriptors in protein structure graphs, thereby providing a valuable tool for disease surveillance and early intervention strategies. Another limitation of this study is that our MD simulations focus solely on the isolated spike protein (chain E) of SARS-CoV-2, without considering its interaction with ACE2 or other ligands. As a result, we were unable to assess the direct impact of mutations on binding affinity or calculate binding energies, such as DeltaG, for the spike-ACE2 complex. Future studies could address this limitation by conducting molecular simulations of the spike protein bound to ACE2, which would provide deeper insights into how specific mutations affect their interaction and binding dynamics. More so, incorporating transmembrane domains and membrane anchoring in such models could further enhance the physiological relevance of the findings. We recognize that the appropriateness and impact of our chosen vertex-weighting scheme (ΔMd per degree) would benefit from systematic benchmarking against more established alternatives, such as hydrophobicity scores, residue centrality measures, and B-factor–based weights. While a comprehensive, side-by-side comparison of these metrics is beyond the scope of the present dataset and analysis, we acknowledge this as a limitation and a key opportunity for future work. Moving forward, we plan to undertake such comparative evaluations to more rigorously justify the selection of our weighting strategy. We also encourage others in the field to explore and refine these benchmarking efforts to better elucidate the strengths and limitations of alternative vertex-weighting methodologies within graph-theoretic modeling of protein mutation effects.

Conclusions

This study provides a detailed computational analysis of key SARS-CoV-2 spike protein mutations, including N501Y, L452R, E484A, and K147N, and their structural and functional impacts. Using dendrogram clustering, Euclidean distance measurements, and MD simulations, the research highlights how these mutations disrupt protein stability and alter RBD. Mutations, such as N501Y and L452R, significantly enhance ACE2 binding affinity, viral transmissibility, and immune evasion, while even milder mutations like E484A and K147N contribute to structural perturbations and reduced antibody recognition. RMSD analysis revealed varying degrees of instability among mutated proteins, with L452R causing the greatest disruption. These findings align with clinical observations of increased infectivity and immune resistance associated with these mutations. While the study underscores the usability of computational models in understanding mutation-driven phenomena, it also highlights areas for future research, such as refining models to better capture the effects of additional mutations like S477N and T478K.

Overall, this research advances our understanding of SARS-CoV-2 evolution and provides critical insights for monitoring viral mutations and developing effective therapeutic strategies.

Authors' Contributions

SK led the entire study design. EKN conducted graph-theoretic modeling under SK's supervision. JAJ performed molecular simulation and dynamics under SK's supervision. SK analyzed and interpreted all results in this study with inputs from JAJ, JGS, and SKS. The manuscript was written by SK, EKN, SMN, KEA, ACY, GO, JGS, DY, EOD, SD, SM, CEM, EY, MS, and SKS. All authors reviewed and approved the final manuscript version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Structural analysis and molecular dynamics of SARS-CoV-2 spike receptor-binding domain mutations.
[DOCX File, 22 KB - bioinform_v6i1e73637_app1.docx]

References

1. Shrestha N, Shad MY, Ulvi O, et al. The impact of COVID-19 on globalization. One Health 2020 Dec 20;11:100180. [doi: 10.1016/j.onehlt.2020.100180] [Medline: 33072836]

2. Msemburi W, Karlinsky A, Knutson V, Aleshin-Guendel S, Chatterji S, Wakefield J. The WHO estimates of excess mortality associated with the COVID-19 pandemic. Nature New Biol 2023 Jan;613(7942):130-137. [doi: 10.1038/s41586-022-05522-2] [Medline: 36517599]

3. Cascella M, Rajnik M, Aleem A, Dulebohn SC, Di Napoli R. Features, Evaluation, and Treatment of Coronavirus (COVID-19): StatPearls Publishing; 2024. [Medline: 32150360]

4. Pollard CA, Morran MP, Nestor-Kalinoski AL. The COVID-19 pandemic: a global health crisis. Physiol Genomics 2020 Nov 1;52(11):549-557. [doi: 10.1152/physiolgenomics.00089.2020] [Medline: 32991251]

5. Huang Y, Yang C, Xu XF, Xu W, Liu SW. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. Acta Pharmacol Sin 2020 Sep;41(9):1141-1149. [doi: 10.1038/s41401-020-0485-4] [Medline: 32747721]

6. Ayyadevara S, Balasubramaniam M, Kakraba S, Alla R, Mehta JL, Shmookler Reis RJ. Aspirin-mediated acetylation protects against multiple neurodegenerative pathologies by impeding protein aggregation. Antioxid Redox Signal 2017 Dec 10;27(17):1383-1396. [doi: 10.1089/ars.2016.6978] [Medline: 28537433]

7. Kakraba S, Ayyadevara S, Penthala NR, et al. A novel microtubule-binding drug attenuates and reverses protein aggregation in animal models of Alzheimer's disease. Front Mol Neurosci 2019;12:310. [doi: 10.3389/fnmol.2019.00310] [Medline: 31920540]

8. Kakraba S, Ayyadevara S, Mainali N, et al. Thiadiazolidinone (TDZD) analogs inhibit aggregation-mediated pathology in diverse neurodegeneration models, and extend c. elegans life- and healthspan. Pharmaceuticals (Basel) 2023 Oct 20;16(10):10. [doi: 10.3390/ph16101498] [Medline: 37895969]

9. Balasubramaniam M, Ayyadevara S, Ganne A, et al. Aggregate interactome based on protein cross-linking interfaces predicts drug targets to limit aggregation in neurodegenerative diseases. iScience 2019 Oct 25;20:248-264. [doi: 10.1016/j.isci.2019.09.026] [Medline: 31593839]

10. Bowroju SK, Mainali N, Ayyadevara S, et al. Design and synthesis of novel hybrid 8-hydroxy quinoline-indole derivatives as inhibitors of Aβ self-aggregation and metal chelation-induced Aβ aggregation. Molecules 2020 Aug 8;25(16):16. [doi: 10.3390/molecules25163610] [Medline: 32784464]

11. Shi Y, Liu W, Yang Y, Ci Y, Shi L. Exploration of the shared molecular mechanisms between COVID-19 and neurodegenerative diseases through bioinformatic analysis. Int J Mol Sci 2023 Mar 2;24(5):4839. [doi: 10.3390/ijms24054839] [Medline: 36902271]

12. Huang P, Zhang LY, Tan YY, Chen SD. Links between COVID-19 and Parkinson's disease/Alzheimer's disease: reciprocal impacts, medical care strategies and underlying mechanisms. Transl Neurodegener 2023 Jan 30;12(1):5. [doi: 10.1186/s40035-023-00337-1] [Medline: 36717892]

13. Li C, Liu J, Lin J, Shang H. COVID-19 and risk of neurodegenerative disorders: a Mendelian randomization study. Transl Psychiatry 2022 Jul 14;12(1):283. [doi: 10.1038/s41398-022-02052-3] [Medline: 35835752]

14. Idrees D, Kumar V. SARS-CoV-2 spike protein interactions with amyloidogenic proteins: potential clues to neurodegeneration. Biochem Biophys Res Commun 2021 May 21;554:94-98. [doi: 10.1016/j.bbrc.2021.03.100] [Medline: 33789211]

15. Bhardwaj T, Gadhave K, Kapuganti SK, et al. Amyloidogenic proteins in the SARS-CoV and SARS-CoV-2 proteomes. Nat Commun 2023 Feb 20;14(1):945. [doi: 10.1038/s41467-023-36234-4] [Medline: 36806058]

16. Bowroju SK, Crooks P, Penthalla N, et al. Novel tdzd analogs as agents that delay, prevent, or reverse age-associated diseases and as anti-cancer and antileukemic agents. 2020 Feb 14 inventors URL: https://pubchem.ncbi.nlm.nih.gov/patent/US-2023125667-A1 [accessed 2025-08-13]

17. Eberle RJ, Coronado MA, Gering I, et al. Tau protein aggregation associated with SARS-CoV-2 main protease. PLoS ONE 2023;18(8):e0288138. [doi: 10.1371/journal.pone.0288138] [Medline: 37603556]

18. Baranova A, Cao H, Zhang F. Causal effect of COVID-19 on Alzheimer's disease: a Mendelian randomization study. J Med Virol 2023 Jan;95(1):e28107. [doi: 10.1002/jmv.28107] [Medline: 36039844]

19. Rai SN, Tiwari N, Singh P, et al. Exploring the paradox of COVID-19 in neurological complications with emphasis on Parkinson's and Alzheimer's disease. Oxid Med Cell Longev 2022:3012778. [doi: 10.1155/2022/3012778] [Medline: 36092161]

20. Baazaoui N, Iqbal K. COVID-19 and neurodegenerative diseases: prion-like spread and long-term consequences. J Alzheimers Dis 2022;88(2):399-416. [doi: 10.3233/JAD-220105] [Medline: 35599487]

21. Andre M, Lau LS, Pokharel MD, et al. From Alpha to Omicron: how different variants of concern of the SARS-coronavirus-2 impacted the world. Biology (Basel) 2023 Sep 21;12(9):1267. [doi: 10.3390/biology12091267] [Medline: 37759666]

22. Thakur S, Sasi S, Pillai SG, et al. SARS-CoV-2 mutations and their impact on diagnostics, therapeutics and vaccines. Front Med (Lausanne) 2022;9:815389. [doi: 10.3389/fmed.2022.815389] [Medline: 35273977]

23. Pondé RAA. Physicochemical effect of the N501Y, E484K/Q, K417N/T, L452R and T478K mutations on the SARS-CoV-2 spike protein RBD and its influence on agent fitness and on attributes developed by emerging variants of concern. Virology (Auckl) 2022 Jul;572:44-54. [doi: 10.1016/j.virol.2022.05.003] [Medline: 35580380]

XSL•FO
RenderX

24. Willett JDS, Gravel A, Dubuc I, et al. SARS-CoV-2 rapidly evolves lineage-specific phenotypic differences when passaged repeatedly in immune-naïve mice. Commun Biol 2024 Feb 16;7(1):191. [doi: 10.1038/s42003-024-05878-3] [Medline: 38365933]

25. Hattab D, Amer MFA, Al-Alami ZM, Bakhtiar A. SARS-CoV-2 journey: from alpha variant to omicron and its sub-variants. Infection 2024 Jun;52(3):767-786. [doi: 10.1007/s15010-024-02223-y] [Medline: 38554253]

26. Huo J, Dijokaite-Guraliuc A, Liu C, et al. A delicate balance between antibody evasion and ACE2 affinity for Omicron BA.2.75. Cell Rep 2023 Jan 31;42(1):111903. [doi: 10.1016/j.celrep.2022.111903] [Medline: 36586406]

27. Chatterjee S, Bhattacharya M, Nag S, Dhama K, Chakraborty C. A detailed overview of SARS-CoV-2 Omicron: its sub-variants, mutations and pathophysiology, clinical characteristics, immunological landscape, immune escape, and therapies. Viruses 2023 Jan 5;15(1):167. [doi: 10.3390/v15010167] [Medline: 36680207]

28. Yang WT, Huang WH, Liao TL, Hsiao TH, Chuang HN, Liu PY. SARS-CoV-2 E484K mutation narrative review: epidemiology, immune escape, clinical implications, and future considerations. Infect Drug Resist 2022;15:373-385. [doi: 10.2147/IDR.S344099] [Medline: 35140483]

29. Tian F, Tong B, Sun L, et al. N501Y mutation of spike protein in SARS-CoV-2 strengthens its binding to receptor ACE2. Elife 2021 Aug 20;10:e69091. [doi: 10.7554/eLife.69091] [Medline: 34414884]

30. Liu Y, Liu J, Plante KS, et al. The N501Y spike substitution enhances SARS-CoV-2 transmission. bioRxiv. 2021 Mar 9. [doi: 10.1101/2021.03.08.434499] [Medline: 33758836]

31. Khandia R, Singhal S, Alqahtani T, et al. Emergence of SARS-CoV-2 Omicron (B.1.1.529) variant, salient features, high global health concerns and strategies to counter it amid ongoing COVID-19 pandemic. Environ Res 2022 Jun;209:112816. [doi: 10.1016/j.envres.2022.112816] [Medline: 35093310]

32. Carabelli AM, Peacock TP, Thorne LG, et al. SARS-CoV-2 variant biology: immune escape, transmission and fitness. Nat Rev Microbiol 2023 Mar;21(3):162-177. [doi: 10.1038/s41579-022-00841-7] [Medline: 36653446]

33. Pickering B, Lung O, Maguire F, et al. Divergent SARS-CoV-2 variant emerges in white-tailed deer with deer-to-human transmission. Nat Microbiol 2022 Dec;7(12):2011-2024. [doi: 10.1038/s41564-022-01268-9] [Medline: 36357713]

34. Hatmal MM, Alshaer W, Al-Hatamleh MAI, et al. Comprehensive structural and molecular comparison of spike proteins of SARS-CoV-2, SARS-CoV and MERS-CoV, and their interactions with ACE2. Cells 2020 Dec 8;9(12):12. [doi: 10.3390/cells9122638] [Medline: 33302501]

35. Shang J, Ye G, Shi K, et al. Structural basis of receptor recognition by SARS-CoV-2. Nature New Biol 2020 May;581(7807):221-224. [doi: 10.1038/s41586-020-2179-y] [Medline: 32225175]

36. Zhang Y, Yan R, Zhou Q. ACE2, B0AT1, and SARS-CoV-2 spike protein: structural and functional implications. Curr Opin Struct Biol 2022 Jun;74:102388. [doi: 10.1016/j.sbi.2022.102388] [Medline: 35584583]

37. Shang J, Wan Y, Luo C, et al. Cell entry mechanisms of SARS-CoV-2. Proc Natl Acad Sci USA 2020 May 26;117(21):11727-11734. [doi: 10.1073/pnas.2003138117]

38. Wang Q, Zhang Y, Wu L, et al. Structural and functional basis of SARS-CoV-2 entry by using human ACE2. Cell 2020 May 14;181(4):894-904. [doi: 10.1016/j.cell.2020.03.045] [Medline: 32275855]

39. Cantón R, De Lucas Ramos P, García-Botella A, et al. New variants of SARS-CoV-2. Rev Esp Quimioter 2021 Oct;34(5):419-428. [doi: 10.37201/req/071.2021] [Medline: 34076402]

40. Wang P, Nair MS, Liu L, et al. Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7. Nature New Biol 2021 May 6;593(7857):130-135. [doi: 10.1038/s41586-021-03398-2]

41. Law JLM, Logan M, Joyce MA, et al. SARS-COV-2 recombinant receptor-binding-domain (RBD) induces neutralizing antibodies against variant strains of SARS-COV-2 and SARS-COV-1. Vaccine (Auckl) 2021 Sep 24;39(40):5769-5779. [doi: 10.1016/j.vaccine.2021.08.081] [Medline: 34481699]

42. Garcia-Beltran WF, Lam EC, St Denis K, et al. Multiple SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity. Cell 2021 Apr 29;184(9):2372-2383. [doi: 10.1016/j.cell.2021.03.013] [Medline: 33743213]

43. Khan MZI, Nazli A, Al-Furas H, et al. An overview of viral mutagenesis and the impact on pathogenesis of SARS-CoV-2 variants. Front Immunol 2022;13:1034444. [doi: 10.3389/fimmu.2022.1034444] [Medline: 36518757]

44. Huuskonen S, Liu X, Pöhner I, et al. The comprehensive SARS-CoV-2 "hijackome" knowledge base. Cell Discov 2024 Dec 9;10(1):125. [doi: 10.1038/s41421-024-00748-y] [Medline: 39653747]

45. Mostefai F, Grenier JC, Poujol R, Hussin J. Refining SARS-CoV-2 intra-host variation by leveraging large-scale sequencing data. NAR Genom Bioinform 2024 Sep;6(4):lqae145. [doi: 10.1093/nargab/lqae145] [Medline: 39534500]

46. Loguercio S, Calverley BC, Wang C, et al. Understanding the host-pathogen evolutionary balance through Gaussian process modeling of SARS-CoV-2. Patterns (N Y) 2023 Aug 11;4(8):100800. [doi: 10.1016/j.patter.2023.100800] [Medline: 37602209]

47. Chen Y, Guo Y, Pan Y, Zhao ZJ. Structure analysis of the receptor binding of 2019-nCoV. Biochem Biophys Res Commun 2020 Apr;525(1):135-140. [doi: 10.1016/j.bbrc.2020.02.071]

48. Xia X. Domains and functions of spike protein in SARS-COV-2 in the context of vaccine design. Viruses 2021 Jan 14;13(1):109. [doi: 10.3390/v13010109] [Medline: 33466921]

49. Yang Z, Zhou H, Srivastav S, et al. Optimizing Parkinson's disease prediction: a comparative analysis of data aggregation methods using multiple voice recordings via an automated artificial intelligence pipeline. Data (Basel) 2025 Jan;10(1):4. [doi: 10.3390/data10010004]

50. Kakraba S, Wenzheng H, Srivastav S, Shaffer J. AI-enhanced multi-algorithm r shiny app for predictive modeling and analytics- a case study of Alzheimer's disease diagnostics. JMIR Preprints. Preprint posted online on Dec 18, 2024. [doi: 10.2196/preprints.70272]

51. Abraham KE, Kakraba S, Srivastav SK, Shaffer JG, Zhou H, Yang X. Parkinson-s-disease-prediction-code-and-data-repository-main. ResearchGate. 2025. URL: https://tinyurl.com/yf6a6h58 [accessed 2025-11-14]

52. Shaffer JG, Srivastav SK, Kakraba S, Han W. Basic SMART Pred R Shiny web application for machine learning and deep learning tasks. ResearchGate. 2024 Dec. URL: https://tinyurl.com/ca6stuxu [accessed 2025-11-14]

53. Dash M, Liu H, Yao J. Dimensionality reduction of unsupervised data. Presented at: Ninth IEEE International Conference on Tools with Artificial Intelligence; Nov 3-8, 1997. [doi: 10.1109/TAI.1997.632300]

54. Forina M, Armanino C, Raggio V. Clustering with dendrograms on interpretation variables. Anal Chim Acta 2002 Mar;454(1):13-19. [doi: 10.1016/S0003-2670(01)01517-3]

55. Negi SS, Schein CH, Braun W. Regional and temporal coordinated mutation patterns in SARS-CoV-2 spike protein revealed by a clustering and network analysis. Sci Rep 2022 Jan 21;12(1):1128. [doi: 10.1038/s41598-022-04950-4] [Medline: 35064154]

56. Huang Q, Zhang Q, Bible PW, et al. A new way to trace SARS-CoV-2 variants through weighted network analysis of frequency trajectories of mutations. Front Microbiol 2022;13:859241. [doi: 10.3389/fmicb.2022.859241] [Medline: 35369526]

57. Sinha A, Sangeet S, Roy S. Evolution of sequence and structure of SARS-CoV-2 spike protein: a dynamic perspective. ACS Omega 2023 Jul 4;8(26):23283-23304. [doi: 10.1021/acsomega.3c00944] [Medline: 37426203]

58. Becker M, Dulovic A, Junker D, et al. Immune response to SARS-CoV-2 variants of concern in vaccinated individuals. Nat Commun 2021;12(1):3109. [doi: 10.1038/s41467-021-23473-6]

59. Kumari M, Lu RM, Li MC, et al. A critical overview of current progress for COVID-19: development of vaccines, antiviral drugs, and therapeutic antibodies. J Biomed Sci 2022 Sep 12;29(1):68. [doi: 10.1186/s12929-022-00852-9] [Medline: 36096815]

60. Kakraba S, Knisley D. A graph-theoretic model of single point mutations in the cystic fibrosis transmembrane conductance regulator. J Adv Biotechnol 2016;6(1):780-786. [doi: 10.24297/jbt.v6i1.4013]

61. Netsey EK, Kakraba S, Naandam SM, Yadem AC. A mathematical graph-theoretic model of single point mutations associated with sickle cell anemia disease. J Adv Biotechnol 2021;9:1-14. [doi: 10.24297/jbt.v9i.9109]

62. Das S, Das A, Bhattacharya DK, Tibarewala DN. A new graph-theoretic approach to determine the similarity of genome sequences based on nucleotide triplets. Genomics 2020 Nov;112(6):4701-4714. [doi: 10.1016/j.ygeno.2020.08.023] [Medline: 32827671]

63. Buphamalai P, Kokotovic T, Nagy V, Menche J. Network analysis reveals rare disease signatures across multiple levels of biological organization. Nat Commun 2021 Nov 9;12(1):6306. [doi: 10.1038/s41467-021-26674-1] [Medline: 34753928]

64. Galindez G, Sadegh S, Baumbach J, Kacprowski T, List M. Network-based approaches for modeling disease regulation and progression. Comput Struct Biotechnol J 2023;21:780-795. [doi: 10.1016/j.csbj.2022.12.022]

65. Cheng TMK, Lu YE, Vendruscolo M, Lio' P, Blundell TL. Prediction by graph theoretic measures of structural effects in proteins arising from non-synonymous single nucleotide polymorphisms. PLoS Comput Biol 2008 Jul 25;4(7):e1000135. [doi: 10.1371/journal.pcbi.1000135] [Medline: 18654622]

66. Zitnik M, Li MM, Wells A, et al. Current and future directions in network biology. Bioinformatics Advances 2024 Jan 5;4(1). [doi: 10.1093/bioadv/vbae099]

67. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. Proc Natl Acad Sci USA 2007 May 22;104(21):8685-8690. [doi: 10.1073/pnas.0701361104]

68. Nicholson DN, Greene CS. Constructing knowledge graphs and their biomedical applications. Comput Struct Biotechnol J 2020;18:1414-1428. [doi: 10.1016/j.csbj.2020.05.017]

69. Valous NA, Popp F, Zörnig I, Jäger D, Charoentong P. Graph machine learning for integrated multi-omics analysis. Br J Cancer 2024 Jul 22;131(2):205-211. [doi: 10.1038/s41416-024-02706-7]

70. Meng Z, Liu S, Liang S, Jani B, Meng Z. Heterogeneous biomedical entity representation learning for gene–disease association prediction. Brief Bioinformatics 2024 Jul 25;25(5). [doi: 10.1093/bib/bbae380]

71. Ratajczak F, Joblin M, Hildebrandt M, Ringsquandl M, Falter-Braun P, Heinig M. Speos: an ensemble graph representation learning framework to predict core gene candidates for complex diseases. Nat Commun 2023 Nov 8;14(1):7206. [doi: 10.1038/s41467-023-42975-z] [Medline: 37938585]

72. Li H, Han Z, Sun Y, et al. CGMega: explainable graph neural network framework with attention mechanisms for cancer gene module dissection. Nat Commun 2024;15(1):5997. [doi: 10.1038/s41467-024-50426-6]

73. Kwon JJ, Pan J, Gonzalez G, Hahn WC, Zitnik M. On knowing a gene: a distributional hypothesis of gene function. Cell Syst 2024 Jun 19;15(6):488-496. [doi: 10.1016/j.cels.2024.04.008] [Medline: 38810640]

74. Gualdi F, Oliva B, Piñero J. Predicting gene disease associations with knowledge graph embeddings for diseases with curtailed information. NAR Genom Bioinform 2024 Jun;6(2):lqae049. [doi: 10.1093/nargab/lqae049] [Medline: 38745993]

75. Callahan TJ, Tripodi IJ, Stefanski AL, et al. An open source knowledge graph ecosystem for the life sciences. Sci Data 2024 Apr 11;11(1):363. [doi: 10.1038/s41597-024-03171-w] [Medline: 38605048]

76. Manipur I, Giordano M, Piccirillo M, Parashuraman S, Maddalena L. Community detection in protein-protein interaction networks and applications. IEEE/ACM Trans Comput Biol Bioinform 2023;20(1):217-237. [doi: 10.1109/TCBB.2021.3138142] [Medline: 34951849]

77. Zhang W, Chien J, Yong J, Kuang R. Network-based machine learning and graph theory algorithms for precision oncology. NPJ Precis Oncol 2017;1(1):25. [doi: 10.1038/s41698-017-0029-7] [Medline: 29872707]

78. Shen R. Graph mining and module detection in protein-protein interaction networks [Dissertation]. : State University of New York at Albany; 2014 URL: https://scholarsarchive.library.albany.edu/legacy-etd/1267/ [accessed 2025-08-13]

79. Knisley D, Knisley J. Predicting protein-protein interactions using graph invariants and a neural network. Comput Biol Chem 2011 Apr;35(2):108-113. [doi: 10.1016/j.compbiolchem.2011.03.003] [Medline: 21555249]

80. Manzoni C, Lewis PA, Ferrari R. Network analysis for complex neurodegenerative diseases. Curr Genet Med Rep 2020 Mar;8(1):17-25. [doi: 10.1007/s40142-020-00181-z]

81. Razi A, Afghah F, Singh S, Varadan V. Network-based enriched gene subnetwork identification: a game-theoretic approach. Biomed Eng Comput Biol 2016;7(Suppl 2):1-14. [doi: 10.4137/BECB.S38244] [Medline: 27081328]

82. Rout T, Mohapatra A, Kar M, Muduly DK. Essential proteins in cancer networks: a graph-based perspective using Dijkstra's algorithm. Netw Model Anal Health Inform Bioinforma 2024;13(1):42. [doi: 10.1007/s13721-024-00477-y]

83. Wang J, Peng X, Peng W, Wu FX. Dynamic protein interaction network construction and applications. Proteomics 2014 Mar;14(4-5):338-352. [doi: 10.1002/pmic.201300257] [Medline: 24339054]

84. Keskin O, Tuncbag N, Gursoy A. Predicting protein–protein interactions from the molecular to the proteome level. Chem Rev 2016 Apr 27;116(8):4884-4909. [doi: 10.1021/acs.chemrev.5b00683] [Medline: 27074302]

85. Kakraba S. A hierarchical graph for nucleotide binding domain 2 [Master's Thesis]. : East Tennessee State University; 2015 URL: https://dc.etsu.edu/cgi/viewcontent.cgi?article=3895&context=etd [accessed 2025-08-13]

86. Brunk E, Rothlisberger U. Mixed quantum mechanical/molecular mechanical molecular dynamics simulations of biological systems in ground and electronically excited states. Chem Rev 2015 Jun 24;115(12):6217-6263. [doi: 10.1021/cr500628b] [Medline: 25880693]

87. Salo-Ahen OMH, Alanko I, Bhadane R, et al. Molecular dynamics simulations in drug discovery and pharmaceutical development. Processes (Basel) 2021;9(1):71. [doi: 10.3390/pr9010071]

88. Tuckerman ME, Ungar PJ, von Rosenvinge T, Klein ML. Ab initio molecular dynamics simulations. J Phys Chem 1996 Aug;100(31):12878-12887. [doi: 10.1021/jp960480+]

89. Lappala A. The next revolution in computational simulations: harnessing AI and quantum computing in molecular dynamics. Curr Opin Struct Biol 2024 Dec;89:102919. [doi: 10.1016/j.sbi.2024.102919] [Medline: 39306949]

90. Yao N, Chen X, Fu ZH, Zhang Q. Applying classical, ab initio, and machine-learning molecular dynamics simulations to the liquid electrolyte for rechargeable batteries. Chem Rev 2022 Jun 22;122(12):10970-11021. [doi: 10.1021/acs.chemrev.1c00904] [Medline: 35576674]

91. Schlick T, Collepardo-Guevara R, Halvorsen LA, Jung S, Xiao X. Biomolecularmodeling and simulation: a field coming of age. Q Rev Biophys 2011 May;44(2):191-228. [doi: 10.1017/S0033583510000284] [Medline: 21226976]

92. Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 2008 Jan 23;9(1):40. [doi: 10.1186/1471-2105-9-40] [Medline: 18215316]

93. Lan J, Ge J, Yu J, et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature New Biol 2020 May 14;581(7807):215-220. [doi: 10.1038/s41586-020-2180-5]

94. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. Nucleic Acids Res 2000 Jan 1;28(1):235-242. [doi: 10.1093/nar/28.1.235] [Medline: 10592235]

95. Pettersen EF, Goddard TD, Huang CC, et al. UCSF ChimeraX: structure visualization for researchers, educators, and developers. Protein Sci 2021 Jan;30(1):70-82. [doi: 10.1002/pro.3943] [Medline: 32881101]

96. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003 Nov;13(11):2498-2504. [doi: 10.1101/gr.1239303] [Medline: 14597658]

97. Sternberg A, Naujokat C. Structural features of coronavirus SARS-CoV-2 spike protein: targets for vaccination. Life Sci 2020 Sep 15;257:118056. [doi: 10.1016/j.lfs.2020.118056] [Medline: 32645344]

98. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. Nat Methods 2015 Jan;12(1):7-8. [doi: 10.1038/nmeth.3213] [Medline: 25549265]

99. Sanyal D, Chowdhury S, Uversky VN, Chattopadhyay K. An exploration of the SARS-cov-2 spike receptor binding domain (RBD) – a complex palette of evolutionary and structural features. bioRxiv. Preprint posted online on Jun 4, 2020. [doi: 10.1101/2020.05.31.126615]

100. Starr TN, Greaney AJ, Hilton SK, et al. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. Cell 2020 Sep 3;182(5):1295-1310. [doi: 10.1016/j.cell.2020.08.012] [Medline: 32841599]

101. Nestey EK. A graph-theoretic model of haemoglobin domain [Thesis]. : University of Cape Coast; 2020 URL: https://ir.ucc.edu.gh/xmlui/handle/123456789/7310 [accessed 2025-11-10]

102. Aleem A, Samad ABA, Vaqar S. Emerging Variants of SARS-Cov-2 and Novel Therapeutics Against Coronavirus (COVID-19): StatPearls Publishing; 2024. URL: https://www.ncbi.nlm.nih.gov/books/NBK570580/ [accessed 2025-11-12]

103. Ahmad A, Fawaz MAM, Aisha A. A comparative overview of SARS-CoV-2 and its variants of concern. Infez Med 2022;30(3):328-343. [doi: 10.53854/liim-3003-2] [Medline: 36148164]

104. Poudel S, Ishak A, Perez-Fernandez J, et al. Highly mutated SARS-CoV-2 Omicron variant sparks significant concern among global experts – what is known so far? Travel Med Infect Dis 2022;45:102234. [doi: 10.1016/j.tmaid.2021.102234] [Medline: 34896326]

105. Alba JMG, Pérez-Martínez Z, Boga JA, et al. Emergence of new SARS-CoV2 Omicron variants after the change of surveillance and control strategy. Microorganisms 2022 Sep 30;10(10):1954. [doi: 10.3390/microorganisms10101954] [Medline: 36296230]

106. Anand U, Pal T, Zanoletti A, et al. The spread of the omicron variant: Identification of knowledge gaps, virus diffusion modelling, and future research needs. Environ Res 2023 May 15;225:115612. [doi: 10.1016/j.envres.2023.115612] [Medline: 36871942]

107. Kakraba S, Yadem AC, Abraham KE. Unraveling protein secrets: machine learning unveils novel biologically significant associations among amino acids. Preprints.org. Preprint posted online on May 3, 2025. [doi: 10.20944/preprints202505.0139.v1]

108. WebGRO for macromolecular simulations. UAMS. URL: https://simlab.uams.edu [accessed 2025-08-13]

109. Abraham MJ, Murtola T, Schulz R, et al. GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 2015 Sep;1-2:19-25. [doi: 10.1016/j.softx.2015.06.001]

110. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. J Mol Graph 1996 Feb;14(1):33-38. [doi: 10.1016/0263-7855(96)00018-5] [Medline: 8744570]

111. pandas-dev/pandas: Pandas. Zenodo. 2020. URL: https://zenodo.org/records/15597513 [accessed 2025-11-10]

112. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020 Mar;17(3):261-272. [doi: 10.1038/s41592-019-0686-2] [Medline: 32015543]

113. Knisley DJ, Knisley JR, Herron AC. Graph-theoretic models of mutations in the nucleotide binding domain 1 of the cystic fibrosis transmembrane conductance regulator. Comput Biol J 2013 Apr 3;2013:1-9. [doi: 10.1155/2013/938169]

Abbreviations

ACE2: angiotensin-converting enzyme 2
AD: Alzheimer disease
hACE2: human angiotensin-converting enzyme 2
I-TASSER: Iterative Threading Assembly Refinement
MD: molecular dynamics
OPLS-AA: Optimized Potentials for Liquid Simulations—All Atom
PDB: Protein Data Bank
RBD: receptor-binding domain
RMSD: root mean square deviation
SCD: sickle cell disease
SEM: standard error of the mean
SPC: simple point charge
SPC/E: simple point charge/extended
VOC: variants of concern

XSL•FO

RenderX

Stacked Deep Learning Ensemble for Multiomics Cancer Type Classification: Development and Validation Study

Amani Ameen[1], BSc; Nofe Alganmi[1,2], PhD; Nada Bajnaid[1], PhD

[1]Faculty of Computing and Information Technology, King Abdulaziz University, P.O.Box 80200, Jeddah, Saudi Arabia
[2]Institute of Genomic Medicine Sciences (IGMS), King Abdulaziz University, Jeddah, Saudi Arabia

Corresponding Author:
Nofe Alganmi, PhD
Faculty of Computing and Information Technology, King Abdulaziz University, P.O.Box 80200, Jeddah, Saudi Arabia

Abstract

Background: Cancer is one of the leading causes of disease burden globally, and early and accurate diagnosis is crucial for effective treatment. This study presents a deep learning–based model designed to classify 5 common types of cancer in Saudi Arabia: breast, colorectal, thyroid, non-Hodgkin lymphoma, and corpus uteri.

Objective: This study aimed to evaluate whether integrating RNA sequencing, somatic mutation, and DNA methylation profiles within a stacking deep learning ensemble improves cancer type classification accuracy relative to the current state-of-the-art multiomics models.

Methods: Using a stacking ensemble learning approach, our model integrates 5 well-established methods: support vector machine, k-nearest neighbors, artificial neural network, convolutional neural network, and random forest. The methodology involves 2 main stages: data preprocessing (including normalization and feature extraction) and ensemble stacking classification. We prepared the data before applying the stacking model.

Results: The stacking ensemble model achieved 98% accuracy with multiomics versus 96% using RNA sequencing and methylation individually, 81% using somatic mutation data, suggesting that multiomics data can be used for diagnosis in primary care settings. The models used in ensemble learning are among the most widely used in cancer classification research. Their prevalent use in previous studies underscores their effectiveness and flexibility, enhancing the performance of multiomics data integration.

Conclusions: This study highlights the importance of advanced machine learning techniques in improving cancer detection and prognosis, contributing valuable insights by applying ensemble learning to integrate multiomics data for more effective cancer classification.

Introduction

Cancer is a complex worldwide health problem associated with high mortality [1]. Recent years have seen the use of a variety of machine learning techniques applied to high-throughput sequencing technology, which has advanced the classification of cancers based on omics data and offered a promising future for precise treatment choices.

Omics data provide a thorough understanding of biological systems, facilitating research into disease pathways, molecular causes, and ecological dynamics. Omics comprises the following fields: metagenomics (eg, microbial genomes), proteomics (eg, protein abundances), metabolomics (eg, small molecule concentrations), epigenomics (eg, DNA methylation patterns), and genomics (eg, DNA sequences and mutations) [2]. RNA sequencing is one type of omics data and is a powerful sequencing-based method that enables researchers to discover, characterize, and quantify RNA transcripts across the entire transcriptome [3]. RNA sequencing can tell us which genes are turned on in the cell, their expression levels, and at what time they are turned on or off [4]. This allows scientists to better understand cell biology and evaluate changes that might indicate disease. These data are characterized as high-throughput and high-dimensional [5]. Methylation, an epigenetic process involving the addition of methyl groups to DNA, plays a vital role in gene expression regulation [6]. Aberrant methylation patterns are pervasive in human cancers, impacting carcinogenesis stages and serving as potential biomarkers for cancer diagnosis and prognosis [7,8]. A somatic mutation is a permanent change that can arise naturally or be brought about by environmental influences in the DNA sequence of a gene or chromosome. It may have an impact on the structure or function of proteins. In cancer research, they are essential markers that

shed light on the genetic causes of carcinogenesis and inform the creation of patient-specific targeted therapy [9].

Studies have shown that while single-genome research has yielded significant results, integrating multiple omics can enhance our understanding of diseases and provide patients with better treatment options. Therefore, integrating data from multiple omics, rather than using single-omic techniques, may provide a better understanding of biological systems and the causes of diseases. This integration improves prediction accuracy and facilitates more efficient identification of therapeutic targets [10,11].

Dealing with omics data poses several challenges, one of which is that sequencing data are high-dimensional. Second, class imbalance in patient data will reduce the model's performance. The third challenge is that the number of patients in the study is still relatively small, which may cause overfitting problems [12]. Based on these challenges, there is a need for development and contribution in this field, including the development of models that can successfully distinguish between types of cancer while considering the 3 challenges.

Recent studies on the analysis of critical data for cancer disorders have used a variety of machine learning strategies, including the multilayer perceptron [13-16]. The multilayer perceptron is a 3-layer system that consists of an input layer, an output layer, and a hidden layer positioned in the middle. A convolutional neural network (CNN) [17,18] is another kind of neural network that is used. It functions similarly to a feed-forward neural network and consists of a convolutional layer that processes the input and outputs the result to the next layer. They also used random forest (RF) [13,19], which is a technique that involves training a large number of decision trees. The final output of the RF is the class that the majority of the trees select. Deep neural architectures for classification have also been used in [18,20,21]. In addition, the support vector machine (SVM) and k-nearest neighbors (KNN), which are typically used for regression and classification, are commonly applied in this field.

Working with omics data presents several challenges, such as overfitting and class imbalance, which we outline below, along with an overview of how previous work has addressed them. Overfitting is common due to the limited amount of data, often resulting in lower model performance. The model's accuracy is directly influenced by the amount of data used. This issue has been noted in several studies where models are excessively trained to fit the training examples. Upon review, some papers overlooked this issue, while others addressed it through approaches such as regularization, cross-validation, and dropout techniques. Class imbalance is another significant issue in this type of data, affecting model training by biasing it toward the class with more data. Summarizing the methods for dealing with this problem involves 2 main approaches. First, oversampling techniques such as SMOTE (Synthetic Minority Oversampling Technique) and undersampling methods such as downsampling are used to balance class distribution in the dataset. Second, another effective strategy is to use ensemble learning, where different models are trained on either different subsets of data or using various algorithms, pooling their predictions for improved overall performance. These methods collectively aim to address the challenges posed by class imbalance in data-driven tasks such as cancer classification using omics data.

The model proposed in this paper uses ensemble learning of 5 common models to classify the 5 most common types of cancer in the Kingdom of Saudi Arabia using 3 types of omics data. The objective is to investigate whether the model's classification accuracy improves upon integrating multiomics data into our stacking model, which combines 5 of the most popular methods in this field.

Methods

Overview

Our proposed model presents a classification of the 5 most common types in the Kingdom of Saudi Arabia, which are breast, colorectal, thyroid, non-Hodgkin lymphoma, and corpus uteri [22], by using deep learning, which in turn extracts features that are believed to have an important role. The model was designed using stacking ensemble learning as shown in Figure 1, which goes through 2 phases: a preprocessing phase that includes normalization and feature extraction (FE), and a classification phase using an ensemble stacking model. Data entered the preprocessing phase, and the output was then directed to the stacking model. We have performed our experiments in Python 3.10 (Python Software Foundation) on the Aziz Supercomputer of King Abdulaziz University, which is the second fastest supercomputer in the Middle East and North Africa region. The following sections explain how the proposed model works, starting with data collection, followed by preprocessing, and ending with the stacking model.

Figure 1. Methodology of the proposed model. ANN: artificial neural network; CNN: convolutional neural network; FE: feature extraction; KNN: k-nearest neighbors; RF: random forest; SVM: support vector machine; TPM: transcripts per million.
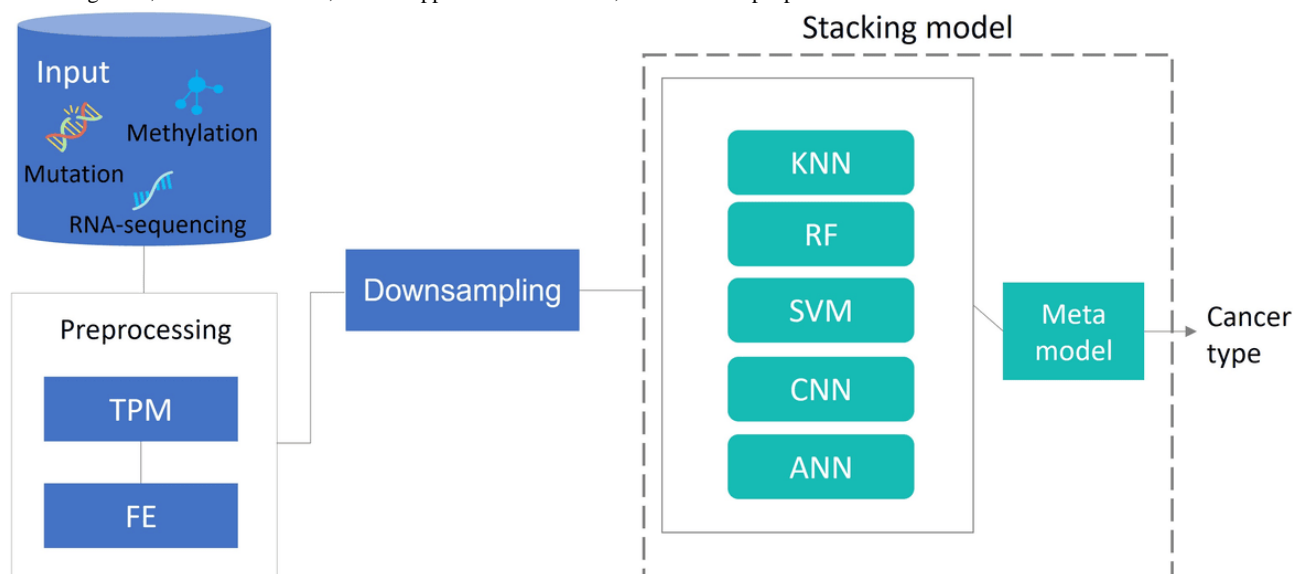
Data Collection and Preprocessing

For RNA sequencing data in this investigation, we used The Cancer Genome Atlas (TCGA) dataset, which is openly accessible to researchers. TCGA comprises approximately 20,000 primary cancer and matched normal samples across 33 cancer types, including the 5 cancer types addressed in our work. Its main goal is to provide scientists with information to improve cancer detection, treatment, and prevention [23]. Furthermore, somatic mutation and methylation data were obtained from the publicly accessible LinkedOmics dataset, which includes

multiomics data from all 32 TCGA cancer types and 10 Clinical Proteomic Tumor Analysis Consortium (CPTAC) cohorts [24].

Figure 2 shows a screenshot of the data types. These are tabular data, with columns representing genes and rows representing cases that are infected by cancer. In Figure 2A, RNA sequencing data capture gene expression levels as continuous values. In Figure 2B, somatic mutation data are sparse and binary (0 or 1), indicating the presence of genomic alterations. In Figure 2C, methylation data provide continuous epigenetic information reflecting gene regulation patterns, with values ranging from −1 to 1.

Figure 2. Show screenshots of the data types: (A) RNA sequencing, (B) somatic mutations, and (C) methylation.

	ENSG00000000003	ENSG00000000005	ENSG00000000419
0	24.175732	2.263355	43.973230
1	15.584672	0.094068	136.057880
2	24.305586	0.274953	32.396744
3	20.707151	3.097024	79.695340

(A)

	ENSG00000121410	ENSG00000148584	ENSG00000078328
0	0	0	0
1	0	0	1
2	0	0	0
3	0	0	0

(B)

	ENSG00000121410	ENSG00000148584	ENSG00000078328
0	−0.3168	0.2011	−0.3811
1	−0.4456	0.2215	−0.4121
2	0.1558	−0.1861	0.1697
3	−0.4865	0.3416	0.3925

(C)

Initially, the data underwent extensive cleaning to ensure the integrity of the model by identifying and removing 7% of cases with missing or duplicate values. Table 1 describes the number of cases of the 5 types of cancer after preprocessing. Regarding

RNA sequencing data, preparation is required before use to provide a precise model evaluation. Thus, 2 processes were carried out in order to preprocess the data: normalization and Feature Extraction (FE).

Table . Show the number of samples in each cancer type after preprocessing.

Cancer type	Abbreviation	RNA sequencing	Somatic mutation	Methylation
Breast	BRCA[a]	1223	976	784
Colorectal	COAD[b]	521	490	394
Thyroid	THCA[c]	568	496	504
Non-Hodgkin lymphoma	NHL[d]	481	240	288
Corpus uteri	UCEC[e]	587	249	432

[a]BRCA: breast invasive carcinoma.

[b]COAD: colon adenocarcinoma.

[c]THCA: thyroid carcinoma.

[d]NHL: non-Hodgkin lymphoma.

[e]UCEC: uterine corpus endometrial carcinoma.

Next, for the normalization step, we used the transcripts per million method to eliminate systematic experimental bias and technical variation while maintaining biodiversity. In addition, it can reduce the bias resulting from the choice of technique used and the conditions tested, or from the experimental procedure, and it can reduce the variance resulting from natural variation and measurement precision [25]. Transcripts per million can be calculated by equation 1 and should be read as "for every 1,000,000 RNA molecules in the RNA-seq sample, x came from this gene/transcript" [26].

$$TPM = 10^6 \times \frac{reads\ mapped\ to\ transcript / transcript\ length}{sum\ (reads\ mapped\ to\ transcript / transcript\ length)} \quad (1)$$

Feature Extraction

RNA sequencing data are high-dimensional. Therefore, to reduce the dimensionality, we use an autoencoder technique based on the results of a study [27] that concluded that autoencoders perform effectively while preserving essential biological properties, allowing for better visualization and interpretation of complex data structures. The architecture of the autoencoder model is composed of an encoder, a code, and a decoder. The encoder compresses the input (features), and the decoder attempts to recreate the input (features) from the compressed version provided by the encoder. The autoencoder model has 5 dense layers, each with 500 nodes and a rectified linear unit (ReLU) activation function. A dropout of 0.3 was applied to handle the overfitting.

Methods for Handling Class Imbalances

In particular, for classes with tiny sample sizes, imbalanced class sizes in the dataset may result in subpar prediction accuracy. Downsampling and SMOTE are 2 methods used to address class imbalances and enhance model performance [28]. In the study by Dittman et al [29], researchers tried class oversampling and class undersampling; then, after evaluating the data, they concluded that undersampling has better results than the oversampling method. Therefore, we decided to apply the downsampling method for the data used in this paper and verified that the data were free of duplicates and then divided into 80% training and 20% test data (Figure 3).

Figure 3. Downsampling for data. NHL had 481 cases in RNA-seq data; 80% (385 cases) were allocated for training and 20% (96 cases) for testing. Somatic mutation types were downsampled to 80% (192 cases) for training and 20% (48 cases) for testing. Methylation data followed suit, with 80% (230 cases) and 20% (58 cases) for training and testing, respectively. BRCA: breast invasive carcinoma; COAD: colon adenocarcinoma; NHL: non-Hodgkin lymphoma; RNA-seq: RNA sequencing; THCA: thyroid carcinoma; UCEC: uterine corpus endometrial carcinoma.
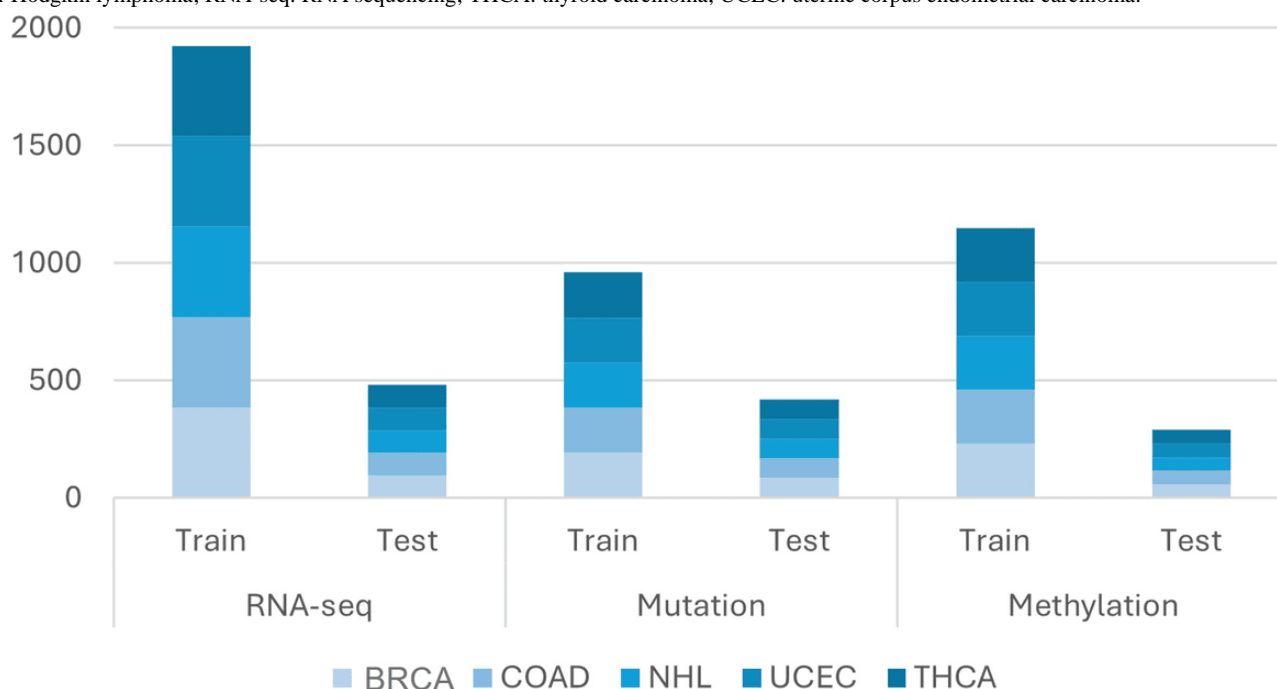
In this dataset, the smallest class (ie, non-Hodgkin lymphoma) included 481 cases in the RNA sequencing data. To balance the dataset, 481 cases were randomly selected from each of the other classes. This resulted in 80% (385 cases) used for training and 20% (96 cases) for testing. For somatic mutations data, each of the 5 types was downsampled to 80% (192 cases) for training and 20% (48 cases) for testing. Similarly, for methylation data, 80% (230 cases) were assigned for training and 20% (58 cases) for testing.

Stacking Ensemble Model

Stacking builds a model with improved performance by training multiple models to come up with the best combination of predictions from these models. The model structure consists of 5 base models and a meta-model that collects the predictions of the base models.

The hyperparameters of each model were described using GridSearchCV (scikit-learn developers), providing a comprehensive configuration for testing and optimization. For the nearest neighbor classifier (BM1), GridSearchCV was used to discover the optimal number of neighbors from values of (1, 3, 5, 10, 5, and 0) and found that the optimal number of neighbors was 10. For the RF classifier (BM2), GridSearchCV was used to explore combinations of "n_estimators" and "min_samples_leaf," achieving the best performance using 500

trees and a minimum of 2 samples per leaf. For the support vector classifier (BM3), the regularization parameter "C" was tuned across a range of values (0.1, 1, 5, 7, and 10), with C=10 achieving the highest accuracy. For CNN (BM4) and artificial neural network (ANN; BM5), GridSearchCV was used to find the optimal activation function from ReLU and softmax, choose dropout rates from 0.1 to 0.6, and finally find the filter value in CNN. Table 2 shows the hyperparameters that we used in each model. Next, the stacking ensemble uses an ANN as the meta-model to combine predictions from BM1 to BM5. The meta-model architecture consists of a neural network with multiple layers. The first dense layer has 32 units and uses a ReLU activation function, followed by a dropout layer with a 50% rate to reduce overfitting. The second dense layer has 16 units and a ReLU activation function, followed by a dropout layer with a 50% rate. The model ends with an output layer that has 5 units and a softmax activation function, suitable for multiclass classification. The model is trained using an Adam optimizer with a learning rate of 0.001 and sparse categorical cross-entropy loss. The integration of the 5 models (SVM, KNN, ANN, CNN, and RF) follows a stacking ensemble approach, where the predictions from each model serve as input features for the meta-model. These base models are trained independently, and their outputs are concatenated to form the input layer of the meta-model.

Table . Hyperparameters of each base model.

Model	Classifier	Hyperparameter
BM1	KNN[a]	Neighbors=10
BM2	RF[b]	n_estimators=500 and min_samples_leaf=2
BM3	SVM[c]	C=10
BM4	CNN[d]	Conv1D with filters= 64, activation="ReLU," optimizer= "adam," loss= "sparse_categorical_crossentropy," and dropout=0.3
BM5	ANN[f]	3 dense layers, activation="ReLU," "softmax," optimizer="adam," loss="sparse categorical crossentropy," and dropout=0.4

[a]KNN: k-nearest neighbor.

[b]RF: random forest.

[c]SVM: support vector machine.

[d]CNN: convolutional neural network.

[e]ReLU: rectified linear unit.

[f]ANN: artificial neural network.

Ethical Considerations

This study exclusively used publicly available datasets obtained from TCGA and LinkedOmics with project names "TCGA-BRCA," "TCGA-COAD," "TCGA-THCA," "TCGA-DLBC," and "TCGA-UCEC". All datasets were fully anonymized and complied with the respective repository's data usage policies.

Results

Overview

In this section, we present the results of our study. First, in the "Performance Evaluation Metrics" section, we analyze critical metrics including the classification report, the confusion matrix, and the receiver operating characteristic (ROC) curve. Second, we present the results of the 5 models individually to compare with our results.

Performance Evaluation Metrics

To assess the effectiveness of the multiclass classification model, various performance metrics were calculated and are shown in Figure 4. The graph shows the performance metrics for a multiclass classification model, including precision, recall, and F_1-score for each class. Precision indicates the accuracy of positive predictions, while recall measures how many actual positives were correctly identified. The F_1-score balances precision and recall. The model achieved an overall accuracy of 98%. Both the macro and weighted averages of the metrics are very similar, reflecting consistent performance across all classes. Subsequently, in Figure 5, we examined the confusion matrix to assess the model's classification performance across the 5 classes. The matrix percentages indicated that the correct classification rates (the diagonal values) were between 91.67% and 100%, showing accurate classification results with error rates (the off-diagonal values) of roughly 8% or less for each class.

Figure 4. Classification report visualizing precision, recall, F_1-score, and support for each class in the stacking ensemble model. BRCA: breast invasive carcinoma; COAD: colon adenocarcinoma; NHL: non-Hodgkin lymphoma; THCA: thyroid carcinoma; UCEC: uterine corpus endometrial carcinoma.

XSL•FO

RenderX
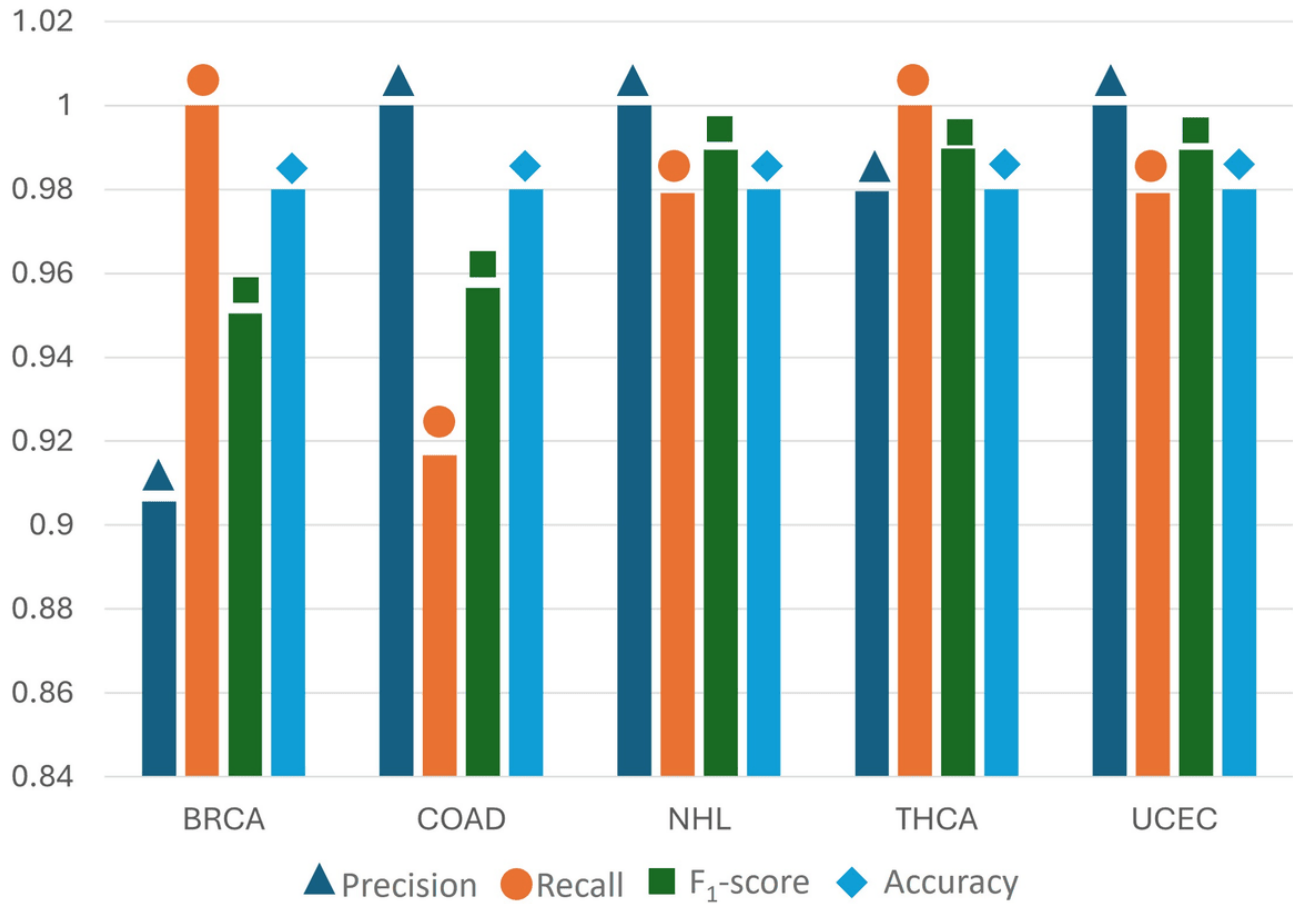
Figure 5. Confusion matrix illustrating the true versus predicted classifications generated by the stacking ensemble model.
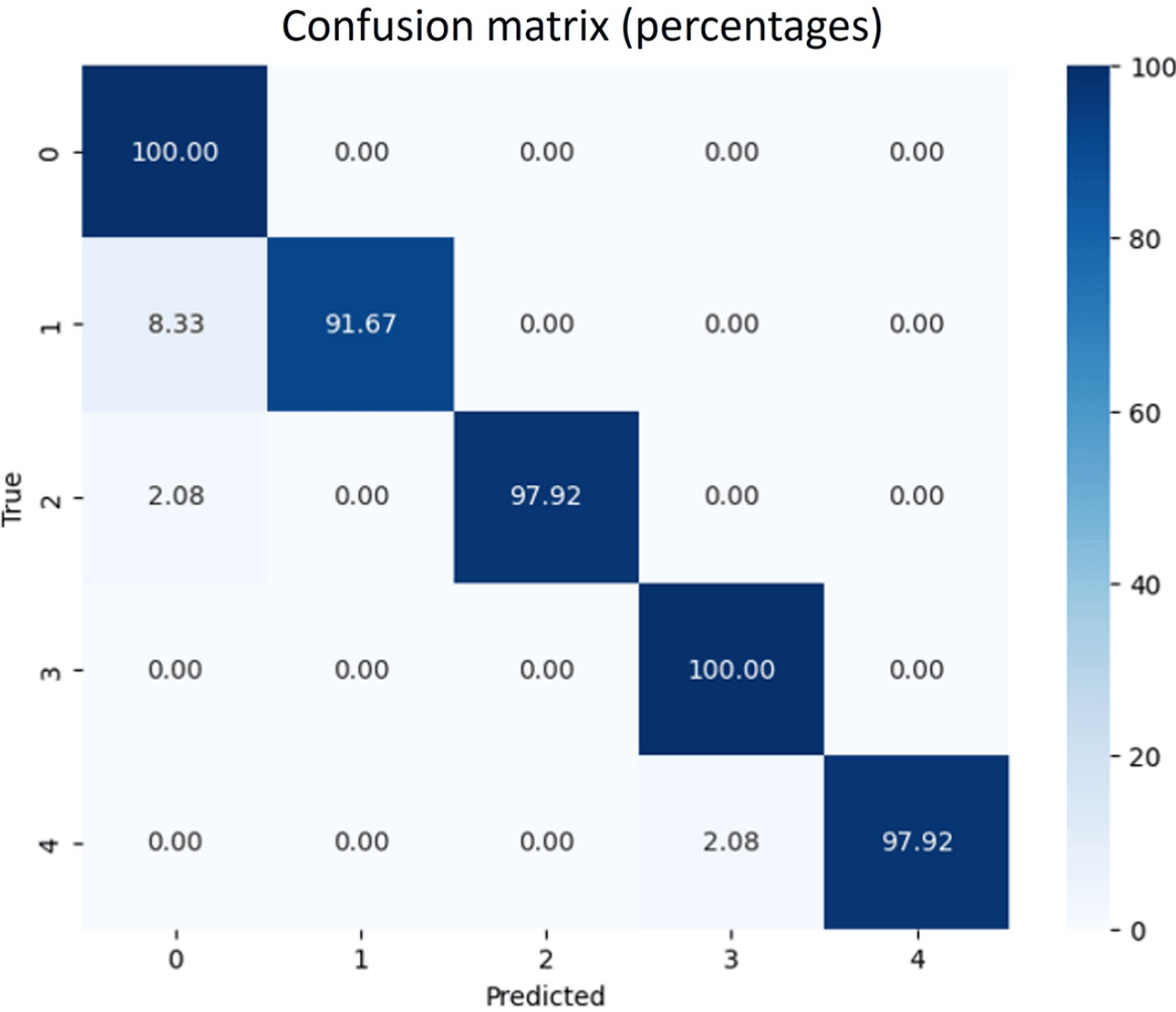
Furthermore, we analyzed the ROC curve, which is a tool for assessing the model's discriminative abilities across multiple classifications. The ROC curve, which provides information about model performance, was modified for our multiclass scenario even though it is usually used in binary classification. In our experiment, we observed compromises between true and false positive rates, which validates the discriminative power of the model. The results, shown in Figure 6, indicate that all classes had consistent performance, as indicated by the area under the curve ranging from 0.90 to 1. These results demonstrate how well the model can classify cases in various classes.

Figure 6. Receiver operating characteristic curve demonstrating the performance of the stacking ensemble model.
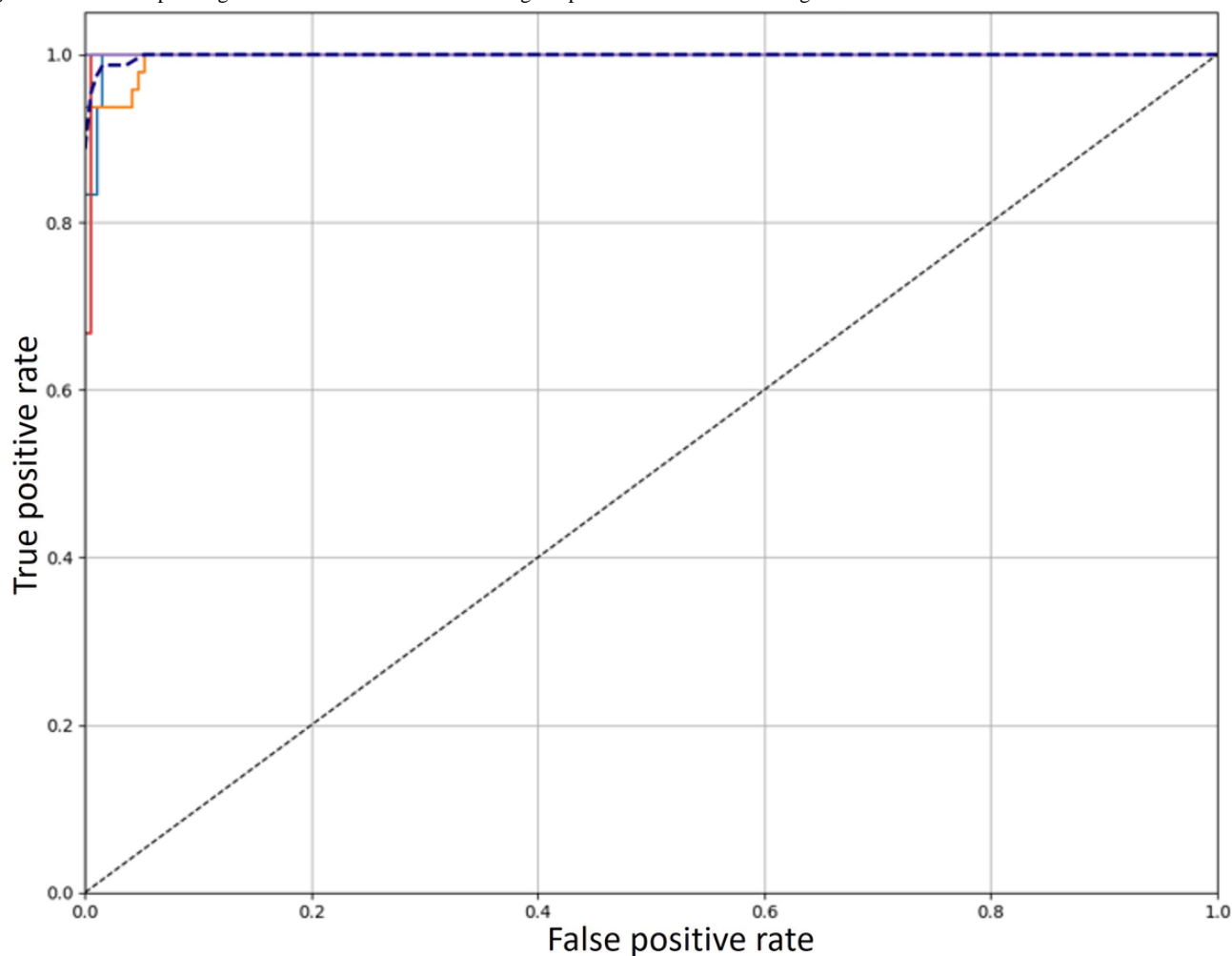
To evaluate the performance of different machine learning approaches on individual omics datasets, we evaluated 5 commonly used classifiers—KNN, RF, SVM, CNN, and ANN—as well as a stacking model composed of all 5 models for each omics type. As shown in Table 3, the RF achieved the highest accuracy on the RNA sequencing dataset (0.98), while the CNN outperformed all other models on the somatic mutations dataset with an accuracy of 0.87. On the methylation dataset, the ANN slightly outperformed the other models with an accuracy of 0.97. The proposed stacking model demonstrated balanced performance across all 3 genome types, achieving accuracies of 0.96 (RNA sequencing), 0.81 (somatic mutations), and 0.96 (methylation). To detail the stacking results, we present Table 4, which shows the performance metrics—precision, F_1-score, recall, and accuracy—for different inputs: RNA sequencing, somatic mutations, methylation separately, and the multiomics approach. For the RNA sequencing input, the model consistently performs well across all 3 folds, with an average precision, F_1-score, recall, and accuracy of 0.96. For the somatic mutations data, the model's accuracy, F_1-score, and recall were relatively low at 0.60, with a slightly higher precision of 0.70. With a mean of 0.97, the accuracy of the model tested on the methylation dataset varied between 0.95 and 0.99 across folds. Similarly, F_1-score and recall averaged 0.96 and 0.97, respectively, while accuracy averaged 0.96. In the multiomics approach, the model achieved an average score of 0.98 across all metrics. Specifically, the model demonstrates near-perfect performance in folds 2 and 3, achieving a precision, recall, and F_1-score of 0.99, reflecting the added value of incorporating multiple data modalities. Overall, the multiomics approach outperforms using each omics type separately, offering a more robust and accurate model across all evaluation metrics. Our analysis showed that some models performed better in terms of recall and precision for certain cancer types when using multiomics, highlighting the importance of combining data to get the most out of the analysis.

Table . Classification accuracy of individual models and the stacking model across RNA sequencing, somatic mutations, and methylation datasets.

Classification model	RNA sequencing	Somatic mutation	Methylation
K-nearest neighbors	0.91	0.72	0.95
Random forest	0.98	0.73	0.96
Support vector machine	0.95	0.79	0.96
Convolutional neural network	0.96	0.87	0.96
Artificial neural network	0.96	0.80	0.97
Stacking with the five model	0.96	0.81	0.96

Table . Performance of the stacking model using RNA sequencing, somatic mutations, methylation, and multiomics data.

Input type and k-fold	Precision	F_1-score	Recall	Accuracy
RNA sequencing				
1	0.95	0.94	0.94	0.94
2	0.97	0.96	0.96	0.96
3	0.98	0.98	0.98	0.98
Avg[a]	0.96	0.96	0.96	0.96
Somatic mutations				
1	0.6	0.6	0.6	0.7
2	0.86	0.85	0.86	0.86
3	0.92	0.91	0.91	0.91
Avg	0.79	0.79	0.79	0.81
Methylation				
1	0.95	0.94	0.94	0.94
2	0.97	0.96	0.97	0.96
3	0.99	0.98	0.99	0.99
Avg	0.97	0.96	0.97	0.96
Multiomics (RNA sequencing, somatic mutations, and methylation)				
1	0.96	0.95	0.95	0.95
2	0.99	0.99	0.99	0.99
3	0.99	0.99	0.99	0.99
Avg	0.98	0.98	0.98	0.98

[a]Avg: average.

Discussion

Principal Findings

The results of this study provide insights into ensemble learning for cancer classification and diagnosis, using 5 different machine learning models. These models were selected based on their proven effectiveness in previous studies and their popularity in the literature, offering a balanced approach to handling the complex nature of multiomics data.

Comparison With Prior Work

Table 5 summarizes several studies that used multiomics data and machine learning techniques to classify and predict various types of cancer. It is worth noting that these studies are not based on the same data but have been reviewed to support our

findings that using multiomics data enhance accuracy. As seen, models from recent studies such as Koh et al [30] and Mohamed and Ezugwu [31] show high area under the curve scores (0.96) and accuracy (97%). Other models, such as Cappelli et al [32] and Jagadeeswara Rao and Sivaprasad [33], also report strong results, typically in the range of 91%‐95%. Overall, these studies highlight the power of integrating multiomics data with advanced machine learning techniques, which consistently led to high accuracy, with models achieving between 91% and 98% accuracy across different cancer types [34]. Although, when comparing the performance of our model with theirs, our approach shows the highest overall accuracy (98%) across a range of cancer types and data modalities. We addressed common challenges in omics data analysis, such as overfitting, class imbalance, and high dimensionality, through the

application of techniques such as dropout, downsampling, and FE. These methods significantly contributed to the robustness of our models, though their effectiveness varied depending on the model and data type.

Table. Comparison of cancer classification performance across multiomics research.

Paper	Year	Data type	Cancer types	Classification model	Overfitting handling	Class imbalance handling	Results (accuracy)
Cappelli et al [32]	2018	RNA sequencing and methylation	BRCA[a], THCA[b], and KIRP[c]	C4.5, RF[d], RIPPER[e], and CAMUR[f]	Feature regularization methods	N/A[g]	95%
Kwon et al [34]	2023	cfDNA[h] and CNVs[i]	LUAD[j]	AdaBoost, MLP[k], and LR[l]	Cross-validation	N/A	91%-98%
Koh et al [30]	2024	Proteomics, RNA sequencing, metabolomics, and targeted immunoassays	Lung	Machine learning	Regularization and QC[m]	Balanced datasets	AUC[n] 0.96
Jagadeeswara Rao and Sivaprasad [33]	2024	RNA sequencing and methylation	PAAD[o]	Ensemble learning	Ensemble techniques	SMOTE[p]	95%
Mohamed and Ezugwu [31]	2024	RNA sequencing, miRNA[q], and DNA methylation	LUAD	CNN[r]	Dropout	SMOTE	97%
Our model	2024	RNA sequencing, methylation, and somatic mutations	BRCA, THCA, NHL[s], UCEC[t], and COAD[u]	Ensemble learning	Cross-validation and dropout	Downsampling	98%

[a]BRCA: breast carcinoma.

[b]THCA: thyroid carcinoma.

[c]KIRP: kidney renal papillary cell carcinoma.

[d]RF: random forest.

[e]RIPPER: Repeated Incremental Pruning to Produce Error Reduction.

[f]CAMUR: Computer Assisted Molecular Unified Receptor.

[g]N/A: not available.

[h]cfDNA: cell-free DNA.

[i]CNV: copy number variation.

[j]LUAD: lung adenocarcinoma.

[k]MLP: multilayer perceptron.

[l]LR: logistic regression.

[m]QC: quality control.

[n]AUC: area under the curve.

[o]PAAD: pancreatic adenocarcinoma.

[p]SMOTE: Synthetic Minority Oversampling Technique.

[q]miRNA: microRNA.

[r]CNN: convolutional neural network.

[s]NHL: non-Hodgkin lymphoma.

[t]UCEC: uterine corpus endometrial carcinoma.

[u]COAD: colon adenocarcinoma.

Typically, deep learning components benefit from graphics processing unit acceleration and need a large amount of computational power, particularly when trained on high-dimensional clinical data. Nevertheless, after training, the model inference time is rather short, allowing for quick predictions that can assist with clinical decisions made in real time. Even while low-resource systems might not be able to support model training, these pretrained models could be used for clinical deployment, particularly in settings with recent computer technology.

Strengths and Limitations

Typically, deep learning components benefit from graphics processing unit acceleration and need a large amount of computational power, particularly when trained on high-dimensional clinical data. Nevertheless, the model inference time is rather short after the ensemble has been trained, allowing for quick predictions that can assist with clinical decisions made in real time. Even while low-resource systems might not be able to support model training, pretrained models can be used for clinical deployment, particularly in settings with recent computer technology.

However, the study has several limitations that must be acknowledged. Data availability constraints limited the scope of our analysis, and the absence of clinical data meant that our findings are based solely on omics data. This restricts the generalizability of our results to real-world clinical settings, where the integration of clinical and omics data is crucial for accurate cancer diagnosis and prognosis. Furthermore, the common limitation in omics data is dataset size, which may result in overfitting. Another restriction is the absence of external validation.

Future Directions

Future research should focus on expanding the types of data used in cancer classification, particularly by incorporating patient clinical data and exploring additional omics layers such as metabolomics and proteomics. Furthermore, the integration of multiomics data with advanced machine learning methods holds promise for deepening our understanding of the molecular mechanisms underlying cancer development. This could lead to more precise cancer staging and prognosis, ultimately improving patient outcomes.

Conclusions

In conclusion, while our study advances the accuracy of cancer classification algorithms, it underscores the need for continuous improvement and validation in diverse and clinically relevant datasets. By addressing these challenges, future research can enhance the applicability of these models in clinical practice, contributing to more effective cancer detection and treatment strategies.

The study aimed to investigate whether incorporating multiomics data into a stacking model that integrates 5 key methods, namely SVM, KNN, ANN, CNN, and RF, enhances the model's ability to classify cancer. With multiomics, the stacking ensemble model obtained 98% accuracy, compared to 96% with RNA sequencing and methylation separately and 81% with somatic mutation data. It emphasizes the importance of integrating advanced machine learning techniques into health care for more effective cancer detection and prognosis. This highlights the need for continuous improvement and validation of classification models in real-world clinical settings to maximize their impact on cancer care. Future research should focus on incorporating clinical metadata and multiomics data to enhance cancer classification, which would improve patient outcomes and clinical applicability.

Conflicts of Interest

None declared.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018 Nov;68(6):394-424. [doi: 10.3322/caac.21492] [Medline: 30207593]

2. Berger B, Peng J, Singh M. Computational solutions for omics data. Nat Rev Genet 2013 May;14(5):333-346. [doi: 10.1038/nrg3433] [Medline: 23594911]

3. Zararsız G, Goksuluk D, Korkmaz S, et al. A comprehensive simulation study on classification of RNA-Seq data. PLoS One 2017;12(8):e0182507. [doi: 10.1371/journal.pone.0182507] [Medline: 28832679]

4. Conesa A, Madrigal P, Tarazona S, et al. A survey of best practices for RNA-seq data analysis. Genome Biol 2016 Jan 26;17(1):13. [doi: 10.1186/s13059-016-0881-8] [Medline: 26813401]

5. Holzinger A, Jurisica I. Knowledge discovery and data mining in biomedical informatics: the future is in integrative, interactive machine learning solutions. In: Holzinger A, Jurisica I, editors. Interactive Knowledge Discovery and Data Mining in Biomedical Informatics: Springer; 2014:1-18.

6. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 2012 May 29;13(7):484-492. [doi: 10.1038/nrg3230] [Medline: 22641018]

7. Pu W, Geng X, Chen S, et al. Aberrant methylation of CDH13 can be a diagnostic biomarker for lung adenocarcinoma. J Cancer 2016;7(15):2280-2289. [doi: 10.7150/jca.15758] [Medline: 27994665]

8. Qiu J, Peng B, Tang Y, et al. CpG methylation signature predicts recurrence in early-stage hepatocellular carcinoma: results from a multicenter study. J Clin Oncol 2017 Mar;35(7):734-742. [doi: 10.1200/JCO.2016.68.2153] [Medline: 28068175]

9. Huang L, Guo Z, Wang F, Fu L. KRAS mutation: from undruggable to druggable in cancer. Signal Transduct Target Ther 2021 Nov 15;6(1):386. [doi: 10.1038/s41392-021-00780-4] [Medline: 34776511]

10. Kang M, Ko E, Mersha TB. A roadmap for multi-omics data integration using deep learning. Brief Bioinform 2022 Jan 17;23(1):bbab454. [doi: 10.1093/bib/bbab454] [Medline: 34791014]

11. Cai Z, Poulos RC, Liu J, Zhong Q. Machine learning for multi-omics data integration in cancer. iScience 2022 Feb 18;25(2):103798. [doi: 10.1016/j.isci.2022.103798] [Medline: 35169688]

12. Zhu W, Xie L, Han J, Guo X. The application of deep learning in cancer prognosis prediction. Cancers (Basel) 2020 Mar 5;12(3):603. [doi: 10.3390/cancers12030603] [Medline: 32150991]

13. Kabir MF, Ludwig SA. Classification models and survival analysis for prostate cancer using RNA sequencing and clinical data. Presented at: 2019 IEEE International Conference on Big Data (Big Data); Dec 9-12, 2019; Los Angeles, CA, USA. [doi: 10.1109/BigData47090.2019.9006036]

14. Feng C, Xiang T, Yi Z, Zhao L, He S, Tian K. An ensemble model for tumor type identification and cancer origins classification. Annu Int Conf IEEE Eng Med Biol Soc 2021 Nov;2021:1660-1665. [doi: 10.1109/EMBC46164.2021.9629691] [Medline: 34891604]

15. Singh NP, Bapi RS, Vinod PK. Machine learning models to predict the progression from early to late stages of papillary renal cell carcinoma. Comput Biol Med 2018 Sep 1;100:92-99. [doi: 10.1016/j.compbiomed.2018.06.030] [Medline: 29990647]

16. Kosvyra A, Maramis C, Chouvarda I. A data-driven approach to build a predictive model of cancer patients' disease outcome by utilizing co-expression networks. Comput Biol Med 2020 Oct;125(2):103971. [doi: 10.1016/j.compbiomed.2020.103971]

17. Nosi V, Luca A, Milan M, et al. MET Exon 14 skipping: a case study for the detection of genetic variants in cancer driver genes by deep learning. Int J Mol Sci 2021 Apr 19;22(8):4217. [doi: 10.3390/ijms22084217] [Medline: 33921709]

18. Mohammed M, Mwambi H, Mboya IB, Elbashir MK, Omolo B. A stacking ensemble deep learning approach to cancer type classification based on TCGA data. Sci Rep 2021 Aug 2;11(1):15626. [doi: 10.1038/s41598-021-95128-x] [Medline: 34341396]

19. Alge O, Gryak J, Hua Y, Najaria K. Classifying osteosarcoma using meta-analysis of gene expression. Presented at: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Dec 3-6, 2018; Madrid, Spain. [doi: 10.1109/BIBM.2018.8621119]

20. Rajpal S, Agarwal M, Kumar V, Gupta A, Kumar N. Triphasic DeepBRCA-a deep learning-based framework for identification of biomarkers for breast cancer stratification. IEEE Access 2021;9:103347-103364. [doi: 10.1109/ACCESS.2021.3093616]

21. Xu J, Wu P, Chen Y, Meng Q, Dawood H, Khan MM. A novel deep flexible neural forest model for classification of cancer subtypes based on gene expression data. IEEE Access 2019;7:22086-22095. [doi: 10.1109/ACCESS.2019.2898723]

22. In the latest Saudi Cancer Registry report issued by the Saudi Health Council. Saudi Health Council. 2016. URL: https://shc.gov.sa/Arabic/MediaCenter/News/Pages/News113.aspx [accessed 2025-07-17]

23. The Cancer Genome Atlas Program (TCGA). Center for Cancer Genomics at the National Cancer Institute. URL: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga [accessed 2025-07-17]

24. Vasaikar SV, Straub P, Wang J, Zhang B. LinkedOmics: analyzing multi-omics data within and across 32 cancer types. Nucleic Acids Res 2018 Jan 4;46(D1):D956-D963. [doi: 10.1093/nar/gkx1090] [Medline: 29136207]

25. Bushel PR, Ferguson SS, Ramaiahgari SC, Paules RS, Auerbach SS. Comparison of normalization methods for analysis of TempO-Seq targeted RNA sequencing data. Front Genet 2020;11:594. [doi: 10.3389/fgene.2020.00594] [Medline: 32655620]

26. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci 2012 Dec;131(4):281-285. [doi: 10.1007/s12064-012-0162-3]

27. Zogopoulos V, Tsotra I, Spandidos D, Iconomidou V, Michalopoulos I. Single-cell RNA sequencing data dimensionality reduction (Review). World Acad Sci J 2025;7(2):27. [doi: 10.3892/wasj.2025.315]

28. Lee W, Seo K. Downsampling for binary classification with a highly imbalanced dataset using active learning. Big Data Research 2022 May;28:100314. [doi: 10.1016/j.bdr.2022.100314]

29. Dittman DJ, Khoshgoftaar TM, Wald R, Napolitano AE. Comparison of data sampling approaches for imbalanced bioinformatics data. Presented at: Proceedings of the 27th International Florida Artificial Intelligence Research Society Conference (FLAIRS 2014); May 21-23, 2014; Pensacola Beach, FL URL: https://cdn.aaai.org/ocs/7850/7850-36780-1-PB.pdf [accessed 2025-07-17]

5. Koh B, Liu M, Almonte R, et al. Multi-omics profiling with untargeted proteomics for blood-based early detection of lung cancer. medRxiv. Preprint posted online on 2024. [doi: 10.1101/2024.01.01.23285841]

31. Mohamed TIA, Ezugwu AES. Enhancing lung cancer classification and prediction with deep learning and multi-omics data. IEEE Access 2024;12:59880-59892. [doi: 10.1109/ACCESS.2024.3394030]

32. Cappelli E, Felici G, Weitschek E. Combining DNA methylation and RNA sequencing data of cancer for supervised knowledge extraction. BioData Min 2018;11:22. [doi: 10.1186/s13040-018-0184-6] [Medline: 30386434]

33. Jagadeeswara Rao G, Sivaprasad A. An integrated ensemble learning technique for gene expression classification and biomarker identification from RNA-seq data for pancreatic cancer prognosis. Int J inf tecnol 2024 Mar;16(3):1505-1516. [doi: 10.1007/s41870-023-01688-8]

34. Kwon HJ, Park UH, Goh CJ, et al. Enhancing lung cancer classification through integration of liquid biopsy multi-omics data with machine learning techniques. Cancers (Basel) 2023 Sep 14;15(18):4556. [doi: 10.3390/cancers15184556] [Medline: 37760525]

Abbreviations

ANN: artificial neural network
CNN: convolutional neural network
FE: feature extraction
KNN: k-nearest neighbors
ReLU: rectified linear unit
RF: random forest
ROC: receiver operating characteristic
SMOTE: Synthetic Minority Oversampling Technique
SVM: support vector machine
TCGA: The Cancer Genome Atlas

XSL•FO
RenderX

Framework for Race-Specific Prostate Cancer Detection Using Machine Learning Through Gene Expression Data: Feature Selection Optimization Approach

David Agustriawan[1*], PhD; Adithama Mulia[1*]; Marlinda Vasty Overbeek[1], MSc; Vincent Kurniawan[1*]; Jheno Syechlo[1*]; Moeljono Widjaja[1*], PhD; Muhammad Imran Ahmad[2*], PhD

[1]Faculty of Engineering and Informatics, Universitas Multimedia Nusantara, Scientia Garden Jalan Boulevard Gading Serpong, Tangerang, Indonesia

[2]Faculty of Electronic Engineering and Technology, Universiti Malaysia Perlis, Perlis, Malaysia

[*]these authors contributed equally

Corresponding Author:
David Agustriawan, PhD
Faculty of Engineering and Informatics, Universitas Multimedia Nusantara, Scientia Garden Jalan Boulevard Gading Serpong, Tangerang, Indonesia

Abstract

Background: Previous machine learning approaches for prostate cancer detection using gene expression data have shown remarkable classification accuracies. However, prior studies overlook the influence of racial diversity within the population and the importance of selecting outlier genes based on expression profiles.

Objective: We aim to develop a classification method for diagnosing prostate cancer using gene expression in specific populations.

Methods: This research uses differentially expressed gene analysis, receiver operating characteristic analysis, and MSigDB (Molecular Signature Database) verification as a feature selection framework to identify genes for constructing support vector machine models.

Results: Among the models evaluated, the highest observed accuracy was achieved using 139 gene features without oversampling, resulting in 98% accuracy for White patients and 97% for African American patients, based on 388 training samples and 92 testing samples. Notably, another model achieved a similarly strong performance, with 97% accuracy for White patients and 95% for African American patients, using only 9 gene features. It was trained on 374 samples and tested on 138 samples.

Conclusions: The findings identify a race-specific diagnosis method for prostate cancer detection using enhanced feature selection and machine learning. This approach emphasizes the potential for developing unbiased diagnostic tools in specific populations.

Introduction

Prostate Cancer Statistics

Prostate cancer is the most common type of organ cancer and the second leading cause of death in the United States among men [1,2]. In 2019, over 893,660 cancer cases were recorded in the United States, with prostate cancer being over 191,930 of them, along with the 2020 estimated number of deaths caused by cancer being 321,160, of which 33,310 were prostate cancer [3-5]. This is likely caused by risk factors found in prostate cancer that include age, family history, and lifestyle. Studies have shown that Asians tend to have a lower risk of prostate cancer than Europeans and Africans due to their genetics and environmental differences [6]. This indicates racial disparity in prostate cancer, which has been extensively documented by numerous studies, with African American men having a higher risk of developing prostate cancer and facing a 2.5-fold higher mortality rate compared to European American men [7,8]. This disparity is attributed to socioeconomic and biological differences, including aggressive tumor phenotypes documented at the molecular level in African American men [9].

Prostate Cancer Detection Methods

In the early 1990s, digital rectal examination was used for screening prostate cancer, which had a significant impact on prostate cancer diagnosis at the time. Digital rectal examination remains beneficial for distinguishing between benign and malignant conditions in the prostate, but it is limited by its low sensitivity and inability to detect cancer at an early stage [3,10,11]. Another screening method is the prostate-specific antigen (PSA) test. While widely used, PSA testing is

XSL•FO
RenderX

controversial due to its susceptibility to false positives, as PSA is a gland-specific biomarker rather than cancer-specific biomarker [10,12]. The lack of a reliable and robust detection method gives rise to the need for a race-based approach to detect prostate cancer.

Machine Learning and Support Vector Machine

In recent years, machine learning applications in health care and biotechnology have grown rapidly, driving advancements in disease diagnostics, personalized medicine, and bioinformatics [13]. In this research, support vector machines (SVMs) were selected for their remarkable performance in classification tasks in the medical field using gene expression data [14-18]. Being a supervised machine learning algorithm that is proficient at distinguishing between 2 sample classes, SVM works by creating a hyperplane that optimally separates sample classes. SVM transforms class data into a higher-dimensional space to effectively identify complex, nonlinear relationships. This makes SVM especially powerful in cases with small sample sizes and high-dimensional data, such as gene expression profiles or genomic datasets. These characteristics made SVM an invaluable algorithm in bioinformatics, where the classification of diseases such as cancer requires robust, data-driven methods to handle variability and heterogeneity [10,15].

Gene Expression Data

Gene expression is a process where information in DNA becomes instructions to make proteins or other molecules [16,19]. The process starts when DNA is copied into mRNA and changed into proteins. Gene expression analysis is typically used for monitoring genetic changes in tissues or single cells under certain conditions. It checks how many DNA transcripts are in a sample to know which genes are active and by how much, including comparing the sequenced reads with the number of base pairs from a DNA piece to a known genome or transcriptome. The process' accuracy depends on the clarity of information obtained, which allows bioinformatics tools to match them to the right genes. However, the gene expression dataset poses an additional challenge due to their high dimensionality, where the ratio of features to samples is high, hindering the performance of classification models. To address this, researchers have used feature selection methods to filter out irrelevant or redundant genes [20,21]. Feature selection has a critical role in improving machine learning models' classification outcomes in high-dimensional datasets, making it a basis for an efficient classification model for cancer detection [22,23].

Racial Dataset Influence in Artificial Intelligence

Racial-based genomic datasets present challenges for machine learning applications. Studies have shown that using race-based genomics data for artificial intelligence algorithms may exhibit biases where trained models favor the majority race in training data, lowering the accuracy on the minority races [8,24]. Racial class imbalance in the dataset, where certain races have more samples, can influence the accuracy of algorithms. However, when the class imbalance is less severe, the algorithms tend to achieve higher balanced accuracy across all racial groups [25].

To mitigate this, an approach that reweighs the minority classes is performed, yet this approach was unreliable when the class imbalance is severe [24,26]. This research uses race-based genomics data instead of a combined race dataset to address the biases that may appear when using a combined dataset.

Prior Research and Objective

Despite significant advancements in machine learning and prostate cancer diagnosis, a gap remains in addressing racial disparities in prostate cancer. A recent study by Alshareef et al [27] introduces artificial intelligence–based feature selection with deep learning model for prostate cancer detection, a newly developed method of prostate cancer detection using deep learning approach using microarray gene expression data with 52 prostate samples and 50 normal samples on 2135 genes [28]. It focuses on feature selection using Chaotic Invasive Weed Optimization and hyperparameter tuning over multiple iterations of the proposed artificial intelligence–based feature selection with deep learning model for prostate cancer detection model which leads to an average accuracy of 97.19%, precision of 97.14%, and F_1-score of 97.28%. Similarly, Ravindran et al [29] proposed a prediction deep learning model for prostate cancer which focuses on data augmentation using the Wasserstein Tabular Generative Adversarial Network technique, which enables powerful discriminators that supply reliable gradient information to the sample generator even with poor sample qualities, allowing for a more stable training process [27]. The research uses a Micro Gene Expression Cancer Dataset (MGECD), of which the prostate cancer MGECD consists of 102 samples and 6033 features, and feature selection based on correlation coefficients with the goal of reducing the features to 1/3 of the initial MGECD by applying a threshold of 0.7. This results in 1833 features being used for the final model that has a 97% accuracy, 98% precision, and 97% recall values, a total of 3.4% accuracy improvement on prostate cancer classification using Wasserstein Tabular Generative Adversarial Network SVM compared to only using SVM. Previous research has demonstrated admirable results with limited amounts of samples, yet the proposed methods do not account for the racial biases that may be present in gene expression data and the number of genes needed to efficiently train machine learning models. To bridge this gap, we use feature selection methods such as differentially expressed gene (DEG) analysis, receiver operating characteristic (ROC) analysis, and MSigDB (Molecular Signature Database) verification. Our goal is to develop a race-based SVM model that improves prostate cancer detection for White populations and provides a novel genomics-based approach for health care professionals.

Methods

Study Design

This study implements data collection, preprocessing, feature selection, and SVM modeling and evaluation as seen in Figure 1. These methods are conducted using Python (version 3.12.3; Python Software Foundation) programming language and the necessary libraries using Visual Studio Code editor (version 1.95.3; Microsoft Corp) [30].

Figure 1. Race-specific prostate cancer detection modeling framework. DEG: differentially expressed gene; GDC: Genomic Data Commons; MSigDB: Molecular Signature Database; ROC: receiver operating characteristic; STAR: Spliced Transcripts Alignment to a Reference; SVM: support vector machine; TCGA: The Cancer Genome Atlas; UCSC: University of California, Santa Cruz.
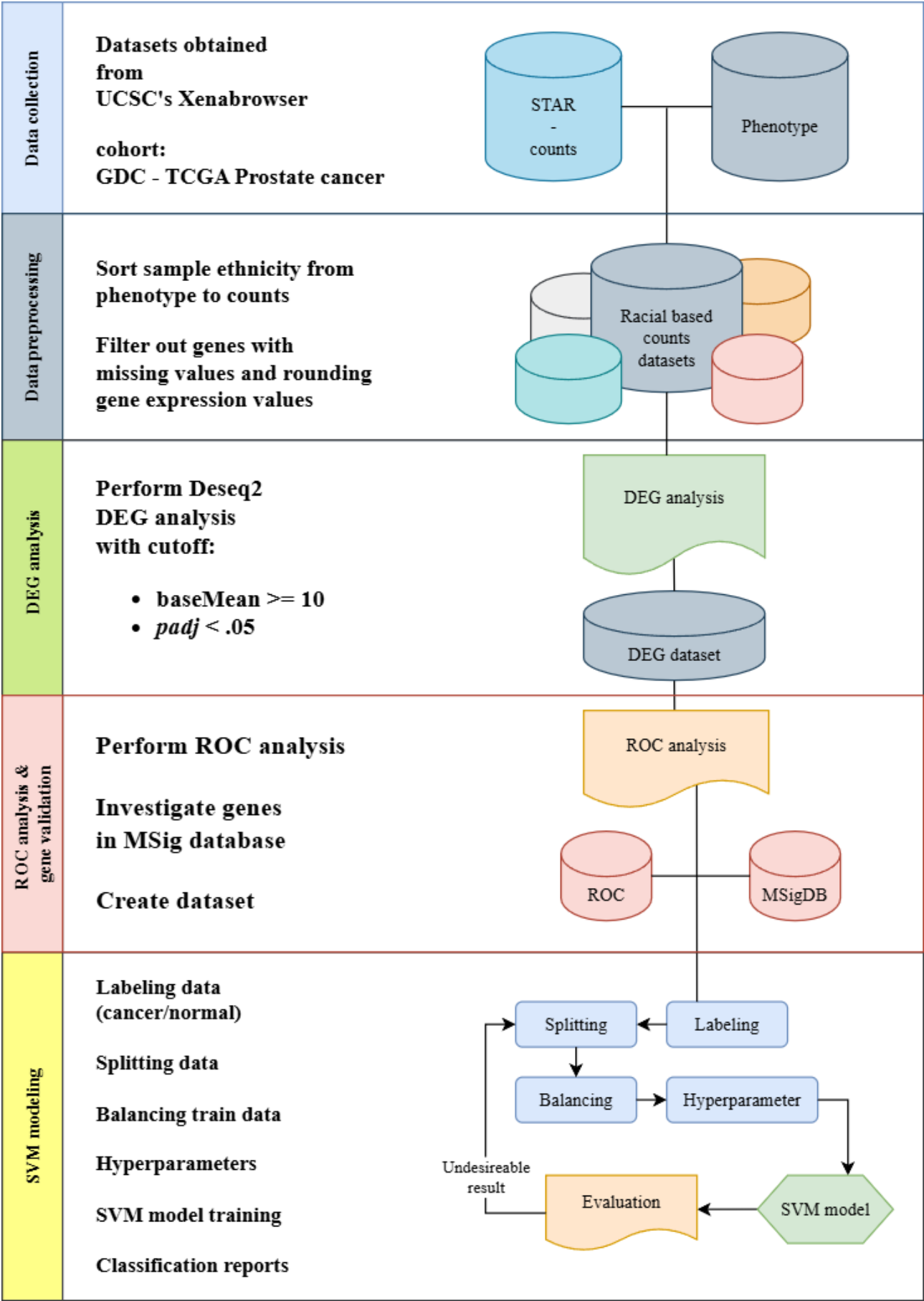
Ethical Considerations

This study used publicly available datasets from the University of California, Santa Cruz Xena [31]. University of California, Santa Cruz Xena allows users to explore functional genomic data sets for correlations between genomic or phenotypic variables. Thus, no ethics approval was required.

Data Collection

This study implements a structured methodology to identify and model significant genes for prostate cancer using gene expression data. There are 2 datasets used and obtained in August 2024 from Xenabrowser's GDC (Genomic Data Commons) TCGA-PRAD (The Cancer Genome Atlas Prostate Adenocarcinoma) cohort, of which 1 contains gene expression counts data, and the other contains the clinical information of the samples [29]. Gene expression dataset has been prenormalized by Xenabrowser using log2(count+1).

Data Preprocessing

Data preprocessing involved separating the counts dataset racially by mapping the samples to their race in the phenotype dataset, filtering samples with missing gene expression values, and labeling samples as normal or cancer via the TCGA (The Cancer Genome Atlas) barcode. These steps were conducted using the Pandas (version 2.2.2; NumFOCUS, Inc) and NumPy (version 1.26.4; NumPy Developers) libraries in a Jupyter Notebook (LF Charities) environment [32-34].

Feature Selection

Feature selection to train the machine learning model was achieved through refining the filtered genes from DEG analysis, performed using the *PyDESeq2* package (version 0.4.10; OWKIN) [35-37]. After creating metadata and the appropriate data frame, we used the DESeqDataSet function to create a suitable dataset for the DESeq2 process. There are 3 parameters used in creating the DESeqDataSet. First is counts, which is where the data frame of gene expression values of each gene ID and sample ID is used. To create metadata for the DeseqDataSet function, we specify the design of the DEG experiment and the factors to be analyzed. The factors in this research are labeled sample IDs with their condition that has been converted to a data frame by using the DeseqStats function. Lastly, we defined the design factor to guide the DEG analysis to focus on the important variables, in this case, the sample conditions. Identifying significant genes is based on the set threshold of baseMean≥10 and *p-adj*<.05. The filtered genes were used to create 5 experimental scenarios, with the first scenario focusing on the outlier genes identified through *PyDESeq2* that met the specified thresholds.

The second and third scenarios were developed by introducing additional thresholds to the DEG results. The additional scenarios further narrowed down the outlier genes by applying log2FoldChange>0.35 and >0.4, respectively.

For the fourth scenario, ROC analysis was performed using the scikit-learn metrics library (version 1.5.1; scikit-learn developers) to isolate genes with high predictive impact [38,39]. Genes were filtered based on a cutoff threshold of area under the curve value above 0.90, and the results were visualized using the matplotlib library (version 3.9.1; The Matplotlib development team) [40]. These genes were then used to create the fourth scenario.

The final scenario involves converting the isolated genes' Ensembl IDs into gene symbols using BioTools.fr for the human species Ensembl format [41] and verifying using gene set enrichment analysis (GSEA). Gene symbols were queried to MSigDB from GSEA to compute overlaps on curated gene sets which enables identification of well-established biological pathways and is widely used in cancer immunology and metabolic research, computational gene sets to complement the curated gene sets by providing unbiased large-scale insights and specific gene expression patterns, oncogenic gene sets that are directly relevant to cancer research and linked to gene expression changes on specific oncogenic events, and False Discovery Rate q-value less than 0.05 to reduce the likelihood of false positives in enrichment results [22,42-46]. Overlaps between the queried genes and the gene sets in MSigDB were analyzed to validate their relevance to prostate cancer. Genes with confirmed prostate cancer relevance were selected for use in the final scenario.

SVM Modeling

The dataset initially shows a strong class imbalance, with a cancer-to-normal ratio of 1:9. To address this class imbalance, the data were split into training and testing sets using various stratified splits: 60%/40%, 70%/30%, and 80%/20%. Stratification ensures that the class distribution among the training data class imbalance was then addressed on all the training data scenarios using oversampling methods, including RandomOverSampler, SVMSMOTE, SMOTEENN, SMOTETomek, ADASYN, BorderlineSMOTE, and KMeansSMOTE from sci-kit libraries with a sampling strategy of 0.3, meaning the training data consists of 66.66% cancer samples and 33.33% normal samples, creating a balanced dataset for model training and preserving the authenticity of the testing data, making a realistic environment for the model to perform in.

Multimedia Appendix 1 (Table S1) and Table 1 show multiple experimental scenarios that were designed to test different parameter combinations and datasets. Two modeling scenarios were used; first, using the default SVC function with linear kernel. Second, conducting hyperparameter tuning to optimize model performance. Hyperparameter tuning was performed using GridSearchCV with a linear kernel SVC classifier and 5-fold cross-validation. The hyperparameters and their ranges were as follows: multiple kernels of the SVC function were used, linear, polynomial, and radial basis function. C values were ranging from 0.01, 0.1, 1, and 10, with gamma values of 0.01, 0.1, and 1, coef0 values of 0 and 1, and lastly class weights of none and balanced.

Evaluation of the model was obtained and inspected using the *classification_report* function, by focusing on harmonization between F_1-score, recall, accuracy, precision, and macro-avg values, we evaluated the models' performance on training and test sets to ensure reliability of the model with no over- or underfitting present. To further validate the results of the obtained machine learning model, we tested the model on a

black dataset with corresponding gene amounts to further investigate the racial differences in prostate cancer. This approach aligns with the goal of improving the identification of prostate cancer within a specific population.

Table . Top 5 models for 4-gene scenario.

Balancing method	Data splitting ratio	Hyper-parameter	White					Black			
			Train accuracy (%)	Test accuracy (%)	F_1-score (%)	Precision (%)	Recall (%)	Test accuracy (%)	F_1-score (%)	Precision (%)	Recall (%)
KMeansS-MOTE	80:20	Yes	94.2	94.6	97	94.3	100	93.7	96.5	94.9	98.2
KMeansS-MOTE	70:30	No	92.8	93.5	96.5	94.6	98.4	93.7	96.5	94.9	98.2
KMeansS-MOTE	80:20	No	92.8	93.5	96.4	95.3	97.6	92.2	95.6	94.8	96.5
SVMSMOTE	80:20	Yes	94.9	92.4	95.8	95.2	96.4	92.2	95.6	96.4	94.7
KMeansS-MOTE	70:30	Yes	93.6	92	95.7	92.5	99.2	90.6	94.9	91.8	98.2

Results

Datasets

Data for this research consists of 2 correlated secondary datasets, obtained through an open-source prostate cancer gene expression database, Xenabrowser GDC TCGA gene expression RNAseq Spliced Transcripts Alignment to a Reference–counts, and Xenabrowser GDC TCGA phenotypes. Gene expression RNAseq Spliced Transcripts Alignment to a Reference–counts contains 550 samples and 60,480 gene IDs in Ensembl format. On the other hand, the phenotype dataset contains 623 rows and 127 samples of clinical information on the samples included, from which sample types and race demographics columns are used to create a dataset based on race demographics. Out of the 550 samples present in the phenotype dataset, 458 were White, 12 were Asian, 1 was American Indian, 64 were African Americans, and 15 were not reported. The filtered-out White race count data that contains 57,429 gene IDs and 458 samples with their respective classes are presented in Multimedia Appendix 1 (Table S2).

Feature Selection

To create a more enhanced feature selection method, several scenarios were made combining multiple methods based on DEG analysis thresholds. These scenarios reveal the most optimal combination of methods to identify genes relevant to prostate cancer.

From DEG analysis, various genes are extracted with several thresholds (Table S3 in Multimedia Appendix 1), the most being 139 genes. This result is further refined with ROC analysis and MSigDB investigation, which reveals 9 of 139 genes to have a direct correlation to prostate cancer.

Of the 139 genes identified through DEG analysis, PCA3 showed the strongest up-regulated correlation with prostate cancer (Table S4 in Multimedia Appendix 1). PCA3 had a baseMean of 12.33, indicating high expression across samples, a log2FoldChange of 0.6198, reflecting increased expression in cancerous tissue, and a *p-adj* value of <.001, confirming statistical significance.

Among the 139 genes identified from DEG analysis, WFDC2 has the strongest down-regulated correlation with prostate cancer (Table S5 in Multimedia Appendix 1). This is evident with a baseMean of 10.17 indicating a moderate expression level across samples, a log2FoldChange of −0.3069 which shows a decrease in expression levels in cancerous tissue compared to normal tissue, and a *p-adj*<.001 indicating high statistical significance after adjustment for multiple testing.

ROC analysis was performed on 139 genes obtained using the White race DEG analysis, applying an area under the curve score threshold above 0.9. This process identified 13 genes as outliers, as shown in Figure 2, significantly narrowing down the initial gene set.

Figure 2. A total of 13 genes were identified to have a strong correlation (AUC>0.9) with prostate cancer obtained through ROC analysis of 139 genes. AUC: area under the curve; ROC: receiver operating characteristic.
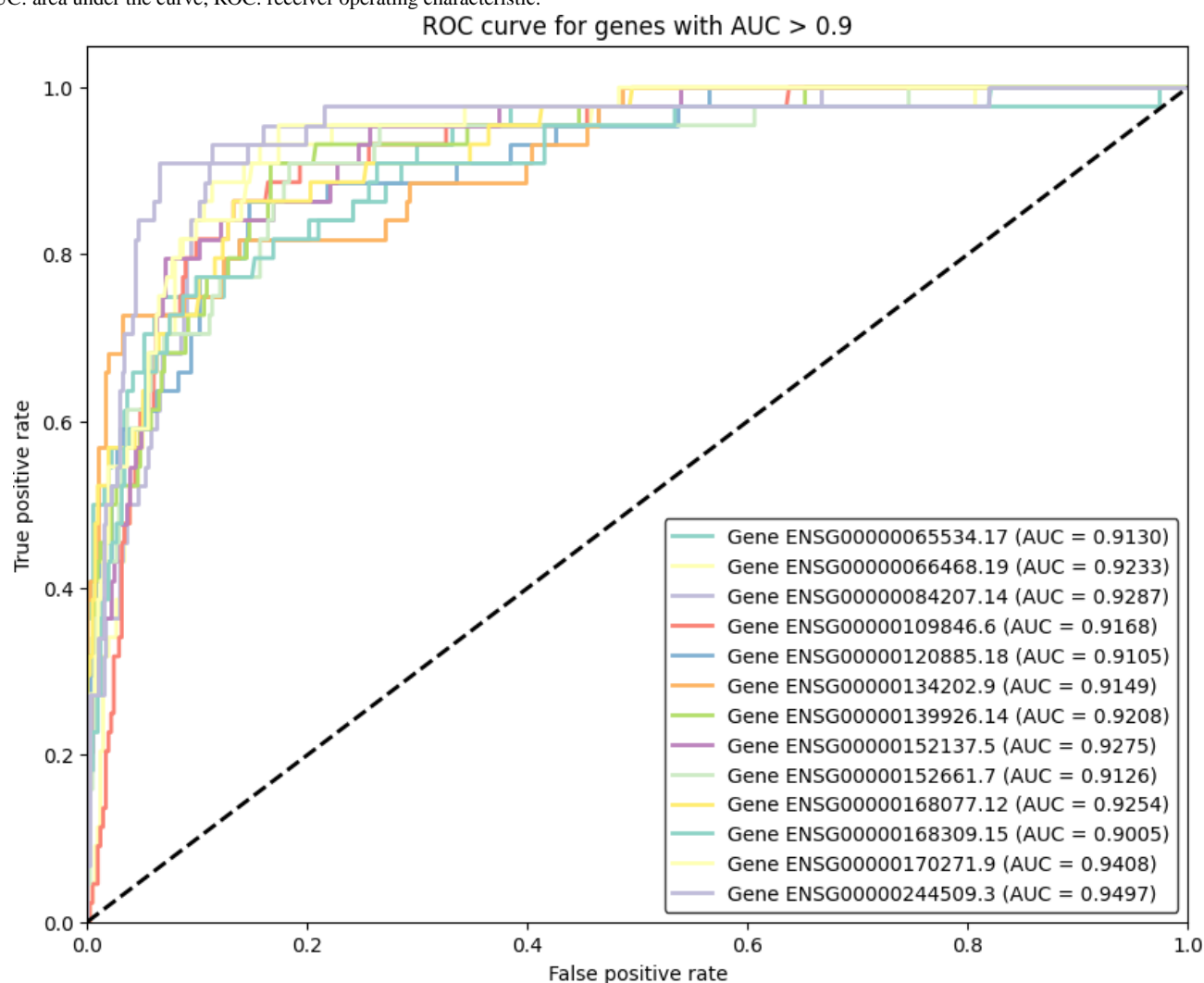
Genes that were identified from ROC analysis were converted from Ensembl format to gene symbol using BioTools.fr (Table S1 in Multimedia Appendix 1) to be verified through MSigDB.

GSEA MSigDB investigation results reveal that the genes' correlation varies between gene sets. We found that out of 13 genes, 9 were found to have a correlation to MSigDBs' LIU_PROSTATE_CANCER_DN gene set with a *P*<.001 and False Discovery Rate q-value of 2.05 e^{-11} as seen in Figure 3.

Figure 3. GSEA MSigDB investigation results of 139 genes selected from DEG analysis reveal 9 genes that are down-regulated in prostate cancer. 3CA / PIK3CA: phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha; AILT: Angioimmunoblastic T-cell lymphoma; CNS: Central Nervous System; DEG: differentially expressed gene; FDR q: False Discovery Rate q-value; GSEA: gene set enrichment analysis; HDAC:Histone Deacetylase; k/K: is a ratio of number of genes in GSEA MSigDB data set (k) divided by the number of genes in the indicated dataset (K); LIU: protein LIU; MSigDB: Molecular Signature Database; PDGFB: Platelet-Derived Growth Factor Subunit B; PTC: papillary thyroid carcinoma; RNAi: RNA interference; U2OS: a human osteosarcoma cell line;

Gene Set Name [# Genes (K)]	Description	in Overlap (k)	k/K	p-value ?	FDRq-value ?
LIU_PROSTATE_CANCER_DN [493]	Genes down-regulated in prostate cancer samples.	9		$2.39\ e^{-15}$	$2.05\ e^{-11}$
PICCALUGA_ANGIOIMMUNOBLASTIC_LYMPHOMA_MA_UP [211]	Up-regulated genes in angioimmunoblastic lymphoma (AILT) compared to normal T lymphocytes.	5		$3.55\ e^{-9}$	$1.53\ e^{-5}$
JOHANSSON_BRAIN_CANCER_EARLY_VS_LATE_DE_DN [43]	Genes down-regulated in early vs late brain tumors induced by retroviral delivery of PDGFB [GeneID=5155].	3		$2.72\ e^{-7}$	$4.65\ e^{-4}$
MODULE_11 [540]	Genes in the cancer module 11.	5		$3.81\ e^{-7}$	$4.65\ e^{-4}$
MODULE_100 [544]	Genes in the cancer module 100.	5		$3.95\ e^{-7}$	$4.65\ e^{-4}$
MODULE_137 [546]	CNS genes.	5		$4.03\ e^{-7}$	$4.65\ e^{-4}$
MODULE_66 [552]	Genes in the cancer module 66.	5		$4.25\ e^{-7}$	$4.65\ e^{-4}$
GAVISH_3CA_MALIGNANT_METAPROGRAM_25_AS_ASTROCYTES [50]	Genes upregulated in subsets of cells of a given type within various tumors	3		$4.32\ e^{-7}$	$4.65\ e^{-4}$
DELYS_THYROID_CANCER_DN [233]	Genes down-regulated in papillary thyroid carcinoma (PTC) compared to normal tissue.	4		$6.01\ e^{-7}$	$5.63\ e^{-4}$
SENESE_HDAC1_AND_HDAC2_TARGETS_DN [238]	Genes down-regulated in U2OS cells (osteosarcoma) upon knockdown of both HDAC1 and HDAC2 [GeneID=3065;3066] by RNAi.	4		$6.54\ e^{-7}$	$5.63\ e^{-4}$

SVM Classifier

Various scenarios with different balancing methods and splitting percentages were implemented for constructing the ideal SVM model, creating minimal but important differences in class counts as seen in Multimedia Appendix 1 (Table S7).

From the various scenarios, we identified the top 5 best-performing models across different feature categories. The model using 139 genes from DEG analysis combined with the SMOTEENN balancing technique achieved the most consistent results, with a training accuracy of 100% and test accuracies of 97% for the White race and 96% for the Black race, alongside strong harmonization across F_1-score, precision, and recall.

Compared to models using 4 and 7 genes, obtained through DEG analysis thresholds of log2FoldChange>0.35 and 0.4, achieved accuracies of 95% or below with unfavorable harmonization, thus the need for more advanced feature selection methods, such as ROC analysis combined with online GSEA. Models with 13 and 9 selected genes obtained through ROC analysis and GSEA demonstrated competitive performance, achieving 97% accuracy for the White race and 95% for the Black race, though slight deviations in precision and recall for the Black race were observed. Detailed metrics for all scenario models can be found from Tables 1-5.

XSL•FO
RenderX

Table . Top 5 models for 7-genes scenario.

Balancing method	Data splitting ratio	Hyper-parameter	White					Black			
			Train accuracy (%)	Test accuracy (%)	F_1-score (%)	Precision (%)	Recall (%)	Test accuracy (%)	F_1-score (%)	Precision (%)	Recall (%)
KMeansS-MOTE	80:20	Yes	94.9	95.6	97.6	95.4	100	95.3	97.4	96.5	98.2
SVMSMOTE	80:20	Yes	97.9	94.6	97	96.4	97.6	90.6	94.6	96.4	93
KMeansS-MOTE	80:20	No	94.4	94.6	97	95.3	98.8	93.7	96.5	94.9	98.2
KMeansS-MOTE	60:40	No	96	92.9	96.1	94.7	97.6	95.3	97.4	96.5	98.2
SVMSMOTE	70:30	Yes	98.7	92.7	96.1	93.9	98.4	95.3	97.4	96.5	98.2

Table . Top 5 models for 9-genes scenario.

Balancing method	Data splitting ratio	Hyper-parameter	White					Black			
			Train accuracy (%)	Test accuracy (%)	F_1-score (%)	Precision (%)	Recall (%)	Test accuracy (%)	F_1-score (%)	Precision (%)	Recall (%)
SVMSMOTE	70:30	Yes	98.4	97.1	98.4	98.4	98.4	95.3	97.3	98.2	96.5
KMeansS-MOTE	80:20	No	96.5	96.7	98.2	98.8	97.6	93.7	96.4	100	93
KMeansS-MOTE	80:20	Yes	95.6	96.7	98.2	98.8	97.6	96.9	98.2	98.2	98.2
SMOTE-Tomek	70:30	Yes	98.7	96.4	98	98.4	97.6	95.3	97.3	98.2	96.5
KMeansS-MOTE	70:30	No	95.7	96.4	98	99.2	96.8	95.3	97.3	98.2	96.5

Table . Top 5 models for 13-genes scenario.

Balancing method	Data splitting ratio	Hyper-parameter	White					Black			
			Train accuracy (%)	Test accuracy (%)	F_1-score (%)	Precision (%)	Recall (%)	Test accuracy (%)	F_1-score (%)	Precision (%)	Recall (%)
KMeansS-MOTE	70:30	No	95.2	97.1	98.4	99.2	97.6	95.3	97.3	98.2	96.5
SMOTE-Tomek	80:20	Yes	98.1	96.7	98.2	97.6	98.8	95.3	97.3	100	94.7
BorderlineS-MOTE	70:30	No	90.4	96.4	97.9	99.2	96.8	95.3	97.3	100	94.7
KMeansS-MOTE	60:40	No	96.9	96.2	97.9	98.2	97.6	92.2	95.5	98.1	93
KMeansS-MOTE	70:30	Yes	95.2	95.6	97.6	98.4	96.8	95.3	97.3	98.2	96.5

Table . Top 5 models for the 139 genes scenario.

Balancing method	Data splitting ratio	Hyper-parameter	White					Black			
			Train accuracy (%)	Test accuracy (%)	F_1-score (%)	Precision (%)	Recall (%)	Test accuracy (%)	F_1-score (%)	Precision (%)	Recall (%)
SMO-TEENN	80:20	No	100	97.8	98.8	98.8	98.8	96.9	98.2	100	96.5
Borderli-neS-MOTE	60:40	Yes	98.8	97.3	98.5	99.4	97.6	96.9	98.2	100	96.5
SMO-TEENN	70:30	Yes	100	97.1	98.4	99.2	97.6	96.9	98.2	100	96.5
SMO-TEENN	70:30	No	100	97.1	98.4	99.2	97.6	96.9	98.2	100	96.5
SMO-TEENN	80:20	Yes	100	96.7	98.2	98.8	97.6	96.9	98.2	100	96.5

Discussion

Principal Results

In this study, we explored multiple feature selection scenarios for race-based SVM classification models aimed at prostate cancer detection using gene expression data. Our findings demonstrate that race-based models with significantly reduced features are capable of achieving competitive performance comparable to models using thousands of genes. The best-performing model, achieved without hyperparameter tuning or cross-validation, demonstrated outstanding results with a training accuracy of 100% and test accuracies of 98% on the White race and 97% on the Black race. Additionally, the model showed strong harmonization across F_1-score, precision, and recall values, which indicates consistent model classification performance. However, models in scenarios with 4 and 7 genes, selected using DEG analysis with thresholds of log2FoldChange>0.35 and 0.4, respectively, showed lower accuracies of 95% or lower, despite noteworthy harmonization between F_1-score, precision, and recall values. This shows the limitations of feature selection solely using DEG analysis thresholds, as it failed to capture the critical biomarkers necessary for reliable classification.

Moreover, models with 9 and 13 selected genes through ROC analysis and GSEA present matched performance, achieving accuracies of 97% on the White race and 95% on the Black race. These models also demonstrated good stability, consistently performing well over different train-test dataset splits. While these reduced-feature models showed strong metrics for the White race, the slight drop in accuracy for the Black race indicates the presence of racial disparities in feature selection. This highlights the need for further research to improve model generalizability across more diverse populations.

Strengths

This study addresses racial disparities in prostate cancer gene expression datasets to create a race-specific SVM classification model with multiple scenarios. Our testing demonstrated greater accuracies on scenarios using 139 genes; however, models with

13 and 9 selected genes also yielded 97% accuracy, highlighting the effectiveness of an optimized feature selection strategy. This feature reduction implies the significance of feature selection along with model construction parameters such as balancing methods, data splitting ratios, and hyperparameter optimization in achieving a robust classification model.

From a clinical standpoint, these results imply significant cost reduction and practical applicability. Reducing the number of genes required for sequencing substantially lowers the financial and computational cost of diagnostic workflows, making this approach more accessible and scalable for routine prostate cancer screening and early detection [47-49].

Comparison With Prior Works

While prior works used feature selection methods with correlation-based and evolutionary algorithm approaches without further validations, our approach used tools such as *PyDESeq2* and MSigDB investigation to further validate the biological relevance of our selected genes to prostate cancer to improve the diagnostic accuracy and provide insights into race-specific prostate cancer biology, an area often neglected by other studies.

Our study achieved comparable accuracies to prior works while significantly reducing the number of features used. For example, Ravindran et al [29] reported a 97% accuracy while using 1833 features selected from the initial 6033 genes through a correlation-based approach [27]. Conversely, our models achieved similar accuracy using only 13 or 9 features, validating the performance of our feature selection method. Additionally, our study integrates racially based datasets to account for racial disparities while achieving robust performance for both the White (98% accuracy) and Black populations (97% accuracy). This further addresses the gap between prior works such as the model by Alshareef et al [27], with 52 prostate cancer samples and 1833 features, which overlook racial disparities [28]. To further appraise our model, we also compared it to a recent study by Xie and Xie [50] using an artificial neural network model on a DEG panel of 220 genes and reporting an accuracy of 78%, our optimized racial-based SVM model outperformed it with higher accuracy and fewer features, while maintaining consistent results across multiple dataset splits. These comparisons

XSL•FO

RenderX

highlight the competitiveness and reliability of our SVM-based framework in prostate cancer detection.

Limitations

However, this study has the following limitations. The datasets used are heavily imbalanced, with an overrepresentation of White individuals and cancer samples compared to normal samples. Only a single dataset source was used due to restricted access to other publicly available datasets, which limits the diversity and variability of the data. Future work should prioritize the inclusion of larger, more diverse populations to enhance the model's generalizability and consider an external independent dataset to validate the model's performance. Additionally, exploring other genomic and epigenomic features, such as DNA methylation patterns, may yield further insights into race-specific prostate cancer biology.

Conclusions

This research used enhanced feature selection methods such as DESeq2 DEG analysis and ROC analysis to reduce feature quantity in machine learning models for prostate cancer detection in specific racial groups. Our findings show that while testing on White race reducing features-maintained model, performance was comparable to studies with larger feature sets. To examine racial disparities, we tested the model on African American data, revealing minimal (~1%) accuracy differences between racial groups. These findings indicate a low influence of racial features on classification while emphasizing the importance of feature selection in developing race-based SVM models for prostate cancer using gene expression data.

Authors' Contributions

Conceptualization: DA (lead), MVO (equal), MW (equal)
Data curation: AM (lead), VK (sopporting), JS (supporting)
Formal analysis: AM
Funding acquisition: DA
Investigation: AM
Methodology: DA (lead), MVO (equal), MW (equal), MIA (equal)
Project administration: AM (lead), VK (sopporting), JS (supporting)
Resources: AM (lead), VK (sopporting), JS (supporting)
Supervision: DA (lead), MVO (equal), MW (equal), MIA (equal)
Validation: DA (lead), MVO (equal), MW (equal), MIA (equal)
Visualization: AM (lead), VK (sopporting), JS (supporting)
Writing – original draft: AM (lead), VK (sopporting), JS (supporting)
Writing – review & editing: AM (lead), VK (sopporting), JS (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Tables on modeling scenarios, dataset, genes, and the machine learning training model.
[DOC File, 88 KB - bioinform_v6i1e72423_app1.doc]

References

1. Cook MB, Beachler DC, Parlett LE, et al. Testosterone therapy in relation to prostate cancer in a U.S. commercial insurance claims database. Cancer Epidemiol Biomarkers Prev 2020 Jan 1;29(1):236-245. [doi: 10.1158/1055-9965.EPI-19-0619]
2. Wang M, Chi G, Bodovski Y, et al. Temporal and spatial trends and determinants of aggressive prostate cancer among Black and White men with prostate cancer. Cancer Causes Control 2020 Jan;31(1):63-71. [doi: 10.1007/s10552-019-01249-0]
3. Iqbal S, Siddiqui GF, Rehman A, et al. Prostate cancer detection using deep learning and traditional techniques. IEEE Access 2021;9:27085-27100. [doi: 10.1109/ACCESS.2021.3057654]
4. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021 May;71(3):209-249. [doi: 10.3322/caac.21660] [Medline: 33538338]
5. Castaldo R, Cavaliere C, Soricelli A, Salvatore M, Pecchia L, Franzese M. Radiomic and genomic machine learning method performance for prostate cancer diagnosis: systematic literature review. J Med Internet Res 2021 Apr 1;23(4):e22394. [doi: 10.2196/22394] [Medline: 33792552]

XSL•FO
RenderX

6. Albertsen PC, Hanley JA, Gleason DF, Barry MJ. Competing risk analysis of men aged 55 to 74 years at diagnosis managed conservatively for clinically localized prostate cancer. JAMA 1998 Sep 16;280(11):975-980. [doi: 10.1001/jama.280.11.975] [Medline: 9749479]

7. Dess RT, Hartman HE, Mahal BA, et al. Association of Black race with prostate cancer-specific and other-cause mortality. JAMA Oncol 2019 Jul 1;5(7):975-983. [doi: 10.1001/jamaoncol.2019.0826] [Medline: 31120534]

8. Lachance J, Berens AJ, Hansen MEB, Teng AK, Tishkoff SA, Rebbeck TR. Genetic hitchhiking and population bottlenecks contribute to prostate cancer disparities in men of African descent. Cancer Res 2018 May 1;78(9):2432-2443. [doi: 10.1158/0008-5472.CAN-17-1550] [Medline: 29438991]

9. Zhang W, Dong Y, Sartor O, Flemington EK, Zhang K. SEER and gene expression data analysis deciphers racial disparity patterns in prostate cancer mortality and the public health implication. Sci Rep 2020 Apr;10(1):6820. [doi: 10.1038/s41598-020-63764-4]

10. Sarkar S, Das S. A review of imaging methods for prostate cancer detection. Biomed Eng Comput Biol 2016;7(Suppl 1):1-15. [doi: 10.4137/BECB.S34255] [Medline: 26966397]

11. Naji L, Randhawa H, Sohani Z, et al. Digital rectal examination for prostate cancer screening in primary care: a systematic review and meta-analysis. Ann Fam Med 2018 Mar;16(2):149-154. [doi: 10.1370/afm.2205] [Medline: 29531107]

12. Barry MJ. Clinical practice. prostate-specific-antigen testing for early diagnosis of prostate cancer. N Engl J Med 2001 May 3;344(18):1373-1377. [doi: 10.1056/NEJM200105033441806] [Medline: 11333995]

13. 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 Jul;5:e56538. [doi: 10.2196/56538]

14. Ng KLS, Mishra SK. De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. Bioinformatics 2007 Jun 1;23(11):1321-1330. [doi: 10.1093/bioinformatics/btm026] [Medline: 17267435]

15. Akinnuwesi BA, Olayanju KA, Aribisala BS, et al. Application of support vector machine algorithm for early differential diagnosis of prostate cancer. Data Sci Manag 2023 Mar;6(1):1-12. [doi: 10.1016/j.dsm.2022.10.001]

16. Alharbi F, Vakanski A. Machine learning methods for cancer classification using gene expression data: a review. Bioengineering (Basel) 2023 Jan 28;10(2):173. [doi: 10.3390/bioengineering10020173] [Medline: 36829667]

17. Khalsan M, Machado LR, Al-Shamery ES, et al. A survey of machine learning approaches applied to gene expression analysis for cancer prediction. IEEE Access 2022;10:27522-27534. [doi: 10.1109/ACCESS.2022.3146312]

18. Xiao J, Mo M, Wang Z, et al. The application and comparison of machine learning models for the prediction of breast cancer prognosis: retrospective cohort study. JMIR Med Inform 2022 Feb 18;10(2):e33440. [doi: 10.2196/33440] [Medline: 35179504]

19. Anna A, Monika G. Splicing mutations in human genetic disorders: examples, detection, and confirmation. J Appl Genet 2018 Aug;59(3):253-268. [doi: 10.1007/s13353-018-0444-7] [Medline: 29680930]

20. Alhenawi E, Al-Sayyed R, Hudaib A, Mirjalili S. Feature selection methods on gene expression microarray data for cancer classification: a systematic review. Comput Biol Med 2022 Jan;140:105051. [doi: 10.1016/j.compbiomed.2021.105051] [Medline: 34839186]

21. Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A. Distributed feature selection: an application to microarray data classification. Appl Soft Comput 2015 May;30:136-150. [doi: 10.1016/j.asoc.2015.01.035]

22. Gomes R, Paul N, He N, Huber AF, Jansen RJ. Application of feature selection and deep learning for cancer prediction using DNA methylation markers. Genes (Basel) 2022 Aug 29;13(9):1557. [doi: 10.3390/genes13091557] [Medline: 36140725]

23. Sheikhpour R, Berahmand K, Mohammadi M, Khosravi H. Sparse feature selection using hypergraph Laplacian-based semi-supervised discriminant analysis. Pattern Recognit DAGM 2025 Jan;157:110882. [doi: 10.1016/j.patcog.2024.110882]

24. Dai B, Xu Z, Li H, Wang B, Cai J, Liu X. Racial bias can confuse AI for genomic studies. Oncologie (Paris) 2022;24(1):113-130. [doi: 10.32604/oncologie.2022.020259]

25. Kapur S. Reducing racial bias in AI models for clinical use requires a top-down intervention. Nat Mach Intell 2021 Jun;3(6):460-460. [doi: 10.1038/s42256-021-00362-7]

26. Monterroso P, Moore KJ, Sample JM, Sorajja N, Domingues A, Williams LA. Racial/ethnic and sex differences in young adult malignant brain tumor incidence by histologic type. Cancer Epidemiol 2022 Feb;76:102078. [doi: 10.1016/j.canep.2021.102078] [Medline: 34896933]

27. Alshareef AM, Alsini R, Alsieni M, et al. Optimal deep learning enabled prostate cancer detection using microarray gene expression. J Healthc Eng 2022;2022:7364704. [doi: 10.1155/2022/7364704] [Medline: 35310199]

28. Zhu L, Wang H, Jiang C, et al. Clinically applicable 53-gene prognostic assay predicts chemotherapy benefit in gastric cancer: a multicenter study. EBioMedicine 2020 Nov;61:103023. [doi: 10.1016/j.ebiom.2020.103023] [Medline: 33069062]

29. Ravindran U, Gunavathi C. Deep learning assisted cancer disease prediction from gene expression data using WT-GAN. BMC Med Inform Decis Mak 2024 Oct 24;24(1):311. [doi: 10.1186/s12911-024-02712-y] [Medline: 39449042]

30. Van Rossum G, Drake FL. Python 3 Reference Manual: CreateSpace; 2009.

31. Welcome to the xena functional genomics explorer. UCSC Xena. URL: https://xenabrowser.net/ [accessed 2025-07-04]

XSL•FO

RenderX

32. Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. Nature New Biol 2020 Sep;585(7825):357-362. [doi: 10.1038/s41586-020-2649-2] [Medline: 32939066]

33. McKinney W. Data structures for statistical computing in python. 2010 Presented at: Python in Science Conference; Jun 28 to Jul 3, 2010; Austin, Texas p. 51-56. [doi: 10.25080/Majora-92bf1922-00a]

34. Kluyver T, Benjain RK, Fernando P, et al. Jupyter notebooks – a publishing format for reproducible computational workflows. In: Loizides F, Schmidt B, editors. Positioning and Power in Academic Publishing: Players, Agents and Agendas: IOS Press; 2016:87-90.

35. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15(12):550. [doi: 10.1186/s13059-014-0550-8] [Medline: 25516281]

36. Risk MC, Knudsen BS, Coleman I, et al. Differential gene expression in benign prostate epithelium of men with and without prostate cancer: evidence for a prostate cancer field effect. Clin Cancer Res 2010 Nov 15;16(22):5414-5423. [doi: 10.1158/1078-0432.CCR-10-0272] [Medline: 20935156]

37. Gunasekaran H, Ramalakshmi K, Arokiaraj ARM, Kanmani SD, Venkatesan C, Dhas CSG. Analysis of DNA sequence classification using CNN and hybrid models. Comput Math Methods Med 2021;2021:1835056. [doi: 10.1155/2021/1835056] [Medline: 34306171]

38. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011 Nov;12:2825-2830 [FREE Full text]

39. Hou C, Zhong X, He P, et al. Predicting breast cancer in Chinese women using machine learning techniques: algorithm development. JMIR Med Inform 2020 Jun 8;8(6):e17364. [doi: 10.2196/17364] [Medline: 32510459]

40. Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng 2007;9(3):90-95. [doi: 10.1109/MCSE.2007.55]

41. da Silva AR, Malafaia G, Menezes IPP. Biotools: an R function to predict spatial gene diversity via an individual-based approach. Genet Mol Res 2017 Apr 13;16(2):2. [doi: 10.4238/gmr16029655] [Medline: 28407196]

42. Goldman MJ, Craft B, Hastie M, et al. Visualizing and interpreting cancer genomics data via the Xena platform. Nat Biotechnol 2020 Jun;38(6):675-678. [doi: 10.1038/s41587-020-0546-8] [Medline: 32444850]

43. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 2005 Oct 25;102(43):15545-15550. [doi: 10.1073/pnas.0506580102]

44. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics 2011 Jun 15;27(12):1739-1740. [doi: 10.1093/bioinformatics/btr260] [Medline: 21546393]

45. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst 2015 Dec 23;1(6):417-425. [doi: 10.1016/j.cels.2015.12.004] [Medline: 26771021]

46. Castanza AS, Recla JM, Eby D, Thorvaldsdóttir H, Bult CJ, Mesirov JP. The molecular signatures database revisited: extending support for mouse data. bioRxiv. Preprint posted online on Oct 25, 2022. [doi: 10.1101/2022.10.24.513539]

47. Pruneri G, De Braud F, Sapino A, et al. Next-generation sequencing in clinical practice: is it a cost-saving alternative to a single-gene testing approach? Pharmacoecon Open 2021 Jun;5(2):285-298. [doi: 10.1007/s41669-020-00249-0] [Medline: 33660227]

48. Stoddard JL, Niemela JE, Fleisher TA, Rosenzweig SD. Targeted NGS: a cost-effective approach to molecular diagnosis of PIDs. Front Immunol 2014;5:531. [doi: 10.3389/fimmu.2014.00531] [Medline: 25404929]

49. Ndiaye M, Prieto-Baños S, Fitzgerald LM, et al. When less is more: sketching with minimizers in genomics. Genome Biol 2024 Oct 14;25(1):270. [doi: 10.1186/s13059-024-03414-4] [Medline: 39402664]

50. Xie Y, Xie J. Integrates differential gene expression analysis and deep learning for accurate and robust prostate cancer diagnosis. ACE 2024;57(1):66-74. [doi: 10.54254/2755-2721/57/20241312]

Abbreviations

DEG: differentially expressed gene
GDC : Genomic Data Commons
GSEA: gene set enrichment analysis
MGECD: Micro Gene Expression Cancer Dataset
MSigDB: Molecular Signature Database
PSA: prostate-specific antigen
ROC: receiver operating characteristic
SVM: support vector machine
TCGA: The Cancer Genome Atlas
TCGA-PRAD: The Cancer Genome Atlas Prostate Adenocarcinoma

XSL•FO
RenderX

Estimating Antigen Test Sensitivity via Target Distribution Balancing: Development and Validation Study

Miguel Bosch[1,2], PhD; Adriana Moreno[2], MSc; Raul Colmenares[2], MSc; Jose Arocha[1], MSc; Sina Hoche[1], PhD; Auris Garcia[1], BSc; Daniela Hall[1], PhD; Dawlyn Garcia[1], MSc; Lindsey Rudtner[1], BSc; Nol Salcedo[1], MSc; Irene Bosch[1], PhD

[1]IDX20 Inc, 166 Clinton Road, Brookline, MA, United States

[2]Info Analytics Innovations LLC, Houston, TX, United States

Corresponding Author:
Irene Bosch, PhD
IDX20 Inc, 166 Clinton Road, Brookline, MA, United States

Abstract

Background: Sensitivity—expressed as percent positive agreement (PPA) with a reference assay—is a primary metric for evaluating lateral-flow antigen tests (ATs), typically benchmarked against a quantitative reverse transcription polymerase chain reaction (qRT-PCR). In SARS-CoV-2 diagnostics, ATs detect nucleocapsid protein, whereas qRT-PCR detects viral RNA copy numbers. Since observed PPA depends on the underlying viral load distribution (proxied by the number of cycle thresholds [Cts], which is inversely related to load), study-specific sampling can bias sensitivity estimates. Cohort differences—such as enrichment for high- or low-Ct specimens—therefore complicate cross-test comparisons, and real-world datasets often deviate from regulatory guidance to sample across the full concentration range. Although logistic models relating test positivity to Ct are well described, they are seldom used to reweight results to a standardized reference viral load distribution. As a result, reported sensitivities remain difficult to compare across studies, limiting both accuracy and generalizability.

Objective: The aim of this study was to develop and validate a statistical methodology that estimates the sensitivity of ATs by recalibrating clinical performance data—originally obtained from uncontrolled viral load distributions—against a standardized reference distribution of target concentrations, thereby enabling more accurate and comparable assessments of diagnostic test performance.

Methods: AT sensitivity is estimated by modeling the PPA as a function of qRT-PCR Ct values (PPA function) using logistic regression on paired test results. Raw sensitivity is the proportion of AT positives among PCR-positive samples. Adjusted sensitivity is calculated by applying the PPA function to a reference Ct distribution, correcting for viral load variability. This enables standardized comparisons across tests. The method was validated using clinical data from a community study in Chelsea, Massachusetts, demonstrating its effectiveness in reducing sampling bias.

Results: Over a 2-year period, paired ATs and qRT-PCR–positive samples were collected from 4 suppliers: A (n=211), B (n=156), C (n=85), and D (n=43). Ct value distributions varied substantially, with suppliers A and D showing lower Ct (high viral load) values in the samples, and supplier C skewed toward higher Ct values (low viral load). These differences led to inconsistent raw sensitivity estimates. To correct for this, we used logistic regression to model the PPA as a function of Cts and applied these models to a standardized reference Ct distribution. This adjustment reduced bias and enabled more accurate comparisons of test performance across suppliers.

Conclusions: We present a distribution-aware framework that models PPA as a logistic function of Ct and reweights results to a standardized reference Ct distribution to produce bias-corrected sensitivity estimates. This yields fairer, more consistent comparisons across AT suppliers and studies, strengthens quality control, and supports regulatory review. Collectively, our results provide a robust basis for recalibrating reported sensitivities and underscore the importance of distribution-aware evaluation in diagnostic test assessment.

Trial Registration: ClinicalTrials.gov NCT05884515; https://clinicaltrials.gov/study/NCT05884515

XSL•FO
RenderX

Introduction

Antigen tests (ATs) have been a common tool utilized to provide evidence for diagnosis and health care decisions and have been used as such for several decades [1-4]. During the COVID-19 pandemic, the worldwide use of ATs demonstrated relevance for disease monitoring and diagnosis of new cases [5,6]. ATs are rapid, economical, and portable; can be self-administered; are quick to develop; and provide direct evidence for the presence of the pathogen in the tested sample—this combination is not equaled by more sophisticated laboratory tests. Due to its increasing use, it is important to accurately evaluate the AT performance for quality and regulatory control.

The common statistic used to evaluate the positive agreement performance of ATs has been sensitivity, calculated over a set of samples known to be positive by a gold standard reference. However, the sensitivity of ATs is known to be strongly dependent on the sample viral load [7-9]. Hence, the sensitivity per se is not an appropriate measure of the AT positive agreement performance because it is largely dependent on the distribution of the viral load of statistical support (ie, the set of samples used to calculate the statistic). Instead, a description of the percent positive agreement (PPA) as a function of viral load is a more accurate measure of positive agreement performance. The PPA function (PPAf) is commonly calculated with a logistic regression of the binary test result on positive agreement (1=agreement, 0=disagreement) against a variable related to the viral load. In the case of COVID-19, the viral load is commonly measured with the quantitative reverse transcription polymerase chain reaction (qRT-PCR) cycle threshold (Ct) result [10,11]; hence, the PPA is a function of the Cts.

Once the PPAf of a given AT supplier's product has been estimated from collected data in real-world application conditions, it is straightforward to estimate the expected sensitivity for any given Ct distribution or Ct sample set. This is particularly useful for equalizing the expected sensitivity to a common standard or reference distribution of Cts for a comparison of the performance across suppliers for product quality or regulatory purposes. This process removes the bias introduced into sensitivity by the circumstantial uneven representation of viral load in the statistical support (ie, the data used to calculate the statistic).

Common methods used to calculate the sensitivity of an AT product, whether for regulatory compliance or clinical diagnostic research purposes, typically involve collecting real-world test results paired with the qRT-PCR gold standard. However, in the case of lateral-flow ATs, sensitivity uncertainty does not adhere to a straightforward Bernoulli process, as the underlying positive agreement probability is not constant and is instead conditioned by viral load. To accurately calculate sensitivity, it would be necessary to segment the collected samples based on the most influential variable affecting underlying probability, using a standard reference histogram. In our case, the influential variable is the viral load, measured as Cts. Yet, implementing such a process in the field with real-world data would be cumbersome and would require a much larger dataset. The proposed balancing method overcomes this challenge by adjusting the raw sensitivity calculation to any desired standardized reference distribution of the viral load, without the need for extended data collection and segmentation.

Methods

Study Description

We conducted a study of AT use in real-world conditions in the city of Chelsea, Massachusetts, during the years 2022‐2023. The objectives of the study were multifold: (1) performing frequent COVID-19 testing at 2 vulnerable population sites (elderly housing), (2) evaluating the performance of ATs from different suppliers in the laboratory and in the real-world context, (3) collecting longitudinal (time series) AT data for qRT-PCR positive samples, and (4) implementing a digital AT data collection platform.

The participants of the study were enrolled after consent. The participant was provided with a single self-testing AT of any of the available 4 participating suppliers. The tests were home tests, and data were self-logged by the participants into an internet-based informatics platform. The participants registered the AT results (their own assessment) and uploaded a photograph of the test after completion (15 min). The fraction of positive AT tests was followed daily with paired qRT-PCR testing. A random number of negative ATs were also analyzed by PCR.

Support personnel were available on scheduled days throughout the week within the community to provide devices, training, and participant follow-up. This minimized potential confounding factors related to test interpretation and self-reporting via the app. In addition to textual reports, participants uploaded an image of their test result, allowing our team to perform retrospective verification. Discrepancies between self-reported results and staff-reviewed interpretations were found to be negligible, occurring in 5 eye assessments of 500 positive tests.

The data analyzed come from ATs provided by 4 different suppliers, labeled A through D, and the corresponding qRT-PCR test results, all of which were processed at a Clinical Laboratory Improvement Amendments–certified laboratory using the same RNA extraction as well as PCR protocol. The total negative qRT-PCR tests were 57 for A, 91 for B, 145 for C, and 114 for D, and positive qRT-PCR tests were 211 for A, 156 for B, 84 for C, and 43 for D. Each participant was tested with a single brand (ie, AT supplier), so the distribution of viral load could be different between the data collected for each brand. These data were used to calculate AT performance statistics and demonstrate the methodology for distribution-balanced sensitivity according to a reference standard distribution. Lineage annotations used in this study for laboratory assays were Delta comparator (B.1.617.2/AY sublineages) using the US WA 1/1 isolate and Omicron (BA.5). Among Chelsea study participants in 2022-2023, we detected BA.1.1, BA.2, BA.2.12.1, BA.5, BA.5.1, BA.5.2, BA.5.2.1, BQ.1.1, BQ.1.1.4, BQ.1.1.5, BQ.1.14, and XBB.1.5. Clinical sample sequences were generated at the Broad Institute.

Ethical Considerations

This study was reviewed and approved by the Advarra Institutional Review Board (IRB) for protocol "Center of Complex Interventions – IDx20-001, Community frequent antigen testing to monitor COVID-19 in senior public housing setup" (Pro00059157). The most recent continuing review approval was granted on November 13, 2023, with approval through November 13, 2024. Advarra is registered with the Office for Human Research Protections/Food and Drug Administration (IRB #00000971) and conducts reviews in accordance with US HHS 45 CFR 46 and FDA 21 CFR 50/56. All procedures adhered to institutional and national ethical standards and the World Medical Association Declaration of Helsinki. In accordance with JMIR Publications requirements, the IRB review outcome is explicitly reported here.

Before any study procedures, participants were informed of the study purpose, procedures, potential risks and benefits, data uses, and their right to withdraw without penalty. Written informed consent was obtained from each participant using IRB-approved consent materials. No monetary compensation was given to participants. All study personnel completed human-subjects protection training prior to participant interaction.

Data were collected and stored under IRB-approved procedures to protect privacy and confidentiality. Only deidentified or aggregated data are presented in this study; no identifiable personal information is reported.

Any protocol amendments, consent-form changes, or reportable events (eg, unanticipated problems; adverse device effects; or protocol deviations that could affect participant rights, safety, or data integrity) were submitted to Advarra in accordance with IRB requirements before implementation.

PPA Function

As commonly used, AT operative reading involves steps of device reaction to the nasal swab sample, a waiting time after the sample is deposited, an observation using the naked eye, and interpretation of the results by the user. For all kits utilized (eg, cassettes), the user observed the presence or absence of a colored line in a test area (*test band*) on a nitrocellulose strip of the lateral-flow test. The result is considered *positive* when the test band visualized (a color band) and can be distinguished from the no color or white background even if the color signal is faint, while a *negative* result is when the test band cannot be visualized by the user. For performance statistics, the result of the AT provided to the user is compared to the gold standard reference test [12], a COVID-19 qRT-PCR test conducted in a state-approved clinical laboratory.

For the positive agreement analysis of each AT supplier dataset, we compare only the AT results that have a paired positive qRT-PCR result (ie, having a positive qRT-PCR result in a parallel swab sample taken the same day and time as the AT swab sample). For the logistic regression analysis, we identified the AT results with a binary variable: 1 for a positive result (agreement with the standard test) and 0 for a negative result (disagreement with the standard test). The outcome of the user assessment can be described by a binary random variable. We modeled the PPAf with a logistic function, having the qRT-PCR Ct as the function domain (ie, the independent variable). Logistic regression is a well-known analysis to estimate the probability as a function of a dependent variable [13]. It has been used to describe the probability of positive agreement in ATs [14].

In addition to estimating the PPAf that characterized each AT supplier, the regression also accounts for the uncertainties of the probability function and parameters. We implement the logistic regression with a Bayesian approach, combining the objectives of (1) fitting the binary observed data and (2) honoring the Clopper-Pearson binomial confidence limits at the raw Ct data sensitivity prediction. Hence, the posterior model uncertainty description ensures compliance with the Clopper-Pearson confidence limits for the sensitivity at the raw Ct data. The numerical calculations are performed by Markov chain Monte Carlo methods.

The logistic model is a well-established method for modeling binary outcomes in which the probability, p_x, varies with a predictor variable, x, particularly when the probability increases monotonically with x. It assumes that the log-odds of the outcome (ie, the logit transformation of the probability) is a linear function of the predictor. Compared to the traditional approach to raw sensitivity estimation—which treats test outcomes as the result of a Bernoulli process with constant (homogeneous) probability—the logistic model offers a first-order improvement by accounting for the dependence of the outcome probability on the target concentration. Although alternative parametric models could be used to describe the relationship between probability and the predictor (eg, probit, splines), the logistic model provides a widely accepted and practical framework for improving sensitivity estimation accuracy. Logistic and probit models both yield nearly identical monotone dose-response curves; we selected the logistic model for its interpretability in terms of log-odds, its ability to directly estimate the concentration at which PPA=0.5, and its widespread use in diagnostics. Spline approaches are not appropriate in this context, as they do not satisfy the boundary conditions of approaching 0 at the lower end and 1 at the upper end of the concentration axis.

Distribution-Balanced Sensitivity Method

The sensitivity, s, is the fraction of the positive agreement cases divided by the total positive cases in the experimental gold standard results (eg, collected real-world AT binary data on positive qRT-PCR cases). Based on the PPAf characterized for each AT supplier data, we can estimate the sensitivity for any set of Ct cases. Let us consider that the experimental data for a given AT supplier involves N cases with qRT-PCR cycle counts, $x=\{x_1,x_2,\ldots x_n,\ldots,x_N\}$. The expected value of the sensitivity is the average of the PPAf, p_x, over the cases:

$$(1) \quad E(s)=\frac{1}{N}\sum_{1}^{N}p(x_n)$$

Likewise, the expected sensitivity over a data support with any Ct probability density function (PDF), $g(x)$, is given by the probability product integration:

$$. (2) \quad E(s)=\int_{x0}^{xf}p(x)g(x)dx$$

Equations 1 and 2 use the PPAf to calculate the expected sensitivity over a specific Ct data or distribution. Adequate

comparison of sensitivity across different AT datasets requires a transformation of raw sensitivity (ie, calculated from the observed data) to the expected sensitivity over a common reference of Ct data values, or a Ct support distribution (ie, histogram or PDF). Considering observed binary data for several AT suppliers, our proposed process to equalize the sensitivity support involves (1) estimating the PPAfs by logistic regression of the observed AT binary data for each one of the supplier's datasets; (2) defining a common reference Ct distribution by a PDF, gx; and (3) calculating for each AT supplier dataset the estimated sensitivity over the common reference Ct support by equation (2).

Equation 2 can be evaluated by discretizing the Ct domain. Alternatively, it can be computed by Monte Carlo integration,

drawing Ct realizations from the target distribution and applying equation 1 to each draw. To balance viral load distributions across assays, we select a reference Ct sample set as the empirical Ct distribution from supplier A and evaluate all other suppliers over this common range. The procedure is as follows: (1) for each nonreference supplier, fit a logistic regression for PPA as a function of Ct using the observed binary AT outcomes and (2) compute expected sensitivity by averaging the fitted PPA over supplier A's Ct values via equation 1.

The described viral load balance processes removed the effect of the Ct distribution on sensitivity, providing a common base for comparison and evaluation of the test performance.

Table 1 presents a glossary of terms and their corresponding meanings.

Table . Glossary of relevant terms and corresponding meanings.

Term	Definition
Antigen test (AT)	Lateral flow antigen test
Target	Specific analyte that the test is designed to detect (protein present in a biological matrix sample)
Cycle threshold (Ct)	qRT-PCR[a] cycle at which the fluorescence signal crosses a set threshold above background
Target concentration	It is the concentration of the target protein, expressed by cycle thresholds or ng mL^{-1}, or by plaque forming units of virus mL^{-1}
Target concentration distribution	The distribution of target concentrations among the tested population; when grouped into concentration ranges, it can be represented as a histogram.
Sensitivity	Proportion of positive cases detected by the antigen tests according to the reference qRT-PCR gold standard. It represents the percent positive agreement.
Percent positive agreement (PPA)	In a Bernoulli process, it is the probability that the test outcome is positive; in the context of antigen tests, this probability varies with the target concentration.
Percent positive agreement function (PPAf)	It is a function assigning the value of the probability of positive agreement for a given target concentration.
Distribution-balanced sensitivity	The modeled value of the sensitivity at a selected (balanced) cycle distribution distribution different from that of the real-world dataset.
Reference distribution	A distribution of the target concentration that is adopted as reference to model the sensitivity.

[a]qRT-PCR: quantitative reverse transcription polymerase chain reaction.

Results

Raw Positive Agreement Statistics

This section describes the basic performance statistics of the ATs of the 4 suppliers analyzed, and the estimated PPAfs, based on the binary data collected from the Chelsea study [15]. The

agreement matrix was determined for each supplier AT, and the common performance agreement fractions were calculated: sensitivity, specificity, positive prediction, negative prediction, and total prediction. Table 2 displays the basic performance statistics for each one of the test suppliers, including the Clopper-Pearson 95% confidence limits [16].

Table . Basic performance statistics for COVID-19 in vitro diagnostics suppliers A, B, C, and D.

Supplier	Positive agreement cases, n	Total cases, n	Value	Lower 95% confidence limit	Upper 95% confidence limit
A					
Sensitivity	177	211	0.84	0.78	0.89
Specificity	55	57	0.96	0.88	1.00
Positive prediction	177	179	0.99	0.96	1.00
Negative prediction	55	89	0.62	0.51	0.71
Total agreement	232	268	0.87	0.82	0.90
B					
Sensitivity	117	156	0.75	0.67	0.82
Specificity	90	91	0.99	0.94	1.00
Positive prediction	177	118	0.99	0.95	1.00
Negative prediction	90	129	0.70	0.61	0.77
Total agreement	207	247	0.84	0.79	0.88
C					
Sensitivity	55	85	0.65	0.54	0.75
Specificity	144	144	1.00	0.97	1.00
Positive prediction	55	55	1.00	0.94	1.00
Negative prediction	144	174	0.83	0.76	0.88
Total agreement	199	229	0.87	0.82	0.91
D					
Sensitivity	35	43	0.81	0.67	0.92
Specificity	107	114	0.94	0.88	0.97
Positive prediction	35	42	0.83	0.69	0.93
Negative prediction	107	115	0.93	0.85	0.97
Total agreement	142	157	0.90	0.88	0.95

Sensitivities show large differences across the suppliers A, B, C, and D, with values 0.84, 0.75, 0.65, and 0.81, respectively. A comparison plot of the raw AT sensitivities for each supplier and confidence limits is shown (Figure 1). Differences are significant with a large departure of 19% (percentage points of the sensitivity) between suppliers A and C. However, the histograms of qRT-PCR Cts supporting the sensitivity calculations have marked differences across the suppliers (Figure 2). Note that suppliers A and D have a larger proportion of low Cts (large viral sample concentration). On the other hand, supplier C has a larger representation of large Cts (low viral sample concentration).

Figure 1. Raw sensitivities resulting from the real-world Chelsea study for lateral flow in vitro diagnostics suppliers A, B, C, and D. The boxes and whiskers indicate the median and 50% and 95% confidence limits calculated using the Clopper-Pearson statistical method. The horizontal dotted line across all histograms indicates the 0.8 sensitivity value, which is acceptable for in vitro diagnostics clinical performance according to regulatory standards at 0.8 sensitivity for COVID-19 antigen tests.
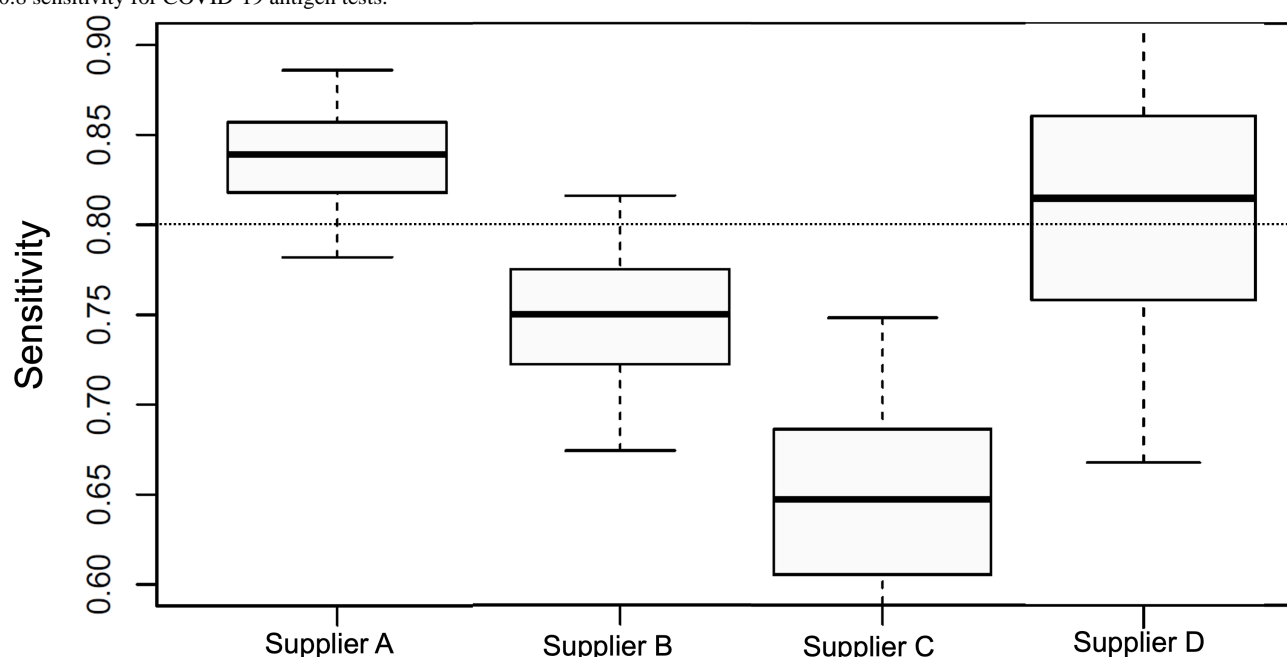
Figure 2. Histogram distributions of cycle thresholds (Cts) for antigen test suppliers A-D using quantitative reverse transcription polymerase chain reaction (qRT-PCR) Cts for each antigen test supplier's dataset. Cts are the average of N and ORFab gene segments of SARS-CoV-2 and reported from Clinical Laboratory Improvement Amendments–certified laboratory using a PerkinElmer SARS-CoV-2 Food and Drug Administration–approved kit.
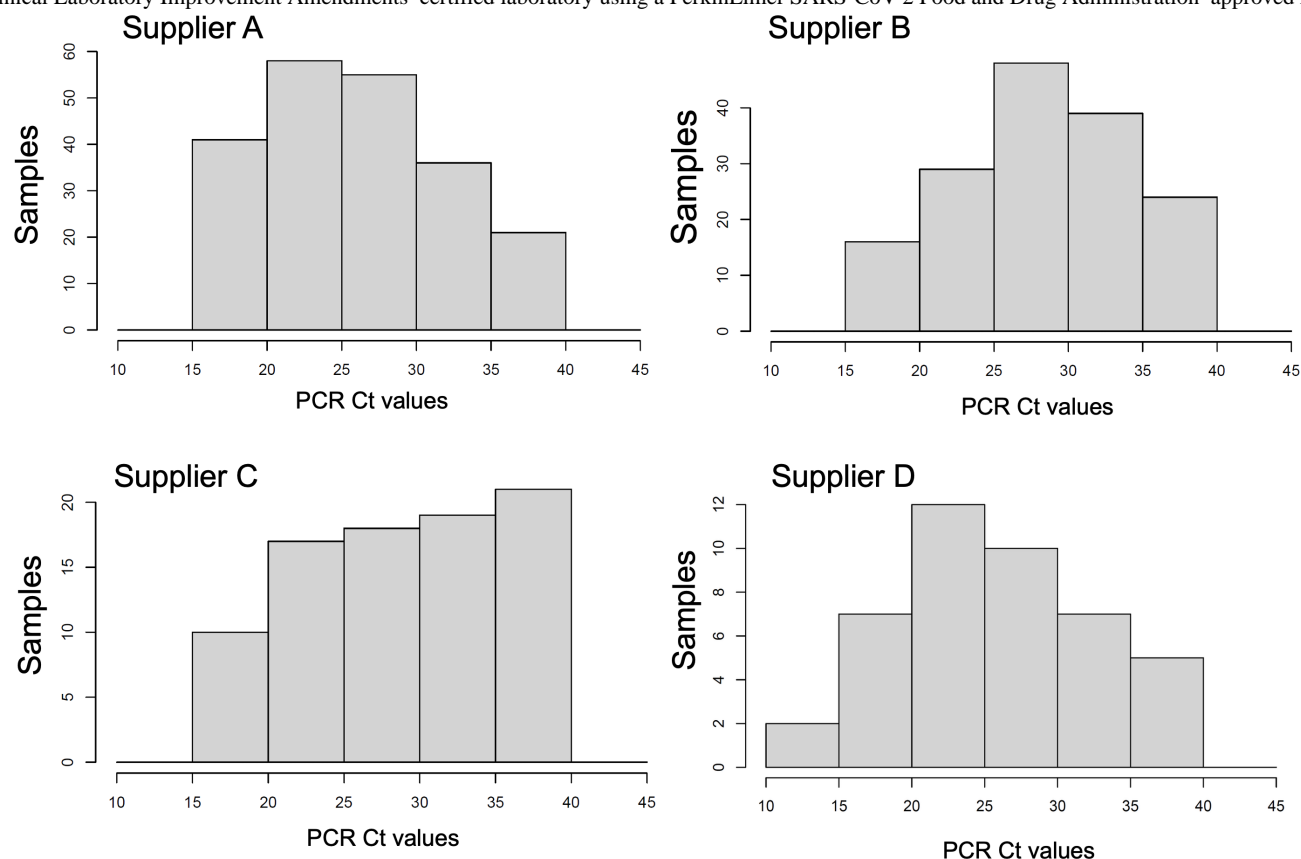
Probability of Positive Agreement Functions

Raw sensitivities (Figure 1) superpose the effects of the viral concentration support to the true performance of the ATs.

Following our method, a first step to decouple the 2 effects is estimating the PPAf from the raw data of each AT supplier by logistic regression, as explained in the previous section. Figure 3 shows the binary data collected for each AT supplier plotted

against the Cts and the estimated PPAf for each test supplier. The estimation of the PPAf by fitting the observed binary AT data also provides the description of the uncertainties in the

PPAf. With the 95% confidence intervals of the PPAf, our formulation estimates the full distribution of the PPA conditioned to the Ct value.

Figure 3. Positive percent agreement (PPA) as a function of the quantitative reverse transcription polymerase chain reaction (qRT-PCR) cycle thresholds (Cts) for naked-eye assessments of the Chelsea project participants after self-application of the antigen tests. Naked-eye assessments of the antigen test result are plotted in the vertical axis with value 1 for positive and 0 for negative. The PPA function is obtained by logistic regression of the binary naked-eye results and shows the strong dependency of the agreement probability with the qRT-PCR Cts.
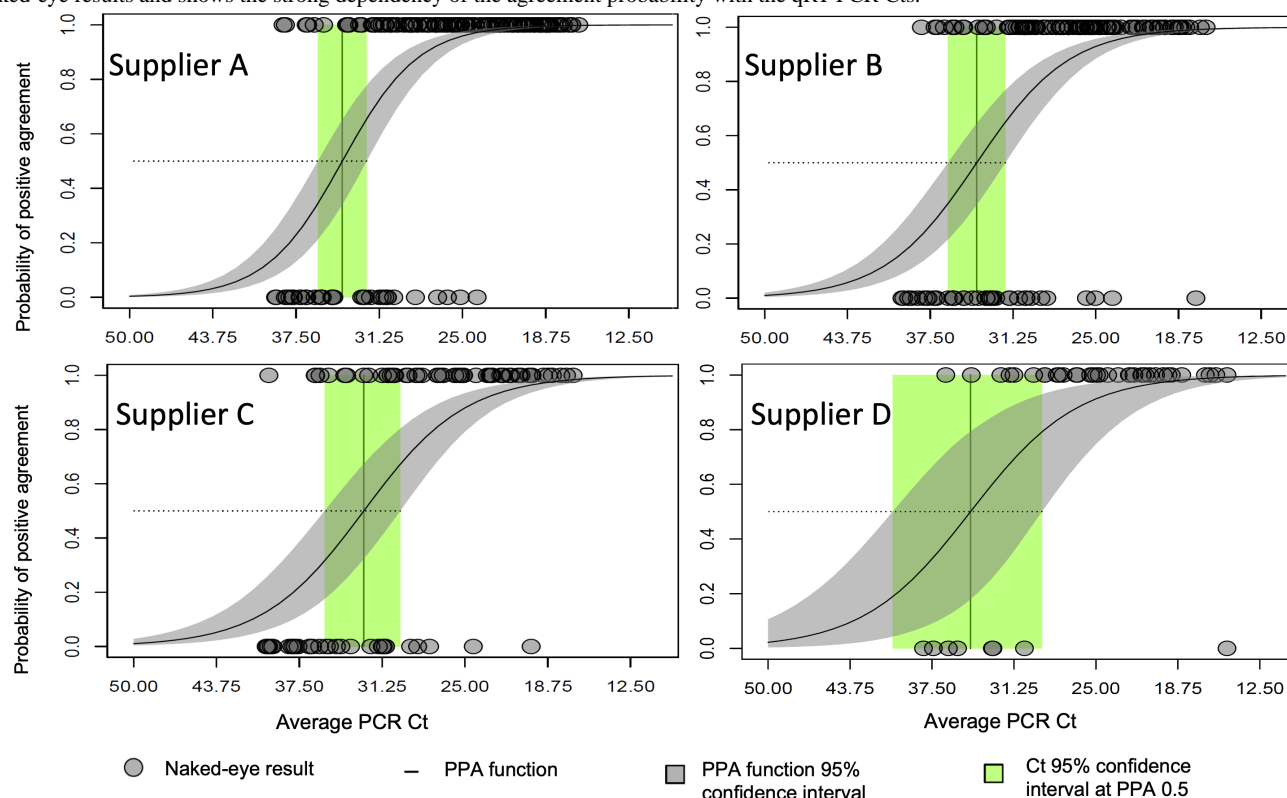
The PPAfs for suppliers A and B are similar (Figure 3). Note that the Ct of median probability (ie, limit of detection at probability *P*=.50) and the slope of the line function are similar to one another. The PPAf for supplier C shows a slightly larger Ct at median probability and lower slope of the function. The PPAf for supplier D is also close to the A and B functions but shows a wider range of uncertainties, also expected from the smaller dataset supporting supplier D.

Several factors influence the estimation of the PPAf: the representation of the Ct range, the balance between positive and negative samples, and the overall sample size. Smaller sample sizes increase the uncertainty of the PPAf, as illustrated in the case of supplier D. However, in this case, the estimation remained reliable because the Ct range and the positive-negative representation were sufficiently well balanced. The Ct values in supplier D's dataset are comparable to those observed in other supplier datasets. Similarly, the distribution of positive and negative samples is appropriate, with a higher frequency of negatives at lower viral loads (ie, higher Ct values), as expected. In contrast, datasets with limited coverage of the Ct range,

underrepresentation of positives or negatives, or a distribution that fails to reflect the expected polarization, with negatives concentrated at higher Ct values and positives at lower Ct values, would not support a reliable estimation.

Reference Distributions of Viral Load

Due to the specific Ct value distributions, the comparison of raw sensitivities is biased by the uneven distribution of the Ct support. This is shown by the corresponding histograms (Figure 2). Utilizing the balance methodology, we computed the sensitivities of the ATs from the 4 suppliers across 4 distinct reference distributions of the Cts. The sample statistics of the PPA exhibit particular sensitivity to low positive samples, that is, those with low viral concentration. Consequently, we opt for a uniform distribution of qRT-PCR cycles spanning 10‐35 Cts, with variable proportions within the 35‐40 range, to underscore the significance of representing low-positive cases in the overall sample PPA (Figure 4A-C). Additionally, we employed the combined Ct distribution of all 4 tests as a reference, that is, the joint positive qRT-PCR Ct counts of the 4 suppliers' data (Figure 4D).

Figure 4. Comparison of selected reference distributions of the raw data. Histograms A, B, and C correspond to 200 dataset points with a range of low virus load (from 5% to 15% of sample at 35‐40 quantitative reverse transcription polymerase chain reaction [qRT-PCR] cycle threshold [Ct] range) and uniform distribution in the range of high-to-moderate virus load (10‐35 qRT-PCR Ct range). Histogram D corresponds to the overlap of the 4 antigen test in vitro diagnostics suppliers' real-world data.
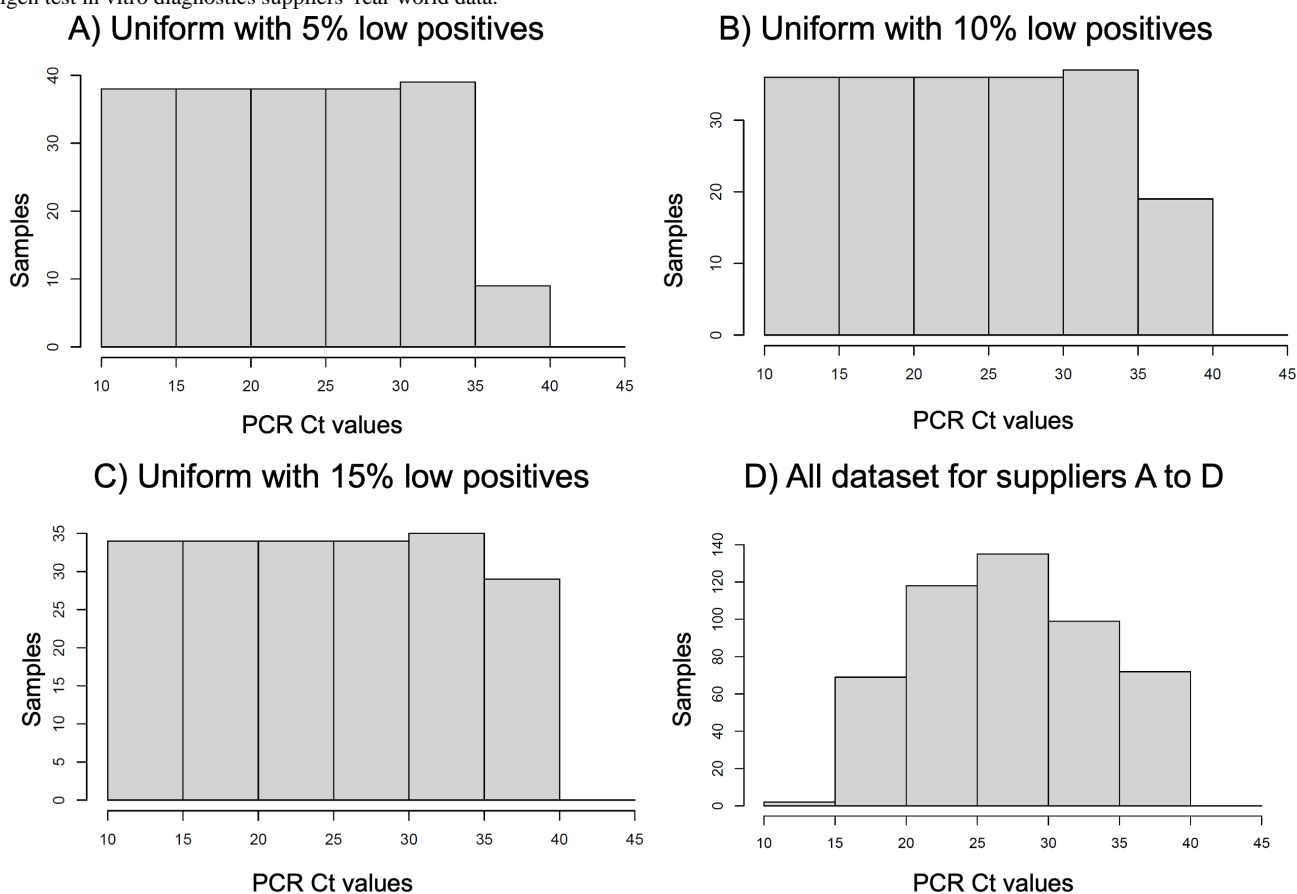
We present 4 different Ct distributions to illustrate how Ct values, and therefore virus load, influence the adjusted sensitivity. The purpose of the distribution-balancing process is 2-fold: (1) to enable accurate comparison of AT sensitivities across studies and (2) to improve real-world sensitivity estimation. First, comparing sensitivities across studies becomes more reliable when distribution-balanced is referenced to a common distribution. Second, clinical studies often have limited numbers of cases and may not adequately reflect the real-world distribution of target concentrations observed in broader populations (eg, regional or national data). In this context, distribution-balanced sensitivity provides a closer approximation of real-world performance. For our clinical dataset, the distribution that best represents the broader population is the overall Ct dataset (Figure 4D), as it includes all cases pooled from the 4 devices.

Sensitivities for the Reference Distributions

According to the described viral load balance method, we estimated the sensitivities of the ATs of the 4 suppliers (Figure 5) over each one of the reference distributions (Figure 4) utilizing the PPAf. It is interesting to compare the results shown (Figure 4) to the raw sensitivities previously calculated (Figure 1). Although the order of performance of the 4 suppliers has been preserved, in order of best to worst performances, the order was supplier A, B, D, and C. The sensitivity differences are smaller among suppliers once the effect of the source distribution is removed. The difference between suppliers A and C is only at 5% instead of the 19% for the raw sensitivity calculation. A large proportion of the raw sensitivity difference between these 2 suppliers originated from the overrepresentation of high viral concentration samples for supplier A and the overrepresentation of low viral concentration samples for supplier C.

Figure 5. Sensitivities of antigen test brands calculated from reference distributions of cycle thresholds as shown in Figure 4. The boxes and whiskers indicate median and 50% and 95% confidence limits. Horizontal dotted line across each panel shows the threshold of 0.8 sensitivity acceptable for in vitro diagnostics clinical performance according to regulatory standards for COVID-19 antigen tests. RW: real world.
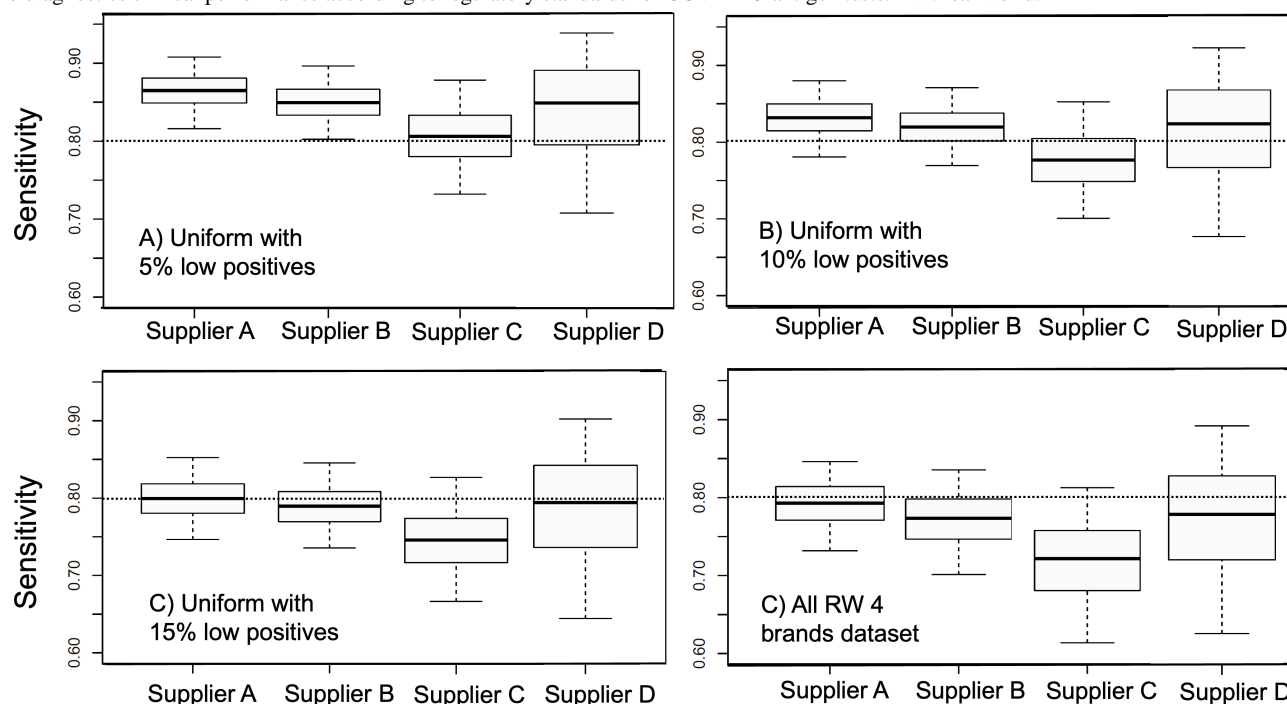
Figure 5 shows the impact of the reference distributions of Ct values on the resulting sensitivity. The absolute value of the sensitivity shows a variation of over 8% difference across the different distributions—larger than the difference across the supplier sensitivities. In particular, the fraction of low virus concentration positives plays an important role, as expected: the PPAfs (Figure 3) show that the probability of positive agreement is very low for low positives (35‐40 Cts) in all suppliers. The indicated line at 0.8 sensitivity in Figure 5 helps to illustrate this point. With the reference distribution including 5% of low positives, the 4 suppliers have sensitivities over the 0.8 threshold. With the reference distribution including 15% of low positives, suppliers B and C are below the threshold, whereas suppliers A and D are borderline at the 0.8 threshold value. With the distribution that combines the 4 suppliers' observed samples, all the suppliers are below the 0.8 threshold.

Discussion

Accurate estimation of AT performance in real-world studies is often confounded by heterogeneous viral load distributions, sample collection conditions, and demographic factors [17]. In particular, the Ct values derived from qRT-PCR, which serve as a surrogate for viral concentration, exhibit considerable variability across study populations. This variability can bias sensitivity—or PPA—estimates when derived directly from unbalanced datasets.

To address the limitation, we introduce a mathematically grounded approach that estimates a PPAf via logistic regression and then recalculates sensitivity over a standardized reference Ct distribution. This method transforms raw sensitivity estimates into a harmonized metric that is independent of the original data's viral load distribution, enhancing comparability across studies and diagnostic platforms. By modeling the entire range

of Ct values, rather than focusing solely on predefined low-positive bins, our approach enables more comprehensive and statistically balanced evaluations. A reference distribution of qRT-PCR Ct values serves as a standardized representation of viral load across a target population, enabling consistent evaluation of diagnostic test sensitivity. Unlike raw distributions derived from individual clinical studies—which are subject to variability in recruitment timing, population demographics, testing strategies, and local epidemiology—a reference distribution is designed to reflect a controlled or representative viral load profile against which diagnostic performance can be compared. The reference Ct distribution may be empirically derived from large, well-characterized datasets collected during peak transmission periods or constructed synthetically based on known viral kinetics in the population. For example, an ideal reference might be a unimodal distribution centered around the Ct range associated with peak transmissibility and highest clinical relevance (eg, Ct 20‐30), or it might reflect the full spectrum of observed viral loads (eg, Ct 10‐40), weighted to mirror realistic clinical case presentations across settings. The purpose of applying such a reference is to enable adjusted sensitivity calculations that are independent of the viral load biases inherent in the original data. This is particularly important when comparing AT performance across different suppliers or studies, where raw sensitivities may differ simply due to the proportion of high- or low-Ct samples in each dataset. In the diagnostic settings, a broader distribution capturing both early and late stages of infection may be more appropriate. Ultimately, the choice of reference distribution must be consistent, allowing for harmonized sensitivity comparisons that reflect diagnostic utility across diverse real-world scenarios.

While regulatory frameworks have moved toward including "lower viral load specimens" in performance assessments to

mitigate overestimation, these approaches often lack generalizability and do not account for the full spectrum of observed Ct values. Our methodology advances this paradigm by providing a distribution-based recalibration mechanism, capturing the full continuum of viral concentrations, and generating sensitivity estimates that are both internally consistent and externally comparable. Critically, FDA Emergency Use Authorization evaluation criteria anchored to qRT-PCR (including use of highly sensitive PCR comparators and mandated proportions of "low-positive" samples) shaped apparent clinical sensitivity and reinforced PCR-first policies for confirmation, which—despite PCR's analytical advantages—dampened uptake of antigen testing for rapid, frequent screening where time-to-result drives transmission control.

Furthermore, genetic strain diversity, specific amino acid mutations in the SARS-CoV-2 nucleoprotein, and host disease–related comorbidities and immunological factors [18,19] may influence AT performance. These factors that could alter antigen detection are also limitations that apply to molecular diagnostics, where ongoing monitoring of performance and vigilance for unexpected results are required. To date, both the original SARS-CoV-2 Delta strain and its descendant Omicron lineages have been detected with comparable efficiency across ATs. From the list of virus lineages we reported, we did not detect differences in AT performance. Kinetics of binding or other more sensitive analysis was outside the scope of this paper. Under the assumption of a relatively genetically uniform circulating viral strain, PCR Ct-based adjustments provided a pragmatic and scientifically valid strategy to reduce bias in estimating AT performance. In contexts where variant heterogeneity becomes relevant, the proposed distribution balancing approach could be extended by calculating variant-specific PPAfs and integrating them into proportionally weighted models. While Ct values serve as quantitative proxies with platform-dependent variability, internal laboratory calibrations (ie, processing of all nasal swabs within the same laboratory and RNA extraction and PCR protocol) mitigate this source of error. In the study, a single RNA extraction and PCR protocol of the Clinical Laboratory Improvement Amendment generated Ct values, thereby minimizing variability and controlling for recalibration effects.

In summary, we propose a robust statistical framework that corrects for real-world sampling biases through distributional modeling. This approach yields adjusted sensitivity estimates that more accurately reflect intrinsic test performance, thereby supporting improved diagnostic evaluation, regulatory decision-making, and public health comparisons across populations and settings.

Data Availability

The data are publicly available at ClinicalTrials.gov (NCT05884515) and upon request at the RADx-UP website for CorDx data. A public access web tool to calculate the percent positive agreement function as well as to execute the distribution-balanced sensitivity methodology is available at [20].

Authors' Contributions

Conceptualization: IB (lead), MB (equal), SH (secondary)
Data curation: AM, DG, DH, IB, JA, MB, RC
Formal analysis: DG (supporting), JA (supporting), LR (supporting), MB (lead), NS (supporting)
Funding acquisition: IB
Investigation: DG, IB, LR, MB, NS, SH
Methodology: IB, MB
Project administration: DH (supporting), IB (lead)
Resources: IB
Supervision: IB, DH, MB, SH

XSL•FO

RenderX

Validation: AM, DG, LR, NS, RC
Writing – original draft: IB (supporting), MB (lead)
Writing – review & editing: IB (lead), MB (supporting)

Conflicts of Interest

IB is a founder of IDX20, a company affiliated with this study. MB is a founder of Info Analytics Innovations, a company also affiliated with this study. The authors declare no additional conflicts of interest.

References

1. Tillmann HL. Hepatitis C virus core antigen testing: role in diagnosis, disease monitoring and treatment. World J Gastroenterol 2014 Jun 14;20(22):6701-6706. [doi: 10.3748/wjg.v20.i22.6701] [Medline: 24944462]
2. Chartrand C, Tremblay N, Renaud C, Papenburg J. Diagnostic accuracy of rapid antigen detection tests for respiratory syncytial virus infection: systematic review and meta-analysis. J Clin Microbiol 2015 Dec;53(12):3738-3749. [doi: 10.1128/JCM.01816-15]
3. Ty Hang V, Minh Nguyet N, The Trung D, et al. Diagnostic accuracy of NS1 ELISA and lateral flow rapid tests for dengue sensitivity, specificity and relationship to viraemia and antibody responses. PLoS Negl Trop Dis 2009;3(1):e360. [doi: 10.1371/journal.pntd.0000360]
4. Lawn SD, Dheda K, Kerkhoff AD, et al. Determine TB-LAM lateral flow urine antigen assay for HIV-associated tuberculosis: recommendations on the design and reporting of clinical studies. BMC Infect Dis 2013 Dec;13(1):407. [doi: 10.1186/1471-2334-13-407]
5. Augustine R, Das S, Hasan A, et al. Rapid antibody-based COVID-19 mass surveillance: relevance, challenges, and prospects in a pandemic and post-pandemic world. J Clin Med 2020 Oct 21;9(10):3372. [doi: 10.3390/jcm9103372] [Medline: 33096742]
6. Peto T, Affron D, Afrough B, et al. COVID-19: rapid antigen detection for SARS-CoV-2 by lateral flow assay: a national systematic evaluation of sensitivity and specificity for mass-testing. EClinicalMedicine 2021 Jun;36:100924. [doi: 10.1016/j.eclinm.2021.100924]
7. Pickering S, Batra R, Merrick B, et al. Comparative performance of SARS-CoV-2 lateral flow antigen tests and association with detection of infectious virus in clinical specimens: a single-centre laboratory evaluation study. Lancet Microbe 2021 Sep;2(9):e461-e471. [doi: 10.1016/S2666-5247(21)00143-9] [Medline: 34226893]
8. Pollock NR, Jacobs JR, Tran K, et al. Performance and implementation evaluation of the Abbott BinaxNOW rapid antigen test in a high-throughput drive-through community testing site in Massachusetts. J Clin Microbiol 2021 Apr 20;59(5):1-10. [doi: 10.1128/JCM.00083-21] [Medline: 33622768]
9. Wagenhäuser I, Knies K, Rauschenberger V, et al. Clinical performance evaluation of SARS-CoV-2 rapid antigen testing in point of care usage in comparison to RT-qPCR. EBioMedicine 2021 Jul;69:103455. [doi: 10.1016/j.ebiom.2021.103455] [Medline: 34186490]
10. Singanayagam A, Patel M, Charlett A, et al. Duration of infectiousness and correlation with RT-PCR cycle threshold values in cases of COVID-19, England, January to May 2020. Euro Surveill 2020 Aug;25(32):2001483. [doi: 10.2807/1560-7917.ES.2020.25.32.2001483] [Medline: 32794447]
11. Rabaan AA, Tirupathi R, Sule AA, et al. Viral dynamics and real-time RT-PCR Ct values correlation with disease severity in COVID-19. Diagnostics (Basel) 2021 Jun 15;11(6):1091. [doi: 10.3390/diagnostics11061091] [Medline: 34203738]
12. Fox MP, Lash TL, Bodnar LM. Common misconceptions about validation studies. Int J Epidemiol 2020 Aug 1;49(4):1392-1396. [doi: 10.1093/ije/dyaa090] [Medline: 32617564]
13. Walker SH, Duncan DB. Estimation of the probability of an event as a function of several independent variables. Biometrika 1967 Jun;54(1):167-179. [doi: 10.2307/2333860] [Medline: 6049533]
14. Corman VM, Haage VC, Bleicker T, et al. Comparison of seven commercial SARS-CoV-2 rapid point-of-care antigen tests: a single-centre laboratory evaluation study. Lancet Microbe 2021 Jul;2(7):e311-e319. [doi: 10.1016/S2666-5247(21)00056-2] [Medline: 33846704]
15. Community study to evaluate the performance of COVID-19 frequent antigen testing at home through digital health reporting. Reagan-Udall Foundation. 2022. URL: https://reaganudall.org/clinical-trial/evaluation-home-use-covid-19-frequent-antigen-testing-and-data-reporting [accessed 2025-10-09]
16. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 1934;26(4):404-413. [doi: 10.1093/biomet/26.4.404]
17. Practical applications of real-world evidence to enable regulatory submissions. Reagan-Udall Foundation. URL: https://reaganudall.org/sites/default/files/2023-09/PIVD%20Commentary%20FINAL3419899_%20bf%20092123.pdf [accessed 2025-10-10]
18. Song Q, Bates B, Shao YR, et al. Risk and outcome of breakthrough COVID-19 infections in vaccinated patients with cancer: real-world evidence from the National COVID Cohort Collaborative. J Clin Oncol 2022 May 1;40(13):1414-1427. [doi: 10.1200/JCO.21.02419] [Medline: 35286152]

XSL•FO
RenderX

19. Del Vecchio C, Cracknell Daniels B, Brancaccio G, et al. Impact of antigen test target failure and testing strategies on the transmission of SARS-CoV-2 variants. Nat Commun 2022 Oct 5;13(1):5870. [doi: 10.1038/s41467-022-33460-0] [Medline: 36198689]
20. IDX20. URL: https://www.idx20.us/ [accessed 2025-10-10]

Abbreviations

AT: antigen test
Ct: cycle threshold
IRB: institutional review board
PDF: probability density function
PPA: percent positive agreement
PPAf: percent positive agreement function
qRT-PCR: quantitative reverse transcription polymerase chain reaction

XSL•FO
RenderX

Systemic Anticancer Therapy Timelines Extraction From Electronic Medical Records Text: Algorithm Development and Validation

Jiarui Yao[1*], PhD; Eli Goldner[1*], MS; Harry Hochheiser[2], PhD; Sean Finan[1], BS; John Levander[2], BS; David Harris[1], BS; Piet C de Groen[3], MD; Elizabeth Buchbinder[4], MD; Danielle Bitterman[5,6], MD; Jeremy L Warner[7,8], MD, MS; Guergana Savova[1], PhD

[1]Computational Health Informatics Program, Boston Children's Hospital, Harvard Medical School, 401 Park Drive, Boston, MA, United States

[2]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, United States

[3]Department of Medicine, Division of Gastroenterology, Hepatology and Nutrition, University of Minnesota, Minneapolis, MN, United States

[4]Department of Medical Oncology, Dana Farber Cancer Institute, Boston, MA, United States

[5]Artificial Intelligence in Medicine (AIM) Program, Mass General Brigham, Harvard Medical School, Boston, MA, United States

[6]Department of Radiation Oncology, Brigham and Women's Hospital/Dana-Farber Cancer Institute, Boston, MA, United States

[7]Center for Clinical Cancer Informatics and Data Science, Legorreta Cancer Center, Brown University, Providence, RI, United States

[8]Brown University Health Cancer Institute, Rhode Island Hospital, Providence, RI, United States

[*]these authors contributed equally

Corresponding Author:
Jiarui Yao, PhD
Computational Health Informatics Program, Boston Children's Hospital, Harvard Medical School, 401 Park Drive, Boston, MA, United States

Abstract

Background: The systemic treatment of cancer typically requires the use of multiple anticancer agents in combination or sequentially. Clinical narrative texts often contain extensive descriptions of the temporal sequencing of systemic anticancer therapy (SACT), setting up an important task that may be amenable to automated extraction of SACT timelines.

Objective: We aimed to explore automatic methods for extracting patient-level SACT timelines from clinical narratives in the electronic medical records (EMRs).

Methods: We used two datasets from two institutions: (1) a colorectal cancer (CRC) dataset including the entire EMR of the 199 patients in the THYME (Temporal Histories of Your Medical Event) dataset and (2) the 2024 ChemoTimelines shared task dataset including 149 patients with ovarian cancer, breast cancer, and melanoma. We explored finetuning smaller language models trained to attend to events and time expressions, and few-shot prompting of large language models (LLMs). Evaluation used the 2024 ChemoTimelines shared task configuration—Subtask1 involving the construction of SACT timelines from manually annotated SACT event and time expression mentions provided as input in addition to the patient's notes and Subtask2 requiring extraction of SACT timelines directly from the patient's notes.

Results: Our task-specific finetuned EntityBERT model achieved 93% F_1-score, outperforming the best results in Subtask1 of the 2024 ChemoTimelines shared task (90%). It ranked second in Subtask2. LLM (LLaMA2, LLaMA3.1, and Mixtral) performance lagged the task-specific finetuned model performance for both the THYME and shared task datasets. On the shared task datasets, the best LLM performance was 77% macro F_1-score, 16% points lower than the task-specific finetuned system (Subtask1).

Conclusions: In this paper, we explored approaches for patient-level timeline extraction through the SACT timeline extraction task. Our results and analysis add to the knowledge of extracting treatment timelines from EMR clinical narratives using language modeling methods.

KEYWORDS

systemic anticancer therapy; electronic medical records; treatment timelines extraction; natural language processing; large language models

XSL•FO
RenderX

Introduction

The systemic treatment of cancer typically requires the use of multiple anticancer agents in combination or sequentially. Systemic anticancer therapy (SACT), which includes traditional cytotoxic chemotherapy, endocrine therapy, targeted therapy, and immunotherapy, has both a low therapeutic index as well as synergistic potential when agents are given in combination. Due to cumulative toxicities, the order in which SACT components are received is much more important than only whether individual drug exposures happened or not, whether in the curative or noncurative setting. Furthermore, patients may receive an extended sequence of treatments across multiple health care settings, systems, and insurance arrangements, making an accurate tally of the totality of treatment using standard structured data resources extremely challenging if not impossible. Meanwhile, clinical narrative texts often contain extensive descriptions of the temporal sequencing of SACT, setting up an important task that may be amenable to automated extraction approaches.

Clinical natural language processing (NLP) is a field that builds computational methods to enable machines to process clinical narratives. Temporality has been a key research area within clinical NLP as it has a wide range of applications including temporal sequencing of SACT [1]. The focus of temporality extraction in clinical NLP has been mainly on instance-level pairwise temporal relation extraction from electronic medical records (EMRs). Instance-level pairwise temporal relations (TLINKs) are the links between an event (EVENT) mention and a temporal expression (TIMEX3) mention or between two event mentions, constituting a triple of the TLINK and the other two components. The set of TLINKs values, that is, type of temporal relations, is CONTAINS, BEFORE, OVERLAP, BEGINS-ON, ENDS-ON, and NOTED-ON [1]. The event that CONTAINS another event is referred to as a narrative container (CONTAINS-1 is the reverse of CONTAINS, meaning an EVENT is contained by the narrative container). In addition, each EVENT has a temporal relation with the document creation time (DocTimeRel), one of BEFORE, BEFORE-OVERLAP, OVERLAP, or AFTER.

The construction of benchmarks, such as THYME (Temporal Histories of Your Medical Event) and i2b2 [1,2], along with the SemEval shared tasks [3-6] on temporality advanced the methodologies and established the state-of-the-art (SOTA) for the task [7-12]. The sophisticated SOTA methods for temporal relation extraction open the door for exploring automatic patient-level timeline construction.

The 2024 ChemoTimelines shared task [13] formulated SACT timeline construction as an information extraction task and provided the deidentified free text documents (except for dates) from the EMRs of 57,520 (breast and ovarian cancer) and 15,946 (melanoma) patients from University of Pittsburgh Medical Center. The documents represented a wide variety of notes, for example, pathology reports, clinical notes, radiology reports, emergency department visits, discharge summaries, etc. A subset of 149 patients was expert-annotated for EVENT mentions, TIMEX3 mentions, and instance-level pairwise temporal relations following the THYME2 schema [1,14] and patient-level timelines of SACT events. The shared task offered 2 subtasks. "Subtask1" involved creating timelines from gold EVENTS and TIMEX3 mentions. "Subtask2" challenged the participants to build end-to-end systems that extracted patient-level SACT timelines directly from the free texts. In this work, "end-to-end" means all text processing is done automatically. Figure 1 summarizes the 2 subtasks. Various approaches were explored by the shared task participants—from supervised finetuning [15,16] to LLM prompting [17,18]. The impressive results (F_1-score=90 for Subtask1 and F_1-score=70 for Subtask2) achieved by the systems from top participants [15] demonstrated the usability and effectiveness of NLP models for this task. The top systems implemented task-specific finetuning of smaller pretrained language models (LMs). Specifically, the LAILab system [15] cast the task as a sequence-to-sequence task, and finetuned Flan-T5-XXL [19] and BART-large [20]. It achieved the best results in the shared task for both subtasks. The Wonder system [16] generated synthetic data using GPT-4 for data augmentation, then finetuned BioLM [21]. The baseline system offered by the organizers [13] also took the supervised finetuning approach with PubMedBERT [22] and secured the second place in both subtasks. In the rest of the paper, for simplicity, we refer to the 2024 ChemoTimelines shared task as the shared task.

In this paper, we further researched SACT timeline extraction using the shared task dataset and adding the dataset of another frequent type of cancer (such as CRC) from another academic medical center. We explored task-specific finetuning approaches and LLM prompting [23-29] to extract SACT timelines. We compared our results on the breast, ovarian, and melanoma datasets from the shared task to the results of the shared task participants. We achieved a new SOTA in Subtask1. We established the SOTA for the CRC dataset as this is a new dataset. Our LLM-based system investigations add to the research of using LLMs for end-to-end SACT treatment timeline extraction from clinical narratives, as only one team explored end-to-end timeline extraction using LLMs in the shared task.

Figure 1. Summary of the 2024 ChemoTimelines shared task. TIMEX3: time expressions; CONTAINS-1: reverse of CONTAINS, meaning "chemotherapy" is contained by "last Thursday"; DocTime: document creation time.
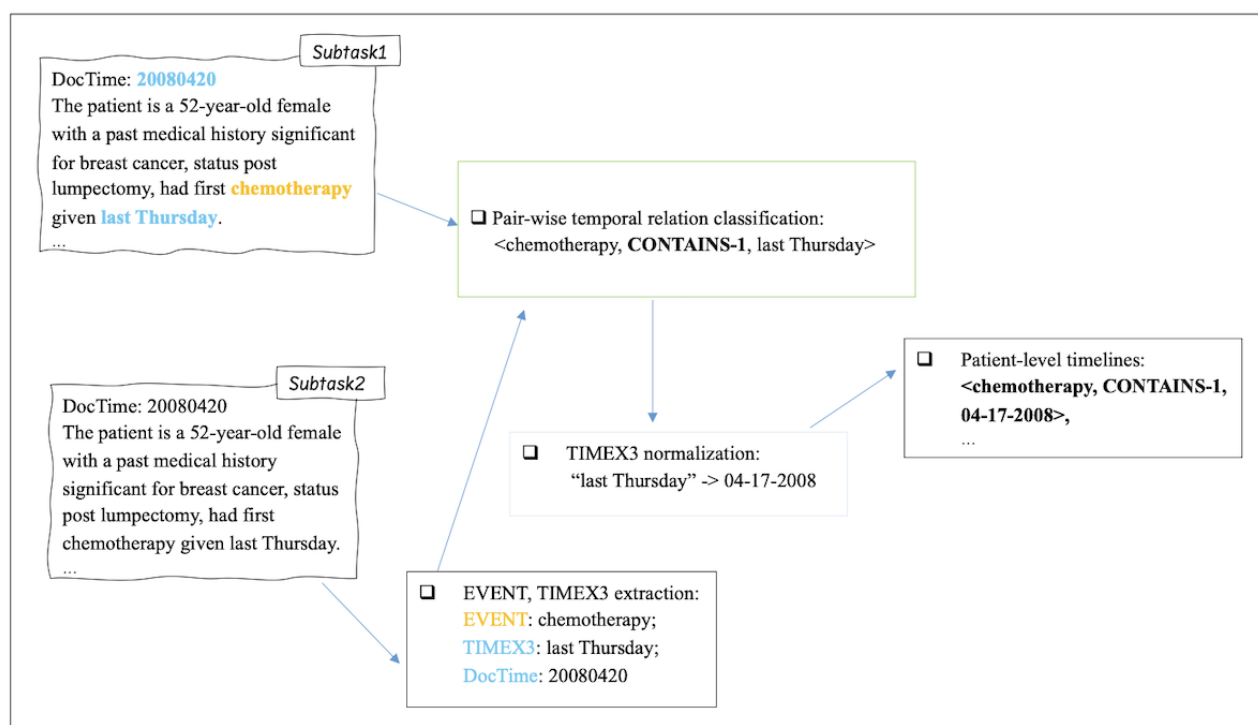
The contributions of this paper are as follows.

First, the approaches for patient-level timeline extraction through the task of SACT timeline extraction. We perform experiments on the 2024 ChemoTimelines shared task as well as on the THYME CRC patients. Our results and analysis on this task add to the knowledge of extracting treatment timelines from EMRs using LLM-based methods.

Second, the SOTA performance of our finetuned LM-based system for Subtask1 of the 2024 ChemoTimelines shared task.

Third, SOTA performance with LLM prompting approaches for Subtask1 and Subtask2 of the 2024 ChemoTimelines shared task outperformed the shared task participant systems that took the approach of prompting LLMs.

Methods

Ethical Considerations

All electronic health record (EHR) data used in this study are deidentified in accordance with the datasets' relevant privacy regulations [1,13,14]. We strictly adhered to the terms outlined in the data use agreement, ensuring that no data were transmitted to any external or public APIs. Ethics approval was not required because the study used secondary data that was aggregated and anonymized before analysis. All experiments were conducted on a secure local machine operating behind a firewall, maintaining full data confidentiality and integrity throughout the study.

Tasks and Datasets

The first dataset we used was from the shared task [13]. The EMR notes of 149 patients with breast, ovarian, and melanoma cancers from the University of Pittsburgh Medical Center were expert-annotated by the shared task organizers for instance-level pairwise temporal relations following the THYME2 schema [1,14] and SACT patient-level timelines.

The second dataset we used included the THYME patients—199 CRC patients from Mayo Clinic. This dataset was NOT part of the 2024 ChemoTimelines shared task. Note that the original THYME corpus consisted of one radiology, one pathology, and one oncology note for each of the 199 CRC patients—not sufficient to extract SACT timelines. Therefore, for the work described in this paper, we obtained the entire EMR documentation for these 199 CRC patients (all manually deidentified except for dates). As with the shared task patients, the CRC patient EMRs were represented by a wide variety of document types. Following the shared task protocol, the CRC notes were expert-annotated for instance-level pairwise temporal relations following the THYME2 schema and SACT patient-level timelines. Table 1 shows the dataset distributions. Table S1 in Multimedia Appendix 1 provides the pairwise label distributions. The label set for the pairwise relations is CONTAINS, BEGINS-ON, ENDS-ON, OVERLAP, and BEFORE. In the final SACT timeline, we converted CONTAINS to CONTAINS-1 so that all triples are structured as <EVENT, TLINK, TIMEX3>, where CONTAINS-1 semantically indicates that the drug was administered on the date specified by the temporal expression (TIMEX3). Note that we did not use i2b2 2012 because we focused on cancer treatment timeline extraction only in this work. Textbox 1 presents a concrete example of patient-level SACT timelines.

As is the established convention, in this paper, we refer to the labels in the shared task and THYME datasets as "gold." All datasets come with predefined training (train), development (dev), and test splits that we used accordingly. Note that the

gold labels of the shared task test set were not publicly available; however, participants could submit their system predictions to the shared task organizers to get evaluation results, thus providing independent evaluation over a held-out, eyes-off dataset.

Table . Dataset summary.

Splits	Patients	Notes	Words[a]	EVENT mentions	TIMEX3[b] mentions	TLINKs[c]
Ovarian cancer (from 2024 ChemoTimelines shared task)						
Train	26	1675	1,183,632	1168	597	494
Dev[d]	8	562	308,814	790	312	226
Test	8	559	257,116	664	381	Not released[e]
Breast cancer (from 2024 ChemoTimelines shared task)						
Train	33	1002	465,644	1023	576	455
Dev	16	499	225,588	279	146	113
Test	35	1333	786,896	2560	1118	Not released
Melanoma (from 2024 ChemoTimelines shared task)						
Train	10	233	124,924	147	78	48
Dev	3	211	178,308	789	261	201
Test	10	229	156,083	398	193	Not released
Colorectal cancer (CRC)						
Train	98	12,990	6,038,431	11,161	6155	5897
Dev	50	6810	3,105,675	3964	2194	1924
Test	51	7357	3,587,387	7552	3612	4403

[a]"Words" denotes tokens delimited by white spaces.

[b]TIMEX3: time expressions.

[c]TLINKs: pairwise temporal relations.

[d]Dev: development set.

[e]Note that the number of test set TLINKs for the 2024 ChemoTimelines shared task was not released publicly.

Textbox 1. An example of a summarized patient-level SACT timeline extracted from the entire patient's EMR chart.

- ['chemotherapy', 'contains-1', '2013-06-20']
- ['carboplatin', 'contains-1', '2013-10-24']
- ['carboplatin', 'contains-1', '2013-09-19']
- ['carboplatin', 'contains-1', '2013-07-18']
- ['carboplatin', 'contains-1', '2013-08-08']
- ['carboplatin', 'contains-1', '2013-08-29']
- ['taxol', 'contains-1', '2013-10-24']
- ['taxol', 'contains-1', '2013-09-19']
- ['taxol', 'contains-1', '2013-07-18']
- ['taxol', 'contains-1', '2013-08-08']
- ['taxol', 'contains-1', '2013-08-29']

Approaches

We explored 2 approaches for the task of SACT timelines extraction: (1) finetuning smaller LMs and (2) prompting LLMs.

Figure 2 shows the complete pipeline of both approaches. We describe each approach in detail in this section.

Figure 2. Methods summary. On the left-hand side, temporal relations are classified via a small finetuned language model (FT LM). On the right-hand side, temporal relation triplets are extracted by prompting large language models (LLMs). In both approaches, EVENTS are extracted using a **B**egin-**I**nside-**O**utside (BIO) tagger. Output for both systems is the same, see Textbox 1. cTAKES: Apache Clinical Text Analysis and Knowledge Extraction System; TIMEX3: time expressions; TLINK: pairwise temporal relation.
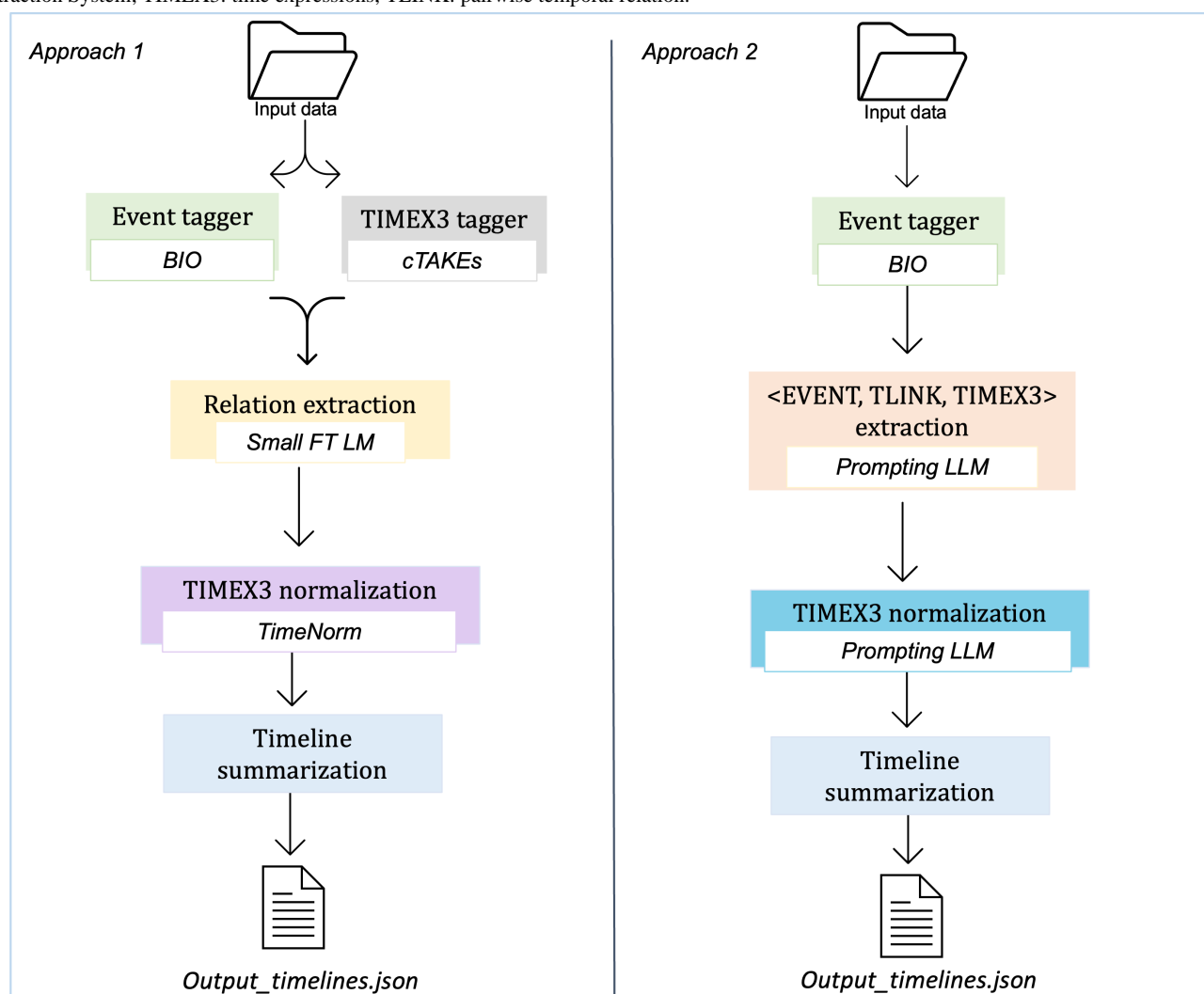
Approach 1: Finetuning LMs for Temporal Relation Extraction

Overview

In this approach, we cast the task of SACT timeline extraction as a pairwise temporal relation extraction task followed by a temporal relation summarization step. Given input texts, we designed a pipeline with the following steps: (1) extracting SACT EVENT mentions, (2) extracting TIMEX3 mentions, (3) classifying pairwise EVENT-TIMEX3 temporal relations, (4) normalizing TIMEX3 mentions, and (5) summarizing and refining patient-level timelines.

Extracting SACT EVENT Mentions

We trained a sequence labeling tagger that marks the beginning, inside, and outside (BIO) of a SACT treatment EVENT mention in the text. The tagger was trained on the train split of the gold labeled data by finetuning a pretrained LM [22,30]. The "Experimental Settings" section shows more details.

Extracting TIMEX3 Mentions

TIMEX3 mentions were extracted by the temporal module of the Apache Clinical Text Analysis and Knowledge Extraction System (cTAKES) [31], a publicly available text processing system. The precision, recall, and F_1-scores of cTAKES for TIMEX3 mention extraction are 57.17%, 83.95%, and 67.25%, respectively; evaluated on the original THYME dataset described in the "Tasks and Datasets" subsection. Different methodologies were used for SACT EVENT mention extraction and TIMEX3 mention extraction because there was no publicly available SACT EVENT extractor with solid performance at the time of the experiments.

Classifying Pairwise EVENT-TIMEX3 Temporal Relations

Given an EVENT-TIMEX3 pair, the task is to determine the temporal relation between them according to a predefined label set of TLINKs (described in the "Introduction" and "Tasks and Datasets" sections). For example, if the patient started a regimen of Taxol on August 1, 2012, the relation between "Taxol" and "August 1, 2012" is BEGINS-ON. Inspired by previous works [11], we finetuned EntityBERT for this step to create an LM specifically trained to attend to EVENT and TIMEX3 mentions.

The input to the model was the EVENT and TIMEX3 mentions within a context window with the EVENT and TIMEX3 mentions highlighted by special tokens, possibly crossing sentence boundaries. We followed the same data preprocessing format as described in [7,9,11]. Concretely, EVENT and TIMEX mentions are highlighted by XML tags "<e>," "</e>," "<t>," and "</t>." The context window that defines the token distances between an EVENT and TIMEX3 in an EVENT-TIMEX3 pair is set to 60 tokens, empirically derived to cover over 95% of the EVENT-TIMEX3 pair instances. The model was trained on the train split of the gold-labeled data for multiclass classification.

Normalizing TIMEX3 Mentions

The goal of this step is to map TIMEX3 mentions to a computable format. We used TimeNorm [32,33] to normalize the TIMEX3 mentions and the document creation time (DocTime) to ISO-TimeML standard [34] (eg, "yesterday" in a note with a DocTime of "2022-04-29" would be normalized to "2022-04-28").

Summarizing and Refining Patient-Level Timelines

A patient's SACT history can be mentioned in multiple notes in different contexts. For example, the physician may discuss the termination of one treatment due to side effects; despite that, in another note, they may say that the therapy will be given to the patient for 3 more cycles. Therefore, after the instance-level temporal relation extraction, deduplication and conflict resolution are necessary to get the final patient-level SACT timelines. For this step, we followed the heuristics from the shared task [13].

Approach 2: Prompting LLMs for SACT Timeline Extraction

Overview

We developed an end-to-end timeline extraction pipeline via LLM prompting. This pipeline involved two steps: Step 1 focused on extracting <EVENT, TLINK, TIMEX3> triplets from clinical texts, and Step 2 was designed for TIMEX3 normalization. We took the approach of in-context learning, which refers to the method of adding exemplars of gold examples with answers to the prompt [25], a common practice in prompt engineering. Textbox 2 provides the prompt templates we used in both steps. For Step 1, we provide 4 exemplars for each TLINK label. For Step 2, we provide 5 exemplars in total. The exemplars are selected from the training split of the data. We explored the discrete prompting strategy where the prompts are created manually, ultimately settling on the prompts with the best performance.

Textbox 2. Prompt templates used in our large language model (LLM) experiments. For Step 1, we provide 4 exemplars for each label. For Step 2, we provide 5 exemplars in total.

- Step 1 prompt: You are a helpful assistant for oncologists. You will read the given PATIENT EHR and summarize the patient's chemotherapy treatment TIMELINES. Please only output TIMELINES in the requested format. Please do not include any other text or reasoning, do not include timelines for any other treatments besides chemotherapy. Please do not use any labels other than the ones given in the examples, i.e., BEGINS-ON, ENDS-ON, CONTAINS. Here are some examples.

- Step 2 prompt: You are asked to decide the date of a time expression. If today was 2013-05-02, what would the date of yesterday be? Please only output the date in the format of "YYYY-MM-DD". Answer "Unknown" if you don't know. Here are some examples.

Step 1: Extracting < EVENT, TLINK, TIMEX3> Triplets

The construction of patient-level treatment timelines requires the system to process all notes of a patient, thus the input can exceed the LLM context window. Current open LLMs have a limited number of tokens they can process per time, for example, LLaMA1 [35] supports up to 2048 tokens and LLaMA2 [23] supports up to 4096 tokens; however, even if the LLM could ingest all the notes of one patient as input per time, which would not be an efficient way of processing texts as transformers' self-attention scales quadratically with input length. Therefore, sending all the notes of a patient to LLMs at one time is not practical. To make this task more feasible for LLMs, we prompted the LLM with only relevant snippets of notes and assembled the timelines afterwards. Specifically, we extracted SACT EVENT mentions using the BIO tagger trained in Approach 1, then fed the LLM the sentences containing the SACT EVENT mentions to extract the triplets. Note, the input to the LLM was a sentence, unlike the context window instances fed to the pairwise classifier in Approach 1. In our initial experiments, we used context window instances with the LLMs as well; however, the partial sentences confused them as tokens outside of the window are discarded. To give LLMs a self-contained input with a reasonable sequence length, we decided to give a complete sentence as input for LLMs instead of a context window as we did in Approach 1.

Step 2: TIMEX3 Normalization With LLMs

We applied in-context learning to normalize the TIMEX3 mentions. For each output triplet from Step 1, we prompted the model to normalize the date of the TIMEX3 mention given the DocTime of the note. We then assembled the final timelines, using the same heuristics as in Approach 1.

Experimental Settings

We explored two approaches for the task of SACT timelines extraction: (1) finetuning smaller LMs and (2) prompting LLMs. For the first approach, we finetuned PubMedBERT base model [22] to train the SACT event tagger. For the temporal relation classification task, we finetuned EntityBERT based on the results reported by Lin et al [11], where they finetuned BioBERT, PubMedBERT, and EntityBERT for clinical temporal relation classification and found that EntityBERT outperformed the other two models. For the experiments with LLMs, we chose LLaMA2-70B [23], LLaMA3.1-70B [36], and Mixtral-8×7B-Instruct-v1 [24], which are current SOTA open LLMs. We did not use proprietary LLMs such as GPT4 [26] because we did not have access to their Health Insurance

Portability and Accountability Act (HIPAA)-compliant versions. The open models we experimented with are reported to have yielded results competitive with those of the proprietary models [24]. Furthermore, we compare our results with those systems in the shared task for the types of cancers included in the shared task. For the CRC dataset (not included in the shared task), we establish the first result that will serve as the baseline for the community. See Table S2 in Multimedia Appendix 1 for details on the computational settings.

We experimented with prompting LLMs for both Subtask1 and Subtask2. In Subtask1, we provided explicit gold SACT events and time expressions in the text, then prompted the LLM to predict the temporal relation between them. The prompt template for this subtask is shown in Table S3 in Multimedia Appendix 1. In Subtask2, we passed to the LLM only plain text as input, then asked the LLM to extract the SACT events, time expressions, and temporal relation between them in 1 step. Textbox 2 lists the prompt template for Subtask2.

Evaluation and Baseline

We used the evaluation metric provided by the shared task, which computed the average F_1-scores across all patients. There were 4 settings with different temporal granularities: strict, relaxed-to-day, relaxed-to-month, and relaxed-to-year. For example, the relaxed-to-month setting required the model to correctly predict the year and month when the therapy was given, while the strict setting required the model to capture the exact date when the patient received the therapy. The official metric for the 2024 shared task was relaxed-to-month scores, which we used as our metric to report the main results in this paper. Results using other metrics are given in Table S4 in Multimedia Appendix 1.

As a baseline, we used the baseline system used in the shared task, which implemented a predefined dictionary as a lookup table for SACT EVENT extraction and a finetuned LM for temporal relation classification. We also compared our results on the 3 types of cancer (breast cancer, ovarian cancer, and melanoma) to the shared task leaderboard results.

Results

In Table 2, we present our results on the development (Dev) and test sets. As the CRC dataset was not available for the shared task, we also present the results of our model finetuned only on the shared task data (under EntityBERT 3 Cr) for a direct comparison with other participating systems. That is, using Approach 1 described above, we trained 2 versions of the model. "EntityBERT" was trained on the shared task data and CRC data. "EntityBERT 3 Cr" was trained only on the shared task data (we combined the training datasets of multiple cancer types into 1 training dataset to train the EntityBERT 3 Cr model and EntityBERT model). Subtask1 in Table 2 shows the results with gold SACT EVENT and TIMEX3 mentions as input. In general, the finetuned EntityBERT and EntityBERT (3 Cr) outperformed LLaMA2, LLaMA3.1, and Mixtral LLMs by a large margin. Among the LLMs, LLaMA achieved higher scores than Mixtral. In Table 2, Subtask2 shows the end-to-end evaluation results. The SACT event extraction evaluation results using the BIO tagger can be found in Table S5 in Multimedia Appendix 1. We note a wide gap between the performance with gold mention input (Subtask1) and the performance with automatically extracted mentions (Subtask2), suggesting that the errors in the mention extraction stage propagate to the relation extraction stage and dramatically affect the overall accuracy of the system. We also notice that the smaller finetuned models outperform LLMs in most cases except for melanoma, the reasons for which we discuss in the Discussion section.

Table . Evaluation results of our systems across 4 types of cancers from 2 academic centers. Scores are macro F_1-score, relaxed-to-month.

Cancer type and models		Subtask1[a], %		Subtask2[b], %	
		Development set	Test set	Development set	Test set
Ovarian cancer					
	EntityBERT[c]	93[e]	95[e]	64	61
	EntityBERT (3 Cr)[c, d]	93[e]	94	67[e]	69[e]
	LLaMA2[f]	70	70	29	42
	LLaMA3.1[g]	75	74	31	56
	Mixtral[h]	60	67	7	27
Breast cancer					
	EntityBERT[c]	97[e]	97	88[e]	63
	EntityBERT (3 Cr)[c]	97[e]	98[e]	87	66[e]
	LLaMA2	81	83	61	50
	LLaMA3.1	79	70	66	48
	Mixtral	66	63	37	25
Melanoma					
	EntityBERT[c]	86[e]	91[e]	43	39
	EntityBERT (3 Cr)[c]	86[e]	88	46	40
	LLaMA2	80	79	47[e]	47[e]
	LLaMA3.1	67	71	26	38
	Mixtral	65	65	4	25
Colorectal cancer (CRC)					
	EntityBERT[c]	90[e]	83[e]	58[e]	56[e]
	LLaMA2	66	77	40	32
	LLaMA3.1	66	68	45	38
	Mixtral	58	66	19	15

[a]Subtask1: input is gold entities (systemic anticancer therapy [SACT] events and time expressions).

[b]Subtask2: entities are automatically generated by the system.

[c]These are systems using small finetuned models.

[d]EntityBERT (3 Cr): EntityBERT model trained only on the shared task data.

[e]These are the best results.

[f]LLaMA2-70B.

[g]LLaMA3.1-70B.

[h]Mixtral-8×7B-Instruct-v1.

Furthermore, unlike the LLM prompting approaches, both our systems based on the smaller finetuned models can be deployed for inference on a laptop without a GPU. Our Subtask1 system is able to process approximately 14 notes/minute. Our Subtask2 system is able to process approximately 10 notes/minute. Assuming a typical patient with 200 notes, our Subtask1 system takes on average 14.5 minutes to process all of the patient's notes, and our Subtask2 system takes on average 20 minutes to process all of the patient's notes. On the other hand, the LLM prompting experiments were conducted on NVIDIA A100 GPUs. It took the LLaMA3.1 70B model approximately 28 minutes for Subtask1 and 13 minutes for Subtask2 to process

200 notes. It took LLMs less time to complete Subtask2 because only sentences containing TIMEX3 mentions needed to be processed in Subtask2.

We position our systems within the broader context of the 2024 ChemoTimelines shared task by comparing them with the shared task participants' systems. If 1 shared task participant has multiple submissions, we take their best result for comparison. Note the official metric for the leader board is relaxed-to-month scores on the Test set. We first compare the result of our EntityBERT (3 Cr) model with the results of the participating systems using similar approaches, that is, finetuning smaller

LMs [13,15,16,18,37]. In Figure 3-Part A, we can see that in Subtask1 our model achieved the best results overall and on the individual cancer types. Our Subtask1 result was 3 points higher than the best shared task score achieved by LAILab [15] (93% vs 90%). In Subtask2 (Figure 3-Part B), our system had the second-best overall scores. However, it is worth noting that LAILab finetuned Flan-T5-XXL [19], a model with 11B parameters, which was much bigger than the EntityBERT model we used that had about 100 million parameters.

Finally, we observe in Table 2 that the model trained only on the breast, ovarian, and melanoma data from the train split of the shared task (ie, EntityBERT 3 Cr) performed slightly better than its counterpart trained on the full train split containing all 4 types of cancer (ie, EntityBERT) in Subtask2. We conjecture that since there was more data for CRC than the other types of cancer within our dataset, the representation of the signal from the CRC data overwhelmed that of the other three cancer types inside the model. The addition of the second dataset (CRC) in this work aims to create a larger pool of datapoints adding a new type of cancer and a different institution as the data source. It also helps answer the questions of whether (1) a model built off data across different EMR sources might be feasible and (2) the quantity of the data matters. Our experiments on these two datasets show that (1) it is likely that institution-specific models capture treatment patterns better but not by a large margin and (2) patterns of the data-rich source likely dominate.

In Figure 4 we compare our LLM-based approaches with the shared task systems that prompted LLMs. With gold mentions as input (Subtask1), our system based on prompting LLaMA2 achieved the highest overall score compared to the shared task systems. When using Mixtral as the starting point, our system and the NLPeers [18] system achieved similar overall scores (65% vs 64%), which are significantly lower than the overall score of LLaMA2 and LLaMA3.1, suggesting that LLaMA family models are more suitable for this subtask than Mixtral. Only 1 team from the shared task explored end-to-end timeline construction using an LLM. In Figure 4-Part B, Subtask2 we can see that the overall performance of the two Mixtral-based systems is similar. Again, we see a performance discrepancy between LLaMA and Mixtral. Jiang et al [24] show that Mixtral performed better than or comparable to LLaMA2 across multiple benchmarks. Our results suggest that the decision of choosing the right LLM should be made empirically. Note that the two LLaMA models we used have the same number of parameters, 70B. Compared to LLaMA2, LLaMA3.1 improved the results on the ovarian dataset, but fell short on the breast and melanoma datasets. Across 64 evaluation settings (4 cancer types, 4 metrics, 2 subtasks, both development and test sets), LLaMA3.1 achieved higher or same F_1-scores as LLaMA2 in 39 cases (61%; see Table S4 in Multimedia Appendix 1). Overall, we observe similar trends across strict, relaxed-to-day, relaxed-to-year evaluation settings as relaxed-to-month setting.

Figure 3. Comparison to finetuning-based models in the shared task [13,15,16,18,37]. Scores are relaxed-to-month macro F_1-score on the test set. "Our EntityBERT, 3 cr" refers to the EntityBERT model trained only on the shared task data. The best-performing team in the shared task was LAILab [15].

Figure 4. Comparison to LLM prompting systems in the shared task [13,17]. Scores are relaxed-to-month macro F_1-scores on test set. "Our LLaMA2" and "Our LLaMA3.1" are LLaMA2-70B and LLaMA3.1-70B, respectively. "Our Mixtral" is the Mixtral-8 × 7B-Instruct-v1 model. FS and ZS refer to few-shot and zero-shot settings.

A: Gold Entities (Subtask1)
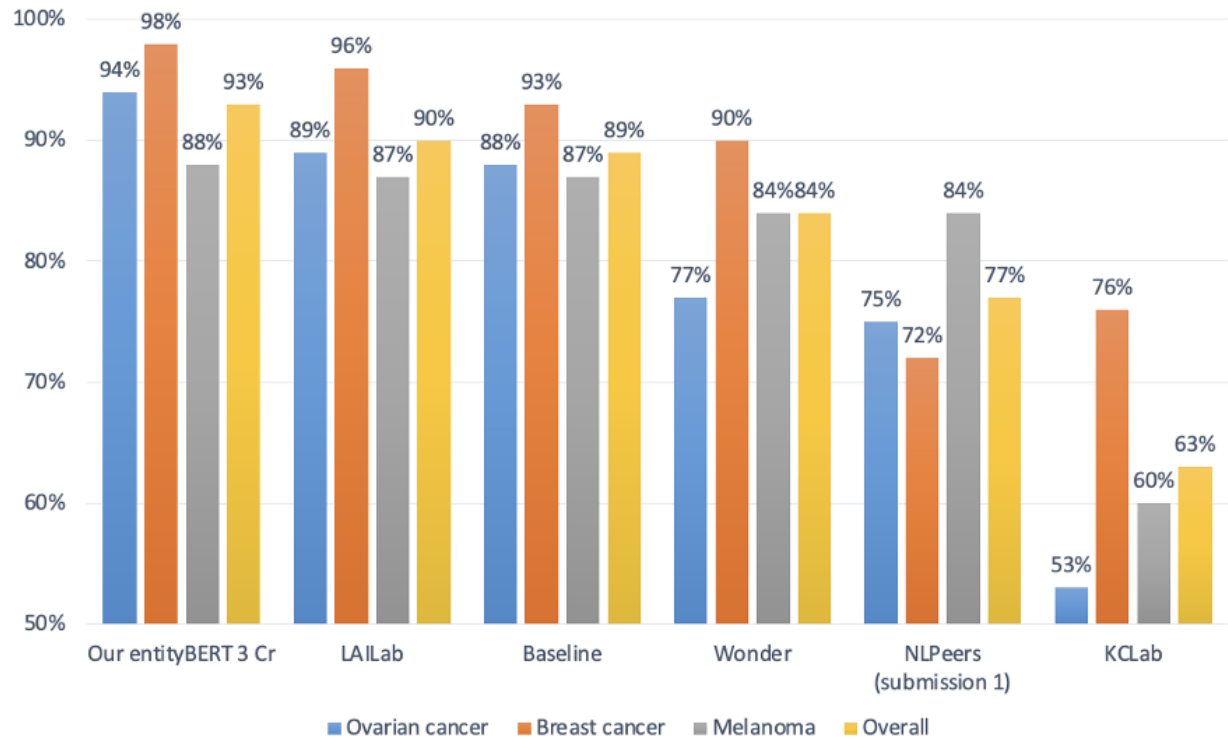
B: Auto entities (Subtask2)
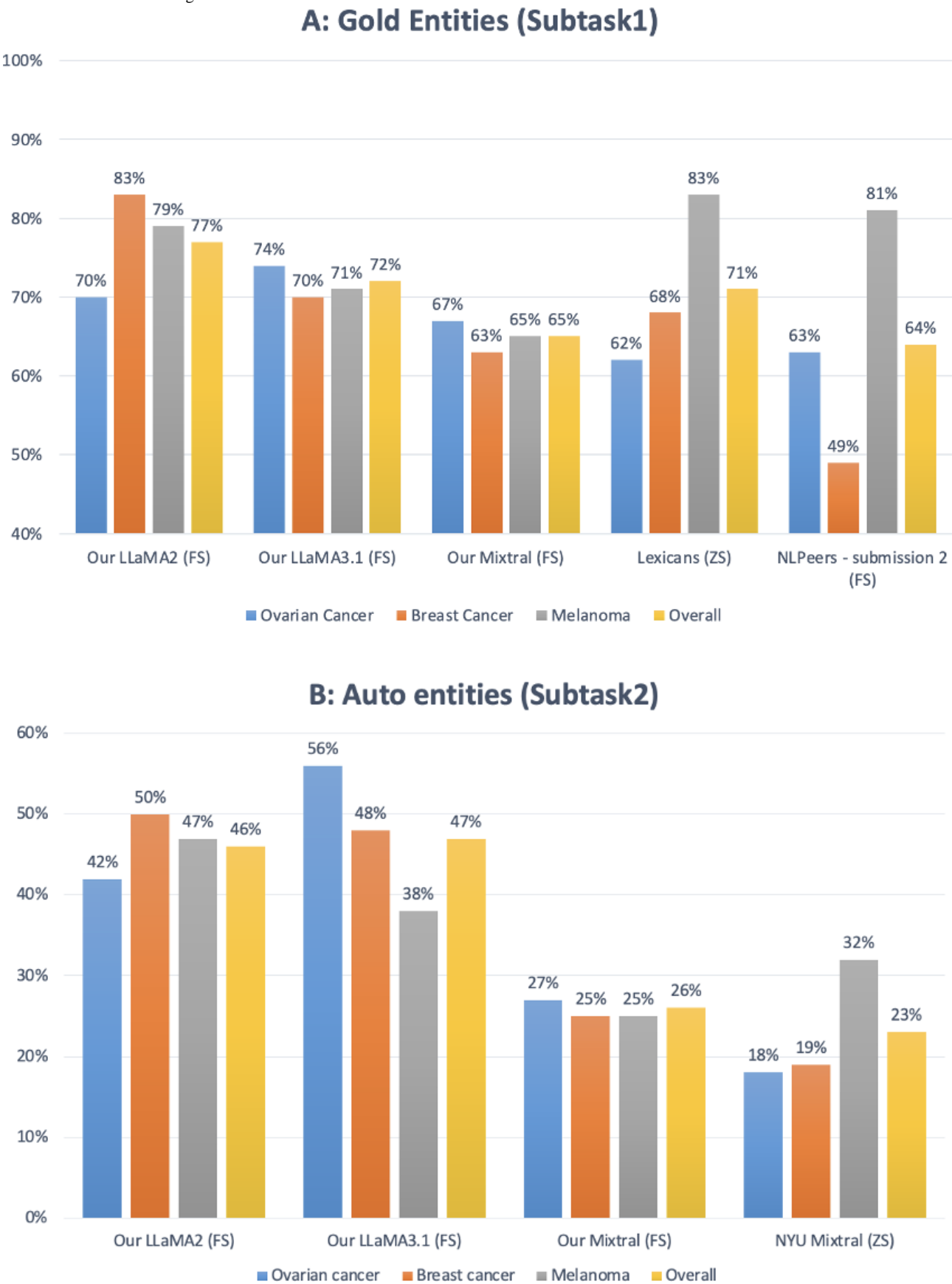
We performed error analysis on the relaxed-to-month output for each cancer type cohort. An incorrect prediction within a predicted patient timeline against a gold patient timeline is either a false positive, that is, a predicted triplet that is not present in

the gold timeline, or a false negative, that is, a triplet in the gold timeline, which is not in the predicted timeline. There is also the possibility of an apparent false positive or false negative being actually correct due to an annotation error, for which we also review. We analyze which of the components in the system pipeline or the annotation process is the root cause of an error in the predicted or gold timelines. For the predicted timeline, this can consist of any combination of one of the extraction components for SACT EVENT mentions (SACT Detection Error) and temporal expression mention (TIMEX3 Detection Error), the TLINK classifier (TLINK Error), and summarization error (Total incorrect summarized predictions). For the gold timeline, this can only consist of an annotation error.

We present the breakdown per error type from the test set in Table S6 in Multimedia Appendix 1. We randomly sampled each type of false positive errors to collect a sample size using a 95% CI, a margin of error of 5%, and a population proportion of 50%. We analyzed the instance-level false positives since

each was associated with a specific TLINK classification instance. The incorrect unsummarized predictions are inputs to the summarization algorithm which result in the incorrect summarized predictions. We found that most of the errors came from incorrect TLINK classification, followed by annotation errors, and finally detection of SACT EVENT and TIMEX3 mentions. We identified the annotation errors for the most part as resulting from likely missed screening of some notes by the expert annotators, as this is a highly cognitively demanding task for a human to perform (see Table 3 for examples). The false negatives tended to be the result of formatting issues, complex reasoning, and some level of hedging around the event. We found that in many notes, there are subsections that start with dates, which are used as the headings for these subsections (see examples in "False negative: formatting" in Table 3); then all events described in that subsection are related to these dates. This is especially challenging as the subsections could consist of multiple sentences.

Table . Types of errors and examples. Note that the dates have been intentionally altered for the purpose of this paper.

Type of error	Text	Explanation
Annotation error	• Anastrozole (Arimidex) 1 mg once a day by mouth [Order Comment : can take am]. • Last dose : 10/18/2033.	• No gold TLINK[a] for "anastrozole (Arimidex)" and "10/18/2033".
Annotation error	• Dr Person17, later today, to discuss management from the standpoint of chemotherapy or hormonal.	• No gold TLINK for "later today" and "chemotherapy".
Annotation error	• Chemo and radiation in 2055.	• No gold link for "chemo" and "2055".
False negative: formatting	• July through December 2055: Completed his 12 cycles of FOLFOX. The first 8 cycles included oxaliplatin and the last 4 cycles were 5-FU/leucovorin.	• No prediction TLINK for "December 2055" and "5-FU/leucovorin". • The dates are used as subsection headings with all events related to them.
False negative: complex reasoning	• November 2055, CEA begins to increase. There is abnormal uptake on a PET scan near the rectosigmoid junction. Patient is then initiated on XELIRI/Avastin in February 2055. [more text..]. • May 2055 through August 2055, managed with observation alone off of all chemotherapeutic administration.	• No prediction TLINK to indicate that XELIRI/Avistin was discontinued May 2055 through August 2055.
False negative: hedging	• We had attempted to treat him with ipilimumab last week; however, when he got the bathroom in the office, he tripped over a wheel of one of the beds and had a fall.	• Gold TLINK is (last week, CONTAINS, ipilimumab). No predicted TLINK due to the expressed uncertainty of whether the event happened.
False positive: complex reasoning	• …cycles of Cytoxan, fludarabine, and Rituxan chemotherapy through July 2055.	• Predicted TLINKs are correct. However, the treatment is associated with the patient's leukemia, not the melanoma which was the targeted extraction.

[a]TLINK: pairwise temporal relations.

Discussion

Principal Findings

The implications of the automatic and faithful extraction of treatment timelines from patients' EMRs affect the spectrum

of patient-physician interactions, decision-making processes, and advances in cancer research. At the point of care, a clinician presented with the patient's treatment timeline would be able to quickly gain insights into the complex disease and treatment process for that patient, especially helpful in oncology where patients come to specialized centers with hundreds of notes.

For research, the automatic generation of timelines opens the door to creating large-scale cohorts to answer important research questions. One such question is related to the treatment regimens as key details in understanding the effects of genetic, epigenetic, and other factors on tumor behavior and responsiveness. As precision oncology progresses, insights into the fine interplay of treatment with tumor molecular characteristics and patient phenotypes become even more critical not only as a source of research data, but as a means of translating findings into patient-tailored therapies similar to those that have been applied to breast cancer and melanoma [38].

Although there is a lot of excitement around LLMs and prompt engineering, there is a major constraint that needs to be factored into engineering decisions—that of the length of the input text. This is especially pronounced for tasks where the entire patient EMR narrative needs to be considered, for example, treatment timeline extraction. When considering the input prompt for LLMs, we first considered sending 1 note at a time to LLMs, or concatenating all the sentences that contain SACT EVENT mentions in a note and sending them to LLMs. However, our experiments showed that extracting timelines from long sequences (even just one patient note) was too challenging for the LLMs we evaluated (although these were the SOTA open LLMs). For example, on the ovarian cancer development set, we saw a 10-point drop in relaxed-to-month scores when we sent multiple sentences from the same document to LLaMA2.

As the error analysis pointed out, the main source of the error is TLINK classification, that is the assignment of the correct temporal relation between an EVENT and TIMEX. The technology we experimented with is LM-based—finetuning smaller LMs and LLM prompting. A path of research to improve TLINK extraction lies in combining the outputs of various technologies into an ensemble with a voting mechanism, for example, majority vote or a classification layer. The ensemble could potentially include the output of LLM-based and non–LLM-based methods such as classic support vector machines [39]. Another potential solution might lie in exploring a 2-stage LLM finetuning strategy, which is a refined ensemble method [40]. The first stage decreases bias and variance iteratively, while in the second stage, a selected fixed-bias model is used to further reduce variance due to optimization in ensembling. Soft prompting [41] might be another viable path to explore, especially given the availability of labeled data.

Our experiments show that LLMs struggle with end-to-end timeline extraction from clinical narratives (see Figure 4B). In Table 4, an examination of label distribution across the development set highlights a strong tendency of the system to overproduce BEGINS-ON and ENDS-ON relations while underrepresenting CONTAINS-1. For example, in colorectal cancer, the system predicted 381 BEGINS-ON and 281 ENDS-ON events, vastly exceeding the gold counts of 82 and 73, respectively. A notable source of error in the system's predictions stems from confusion in relation directionality, particularly with the CONTAINS-1 relation. By design, all triples are structured as <EVENT, TLINK, TIMEX3>, where CONTAINS-1 semantically indicates that the drug was administered on the date specified by the TIMEX3 (see the Tasks and Datasets subsection in the Methods section). However, the system frequently reversed this logic, producing incorrect <EVENT, CONTAINS, TIMEX3> triples. Such mispredictions not only result in spurious labels (captured under the CONTAINS category in the label distribution) but also reflect a deeper modeling issue: the model's difficulty in internalizing fine-grained relational semantics. To mitigate this, future work could incorporate explicit prompt instruction or soft constraints to enforce the expected directionality of relations during inference in the spirit of constrained decoding [42]. In addition, postprocessing steps could validate predicted relations by checking for allowable type-direction combinations, correcting or filtering those that violate domain-specific rules.

Table . Label distribution across the gold timelines and large language model (LLM) predicted timelines (LLAMA2 70B model, end-to-end setting) on the development set.

Cancer type	Gold timelines, n			System timelines, n			
	CONTAINS-1	BEGINS-ON	ENDS-ON	CONTAINS	CONTAINS-1	BEGINS-ON	ENDS-ON
Breast cancer	16	11	12	1	2	49	21
Ovarian cancer	65	8	12	7	11	104	38
Melanoma	39	5	1	2	8	47	22
Colorectal cancer	97	82	73	87	0	381	281

The error analysis also revealed incorrect annotations in the gold labels. We identified 30 annotation errors in the sample of the shared task dataset (~3.5 million words). The number of annotation errors in the CRC dataset sample is higher, but this is also the largest dataset (12 million+ words). Thus, as a proportion, the estimated annotation error rates across the independent datasets are similar. Annotation error is a standard hazard of the annotation process, especially for a highly cognitively demanding task as the timeline extraction from the entire patient's chart. One has to review every single document from the patient's chart, which for oncology patients translates into hundreds, if not thousands, of notes. Human errors are bound to happen. This further underscores the importance of developing methods for automatic and faithful timeline extraction.

A curious result emerges on the melanoma dataset. As shown in Table 2, the performance on the melanoma dataset is lower than the performance on other types of cancer using task-specific finetuned model. We believe this is caused by the data scarcity in the melanoma dataset because (1) SACT is not the main treatment modality for most melanoma presentations; therefore,

XSL·FO

RenderX

there are fewer instances of SACT in the melanoma data and (2) the melanoma test set is the smallest of the 4 datasets. As the evaluation script computed the average F_1-scores across all patients, the overall performance on the melanoma test set fluctuated greatly with the score of individual patients.

In this work, we focus on cancer treatment timeline extraction. However, the methodology described in this work can be applied to treatment timelines extraction of other diseases. For instance, if gold standard datasets are available for an out-of-domain disease type, one can finetune a small LM for temporal relation extraction. If gold annotations are not available for a type of disease, prompting LLMs with a few domain-specific examples would be a viable solution.

Limitations

In this work, we did not use powerful, but proprietary LLMs such as GPT-4 [26] or Gemini [43], as we do not have access to nonretaining versions of these models for large scale processing. Despite the fact that our dataset was deidentified per HIPAA requirements, we did not feel that it was ethically appropriate to submit patient-derived data to a retaining LLM. However, experimenting with open models presents a realistic scenario for the average academic center as experimenting with proprietary LLMs comes at a significant cost. The LLMs we selected in our study were those reported to have competitive performance to proprietary models [24,36]. During paper revision, the DeepSeek-R1 [44] open model was released which outperformed the proprietary models on several general benchmarks. We leave experimentation with it as a future study. We did not use prompting techniques such as chain-of-thought [45] because it is not clear how to directly convert a complex task such as timeline extraction from the entire EMR clinical narrative into a series of reasoning steps. We leave the exploration of using HIPAA-compliant versions of proprietary LLMs (access-dependent) and other prompting methods such as prompt-tuning [46-48] for future research. Another limitation is that the datasets represent 2 medical centers and thus may introduce institutional or regional biases. However, to the best

of our knowledge, these datasets are the only ones on cancer treatment timelines available to the community. In addition, this study focuses on colorectal cancer, breast cancer, ovarian cancer, and melanoma. While these common cancer types are broadly representative, future work should extend the SACT timeline extraction task to other cancer types. We should note that such pan-cancer extensions necessitate significant resources for the creation of the gold annotations. We also acknowledge that some cancer journeys are complex, with lines of therapy containing SACT interspersed with other therapeutic modalities such as radiation; these complexities are out of scope for the current approach but should be a focus of future work. Finally, this work uses an established set of predefined temporal relations (CONTAINS, BEGINS-ON, ENDS-ON, OVERLAP, and BEFORE) and preexisting annotations. We acknowledge that modeling more complex and nuanced temporal scenarios might potentially provide additional insights; however, this is the core set the clinical temporal information extraction community has converged on with some minor nuances [1,2,14].

Conclusions

In this paper, we explored approaches for patient-level timeline extraction through the task of SACT timeline extraction. We performed experiments on the 2024 ChemoTimelines shared task as well as on the THYME dataset, thus the data represented 4 types of cancer across two institutions. We finetuned an LM that was specifically trained to attend to EVENT and TIMEX3 mentions. In that, we achieved higher scores than all shared task participants in Subtask1. We also explored LLM-based systems via prompting. In both subtasks, our LLM-based systems outperformed the shared task participant systems that took the approach of prompting LLMs. Our results contribute to the body of work that shows that task-specific finetuning based on rich, disease-specific datasets outperforms prompting the current generalist LLMs. We believe our results and analysis on this task add to the knowledge of extracting treatment timelines in EMRs using NLP methods. Our code will be released publicly upon acceptance.

Data Availability

The colorectal cancer dataset analyzed during this study is available to those involved in natural language processing research under a data use agreement (DUA) with Mayo Clinic. The corpus is distributed through the hNLP Center (center.healthnlp.org). The breast cancer, ovarian cancer, and melanoma datasets analyzed during this study are available under a DUA. Please contact author HH (email: harryh@pitt.edu) for details.

Authors' Contributions

JY contributed to conceptualization and methodology, performed the experiments with the assistance of EG and GS, and contributed to formal analysis, writing the original draft, review, and editing. EG contributed to conceptualization, methodology, software,

XSL•FO

RenderX

formal analysis, writing the original draft, review, and editing. HH and SF contributed to data acquisition, writing – review and editing. JL contributed to data acquisition. DH contributed to data curation and error analysis. PCdeG contributed to subject matter expertise, data curation, and writing – review and editing. EB contributed subject matter expertise. DB contributed to subject matter expertise and writing – review and editing. JLW contributed to subject matter expertise and writing – review and editing. GS contributed to conceptualization, data curation, methodology, formal analysis, writing the original draft, review and editing, funding acquisition, and project administration. All authors discussed the results and commented on the manuscript.

Conflicts of Interest

EB serves as a consultant/advisory board member for Pfizer, Werewolf Pharma, Merck, Iovance, Sanofi, Xilio, and Novartis. Clinical trial support from Lilly, Novartis, Partners Therapeutics, Genentech, and BVD. JLW reports consulting for Westat, The Lewin Group, and ownership in HemOnc.org LLC. He is also editor-in-chief of JCO CCI. DB reports Scientific Advisory Board membership for MercurialAI. She is also associate editor of Radiation Oncology, HemOnc.org (no financial compensation, unrelated to this work) and an associate editor of JCO CCI; funding from American Association for Cancer Research (unrelated to this work). None declared by the other authors.

Multimedia Appendix 1
More details on data statistics, experimental settings, results, and error analysis.
[DOCX File, 44 KB - bioinform_v6i1e67801_app1.docx]

References

1. Styler WF 4th, Bethard S, Finan S, et al. Temporal annotation in the clinical domain. Trans Assoc Comput Linguist 2014 Apr;2:143-154. [doi: 10.1162/tacl_a_00172] [Medline: 29082229]
2. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. J Am Med Inform Assoc 2013;20(5):806-813. [doi: 10.1136/amiajnl-2013-001628] [Medline: 23564629]
3. Bethard S, Derczynski L, Savova G, Pustejovsky J, Verhagen M. SemEval-2015 task 6: clinical tempeval. Presented at: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015); Jun 4-5, 2015; Denver, Colorado. [doi: 10.18653/v1/S15-2136]
4. Bethard S, Savova G, Chen WT, Derczynski L, Pustejovsky J, Verhagen M. SemEval-2016 task 12: clinical tempeval. Presented at: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016); Jun 16-17, 2016; San Diego, California. [doi: 10.18653/v1/S16-1165]
5. Bethard S, Savova G, Palmer M, Pustejovsky J. SemEval-2017 task 12: clinical tempeval. Presented at: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017); Aug 3-4, 2017; Vancouver, Canada. [doi: 10.18653/v1/S17-2093]
6. Laparra E, Xu D, Elsayed A, Bethard S, Palmer M. SemEval 2018 task 6: parsing time normalizations. Presented at: Proceedings of The 12th International Workshop on Semantic Evaluation; Jun 5-6, 2018; New Orleans, Louisiana. [doi: 10.18653/v1/S18-1011]
7. Dligach D, Miller T, Lin C, Bethard S, Savova G. Neural temporal relation extraction. Presented at: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers; Apr 3-7, 2017; Valencia, Spain. [doi: 10.18653/v1/E17-2118]
8. Lin C, Miller T, Dligach D, Amiri H, Bethard S, Savova G. Self-training improves recurrent neural networks performance for temporal relation extraction. Presented at: Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis; Nov 1, 2018; Brussels, Belgium. [doi: 10.18653/v1/W18-5619]
9. Lin C, Miller T, Dligach D, Bethard S, Savova G. A BERT-based universal model for both within- and cross-sentence clinical temporal relation extraction. Presented at: Proceedings of the 2nd Clinical Natural Language Processing Workshop; Jun 9, 2019; Minneapolis, Minnesota, USA. [doi: 10.18653/v1/W19-1908]
10. Lin C, Miller T, Dligach D, Sadeque F, Bethard S, Savova G. A BERT-based one-pass multi-task model for clinical temporal relation extraction. Presented at: Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing; Jun 9, 2020. [doi: 10.18653/v1/2020.bionlp-1.7]
11. Lin C, Miller T, Dligach D, Bethard S, Savova G. EntityBERT: entity-centric masking strategy for model pretraining for the clinical domain. Presented at: Proceedings of the 20th Workshop on Biomedical Language Processing; Jun 11, 2021. [doi: 10.18653/v1/2021.bionlp-1.21]
12. Yuan C, Xie Q, Ananiadou S. Zero-shot temporal relation extraction with chatgpt. Presented at: The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks; Jul 13, 2023; Toronto, Canada. [doi: 10.18653/v1/2023.bionlp-1.7]
13. Yao J, Hochheiser H, Yoon W, Goldner E, Savova G. Overview of the 2024 shared task on chemotherapy treatment timeline extraction. Presented at: Proceedings of the 6th Clinical Natural Language Processing Workshop; Jun 21, 2024; Mexico City, Mexico. [doi: 10.18653/v1/2024.clinicalnlp-1.53]

14. Wright-Bettner K, Lin C, Miller T, et al. Defining and learning refined temporal relations in the clinical narrative. Presented at: Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis; Nov 20, 2020. [doi: 10.18653/v1/2020.louhi-1.12]

15. Haddadan S, Le TD, Duong T, Thieu T. LAILab at chemotimelines 2024: finetuning sequence-to-sequence language models for temporal relation extraction towards cancer patient undergoing chemotherapy treatment. Presented at: Proceedings of the 6th Clinical Natural Language Processing Workshop; Jun 21, 2024; Mexico City, Mexico. [doi: 10.18653/v1/2024.clinicalnlp-1.37]

16. Wang L, Lu Q, Li R, Fu S, Liu H. Wonder at chemotimelines 2024: medtimeline: an end-to-end NLP system for timeline extraction from clinical narratives. Presented at: Proceedings of the 6th Clinical Natural Language Processing Workshop; Jun 21, 2024; Mexico City, Mexico. [doi: 10.18653/v1/2024.clinicalnlp-1.48]

17. Sharma V, Fernandez A, Ioanovici A, Talby D, Buijs F. Lexicans at chemotimelines 2024: chemotimeline chronicles - leveraging large language models (llms) for temporal relations extraction in oncological electronic health records. Presented at: Proceedings of the 6th Clinical Natural Language Processing Workshop; Jun 21, 2024; Mexico City, Mexico. [doi: 10.18653/v1/2024.clinicalnlp-1.38]

18. Bannour N, Andrew JJ, Vincent M. Team nlpeers at chemotimelines 2024: evaluation of two timeline extraction methods, can generative LLM do it all or is smaller model fine-tuning still relevant? Presented at: Proceedings of the 6th Clinical Natural Language Processing Workshop; Jun 21, 2024; Mexico City, Mexico. [doi: 10.18653/v1/2024.clinicalnlp-1.39]

19. Chung HW, Hou L, Longpre S, Zoph B, Tai Y, Fedus W, et al. Scaling instruction-finetuned language models. J Mach Learn Res 2024;25:1-53 [FREE Full text]

20. Lewis M, Liu Y, Goyal N, et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. Presented at: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; Jul 5-10, 2020. [doi: 10.18653/v1/2020.acl-main.703]

21. Lewis P, Ott M, Du J, Stoyanov V. Pretrained language models for biomedical and clinical tasks: understanding and extending the state-of-the-art. Presented at: Proceedings of the 3rd Clinical Natural Language Processing Workshop; Nov 19, 2020. [doi: 10.18653/v1/2020.clinicalnlp-1.17]

22. Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthcare 2022 Jan 31;3(1):1-23. [doi: 10.1145/3458754]

23. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: open foundation and fine-tuned chat models. arXiv. Preprint posted online on Jul 19, 2023. [doi: 10.48550/arXiv.2307.09288]

24. Jiang AQ, Sablayrolles A, Roux A, Mensch A, Savary B, Bamford C, et al. Mixtral of experts. arXiv. Preprint posted online on Jan 8, 2024. [doi: 10.48550/arXiv.2401.04088]

25. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv. Preprint posted online on May 28, 2020. [doi: 10.48550/arXiv.2005.14165]

26. OpenAI. GPT-4 technical report. arXiv. Preprint posted online on Mar 15, 2023. [doi: 10.48550/arXiv.2303.08774]

27. Raffel C, Shazeer NM, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv. Preprint posted online on Oct 23, 2019. [doi: 10.48550/arXiv.1910.10683]

28. Kweon S, Kim J, Kim J, et al. Publicly shareable clinical large language model built on synthetic clinical notes. In: Ku LW, Martins A, Srikumar V, editors. Presented at: Findings of the Association for Computational Linguistics ACL 2024; Aug 11-16, 2024; Bangkok, Thailand. [doi: 10.18653/v1/2024.findings-acl.305]

29. Team G, Mesnard T, Hardin C, et al. Gemma: open models based on gemini research and technology. arXiv. Preprint posted online on Apr 16, 2024. [doi: 10.48550/ARXIV.2403.08295]

30. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. Presented at: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; Jun 2-7, 2019; Minneapolis, Minnesota. [doi: 10.18653/v1/N19-1423]

31. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507-513. [doi: 10.1136/jamia.2009.001560] [Medline: 20819853]

32. Bethard S. A synchronous context free grammar for time normalization. Presented at: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing; Oct 18-21, 2013; Seattle, Washington, USA. [doi: 10.18653/v1/D13-1078]

33. Bethard S, Parker J. A semantically compositional annotation scheme for time normalization. Presented at: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16); May 23-28, 2016; Portorož, Slovenia URL: https://aclanthology.org/L16-1599/ [accessed 2025-07-31]

34. Pustejovsky J, Lee K, Bunt H, Romary L. ISO-timeml: an international standard for semantic annotation. Presented at: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10); May 17-23, 2010; Valletta, Malta URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/55_Paper.pdf [accessed 2025-07-31]

35. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. LLaMA: open and efficient foundation language models. arXiv. Preprint posted online on Feb, 2023. [doi: 10.48550/arXiv.2302.13971]

36. Dubey A, et al. The llama 3 herd of models. arXiv. Preprint posted online on Aug 15, 2024. [doi: 10.48550/arXiv.2407.21783]

37.   Tan Y, Dede M, Chen K. KCLab at chemotimelines 2024: end-to-end system for chemotherapy timeline extraction –
      subtask2. Presented at: Proceedings of the 6th Clinical Natural Language Processing Workshop; Jun 21, 2024; Mexico
      City, Mexico. [doi: 10.18653/v1/2024.clinicalnlp-1.40]
38.   Shin SH, Bode AM, Dong Z. Addressing the challenges of applying precision oncology. NPJ Precis Oncol 2017;1(1):28.
      [doi: 10.1038/s41698-017-0032-z] [Medline: 29872710]
39.   Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995 Sep;20(3):273-297. [doi: 10.1007/BF00994018]
40.   Wang L, Li Y, Miller T, Bethard S, Savova G. Two-stage fine-tuning for improved bias and variance for large pretrained
      language models. Presented at: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
      (Volume 1; Jul 9-14, 2023; Toronto, Canada. [doi: 10.18653/v1/2023.acl-long.877]
41.   Yao J, Perova Z, Mandloi T, Lewis E, Parkinson H, Savova G. Extracting knowledge from scientific texts on patient-derived
      cancer models using large language models: algorithm development and validation. bioRxiv. Preprint posted online on Jan
      29, 2025. [doi: 10.1101/2025.01.28.634527]
42.   Geng S, Josifoski M, Peyrard M, West R. Grammar-constrained decoding for structured NLP tasks without finetuning. In:
      Bouamor H, Pino J, Bali K, editors. Presented at: Proceedings of the 2023 Conference on Empirical Methods in Natural
      Language Processing; Dec 6-10, 2023; Singapore. [doi: 10.18653/v1/2023.emnlp-main.674]
43.   Saab K, Tu T, Weng WH, Tanno R, Stutz D, Wulczyn E, et al. Capabilities of gemini models in medicine. arXiv. Preprint
      posted online on May 1, 2024. [doi: 10.48550/arXiv.2404.18416]
44.   DeepSeek-AI GD, Yang D, et al. DeepSeek-R1: incentivizing reasoning capability in llms via reinforcement learning.
      arXiv. Preprint posted online on Jan 22, 2025. [doi: 10.48550/arXiv.2501.12948]
45.   Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. arXiv. Preprint
      posted online on Jan 28, 2022. [doi: 10.48550/arXiv.2201.11903]
46.   Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning. In: Moens MF, Huang X,
      Specia L, Yih SW, Eds. O, Cana P, editors. Presented at: Proceedings of the 2021 Conference on Empirical Methods in
      Natural Language Processing; Nov 7-11, 2021; Punta Cana, Dominican Republic. [doi: 10.18653/v1/2021.emnlp-main.243]
47.   Li XL, Liang P. Prefix-tuning: optimizing continuous prompts for generation. In: Zong C, Xia F, Li W, Navigli R, editors.
      Presented at: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th
      International Joint Conference on Natural Language Processing; Aug 1-6, 2021. [doi: 10.18653/v1/2021.acl-long.353]
48.   Liu X, Zheng Y, Du Z, et al. GPT understands, too. AI Open 2024;5:208-215. [doi: 10.1016/j.aiopen.2023.08.012]

Abbreviations

BIO: beginning, inside, outside
CRC: colorectal cancer
DocTime: Document Creation Time
DocTimeRel: relation with the Document Creation Time
EMR: electronic medical record
HIPAA: Health Insurance Portability and Accountability Act
LLM: large language model
LM: language model
NLP: natural language processing
SACT: systemic anticancer therapy
SOTA: state-of-the-art
THYME: Temporal Histories of Your Medical Event
TIMEX3: time expressions
TLINK: pairwise temporal relation

Paired-Sample and Pathway-Anchored MLOps Framework for Robust Transcriptomic Machine Learning in Small Cohorts: Model Classification Study

Mahdieh Shabanian[1]; Nima Pouladi[1]; Liam Wilson[1]; Mattia Prosperi[2]; Yves A Lussier[1,3,4]

[1]Biomedical Informatics, University of Utah, 421 Wakara Suit 140, Salt Lake City, UT, United States

[2]College of Public Health and Health Professions Dean's Office, University of Florida, Gainesville, FL, United States

[3]Center for Genomic Medicine, University of Utah, Salt Lake City, UT, United States

[4]Hunstman Cancer Cetner, University of Utah, Salt Lake City, UT, United States

Corresponding Author:
Yves A Lussier
Biomedical Informatics, University of Utah, 421 Wakara Suit 140, Salt Lake City, UT, United States

Abstract

Background: Approximately 90% of the 65,000 human diseases are infrequent, collectively affecting ~400 million people, substantially limiting cohort accrual. This low prevalence constrains the development of robust transcriptome-based machine learning (ML) classifiers. Standard data-driven classifiers typically require cohorts of more than 100 participants per group to achieve clinical accuracy while managing high-dimensional input (~25,000 transcripts). These requirements are infeasible for microcohorts of ~20 individuals, where overfitting becomes pervasive.

Objective: To overcome these constraints, we developed a classification method that integrates three enabling strategies: (i) paired-sample transcriptome dynamics, (ii) N-of-1 pathway-based analytics, and (iii) reproducible machine learning operations (MLOps) for continuous model refinement.

Methods: Unlike ML approaches relying on a single transcriptome per subject, within-subject paired-sample designs—such as pre- versus post-treatment or diseased versus adjacent-normal tissue—effectively control intraindividual variability under isogenic conditions and within-subject environmental exposures (eg, smoking history, other medications, etc), improve signal-to-noise ratios, and, when pre-processed as single- studies (N-of-1), can achieve statistical power comparable with that obtained in animal models. Pathway-level N-of-1 analytics further reduces each sample's high-dimensional profile into ~4000 biologically interpretable features, annotated with effect sizes, dispersion, and significance. Complementary MLOp practices—automated versioning, continuous monitoring, and adaptive hyperparameter tuning—improve model reproducibility and generalization.

Results: In two case studies of distinct diseases, human rhinovirus infection (HRV) versus matched healthy controls (n=16 training; n=3 test) and breast cancer tissues harboring *TP53* or *PIK3CA* mutations versus adjacent normal tissue (n=27 training; n=9 test)—this approach achieved 90% precision and recall on an unseen breast cancer test set and 92% precision with 90% recall in rhinovirus fivefold cross-validation. Incorporating paired-sample dynamics boosted precision by up to 12% and recall by 13% in breast cancer and by 5% each in HRV. MLOps workflows yielded an additional ~14.5% accuracy improvement compared to traditional pipelines. Moreover, our method identified 42 critical gene sets (pathways) for rhinovirus response and 21 for breast cancer mutation status, selected as the most important features (mean decrease impurity) of the best-performing model, with retroactive ablation of top 20 features reducing accuracy by ~25%.

Conclusions: These proof-of-concept results support the utility of integrating intrasubject dynamics, "biological knowledge"-based feature reduction (pathway-level feature reduction grounded in prior biological knowledge; eg, N-of-1-pathway analytics), and reproducible MLOp workflows can overcome cohort size limitations in infrequent disease, offering a scalable, interpretable solution for high-dimensional transcriptomic classification. Future work will extend these advances across various therapeutic and small cohort designs.

XSL•FO

RenderX

Introduction

Precision medicine seeks to personalize health care by accounting for individual differences in genetic makeup, environmental exposures, and lifestyle factors. This tailored approach becomes especially challenging when analyzing high-dimensional transcriptomic data derived from small patient cohorts (microcohorts), a scenario frequently encountered in studies of rare or infrequent diseases. Microcohorts typically involve datasets characterized by high dimensionality (approximately 25,000 transcriptomic features) juxtaposed against limited sample sizes (approximately 20 persons), conditions that commonly induce overfitting in traditional machine learning models. Advanced analytical methodologies have thus become essential in identifying robust and clinically meaningful biomarkers from these small-scale studies to facilitate personalized patient care.

A large share of the ~65,000 known human diseases are infrequent—neither rare nor common—making it difficult to assemble statistically robust cohorts without multiyear, multicenter efforts. Around 5.9% of the global population is affected by rare diseases [1], highlighting their substantial impact on global health.

Moreover, finely stratified subtypes of otherwise common diseases present similar challenges as their reduced prevalence within heterogeneous populations undermines statistical power. For example, in highly heterogeneous diseases, such as cancer, where tumor subtypes and genetic mutation profiles can vary substantially between individuals, conventional machine learning approaches often suffer from insufficient statistical power and heightened risk of overfitting. To mitigate these challenges, single-subject (N-of-1) transcriptome analytics has emerged as an innovative approach, allowing individuals to serve effectively as their own controls. By measuring within-subject transcriptomic changes and integrating these measurements into biologically interpretable pathway-level features, N-of-1 analyses significantly reduce noise and enhance the detection of biologically meaningful signals, even amidst substantial intersubject variability [2-8].

Concurrently, the emergence of machine learning operations (MLOps), inspired by DevOps practices, has significantly improved the deployment, optimization, and monitoring of machine learning (ML) models. MLOps leverage automated experiment tracking, hyperparameter tuning, and continuous integration, enhancing workflow efficiency, reliability, reproducibility, and scalability—factors essential for developing robust and maintainable models in biomedical research [9-15].

We hypothesized that integrating three complementary strategies would enhance classification accuracy and robustness in microcohort scenarios: (i) implementing MLOp frameworks to achieve robust and reproducible model performance and (ii) leveraging transcriptomic dynamics observed between paired biological samples (eg, diseased versus healthy tissues from the same individual). Paired-sample information can be incorporated in two distinct ways: (ii-a) as continuous fold-change values between matched samples or (ii-b) through single-subject (N-of-1) pathway analysis, which aggregates paired gene-level signals into biologically interpretable, ternary pathway features (upregulated, downregulated, or unchanged) across ~4000 human curated biological pathways annotated along with their respective effect sizes and significance levels.

To empirically test this hypothesis, we conducted a proof-of-concept analysis on two distinct human microcohorts, one in breast cancer (BC) (TP53 vs PIK3CA tumors) and one in human rhinovirus (HRV) infection (symptomatic vs asymptomatic), each comprising paired biological samples representing two different tissue conditions per subject. For each cohort, we systematically evaluated three distinct data transformation strategies: [i] conventional analysis using only the affected tissue per subject [ii], fold-change transformation involving the ratio of affected tissue mRNA expression to paired control tissue expression for each subject, and [iii] N-of-1-pathway transformation, summarizing individual subject-level pathway effect sizes and P values. The TP53–PIK3CA contrast provides a clinically relevant and mechanistically distinct testbed: both genes are frequent drivers in BC, associated with divergent transcriptomic programs and prognostic implications across the Cancer Genome Atlas and independent cohorts. Their prevalence and biological differences make them suitable paired-sample targets to evaluate whether within-subject transformations amplify signal over baseline variability.

Each of these 3 data transformations was subjected to classification modeling both with and without incorporating MLOps, resulting in a total of 12 experimental conditions across both cohorts. To further validate the robustness and relevance of features selected by the best-performing classifier, we conducted a rigorous retrospective ablation analysis. Specifically, in ablation analysis, we masked the top 20 y discriminative features from the dataset and assessed the resulting impact on classification accuracy and stability. This comprehensive analysis framework allowed us to quantify the individual contributions of key biomarkers to the model's predictive performance.

Methods

Ethical Considerations

All transcriptomes were obtained as expression files from public published USA NIH datasets (gene Expression Omnibus and TCGA). Such expression data are not considered protected human information under HIPAA.

Human Cohort Datasets

Two distinct human cohorts, spanning cancer and infection, were selected to test our framework: a BC cohort (oncogene drivers TP53 vs PIK3CA) and a HRV infection cohort (symptomatic vs asymptomatic). Both cohorts were characterized by small sample sizes, varying heterogeneity, and paired tissue samples per subject (Table 1). Processing followed published methods, ensuring prior studies' comparability [16,17].

Table . Description of the two human cohort datasets[a].

Dataset	HRV[b] Dataset	BC[c] Dataset
Source-reference	GSE17156 (downloaded 9/17/2014) [18]	The Cancer Genome Atlas (TCGA-BRCA (downloaded 03/05/2019) [19,20]
Platform	Microarrays: Human Gene U133A 2.0	Illumina Hi-Seq 2000 (version 2 analyses)
Paired tissues	PBMC[d] samples drawn before and 48 hours after HRV nasal inoculation	Primary breast carcinoma biopsies (affected) versus unaffected breast tissue margins
Experimental groups	Symptom measures before and after successful inoculation (virus present in sputum confirmed): [1] asymptomatic vs [2] symptomatic (headache, throat ache, rhinorrhea, and/or mild fever)	Somatic (tumor) mutations in either [1] *TP53* or [2] *PIK3CA* (cases with both mutations or none of these excluded)
Individuals	19 healthy adult volunteers	42 patients
#total	10 symptomatic for common cold	*TP53* (23 patients)
	9 asymptomatic	*PIK3CA* (19 patients)
Sample	38 gene expression microarray files	84 RNAseq count files

[a] #: count of individuals.

[b]HRV: human rhinovirus

[c]BC: breast cancer.

[d]PBMC: peripheral blood mononuclear cell

The classification task in BC was to identify one of two oncogene drivers that influence the treatment and prognosis, because in primary, early-stage, nonmetastatic breast carcinoma, *TP53*-mutated and *PIK3CA*-mutated tumors are generally not reliably distinguishable by histopathology alone—that is, without molecular (immunochemistry or genetic) assays. In addition, the *TP53*-driven subtype has substantially poorer 5-year survival and presents substantial resistance to therapy [21]. While the classification task in HRV was classifying symptomatic infected individuals versus asymptomatic infected ones.

Additionally, we downloaded Gene Ontology (GO) Biological Process and their gene annotations termed "gene sets," downloaded from Ashburner et al on January 3, 2024 [22].

Dataset Transformations

One Affected Tissue Transcriptome Per Individual

Most conventional transcriptome classifiers typically analyze a single transcriptome derived from the affected tissue of each individual. To evaluate the accuracy achievable with traditional classification methods using one sample per individual, we used the affected tissue of the datasets and did not use the paired control tissue. The BC cohort [16]included 22,279 TMM (trimmed mean of M values) normalized gene expression [23] values from 42 individuals, and two samples per individual (BC and unaffected margins). The HRV cohort [16]included 20,502 RMA-normalized Affymetrix GeneChip expressions of probe sets from 19 individuals and two samples per individual (peripheral blood mononuclear cells 48 h before HRV inoculation and after successful inoculation and shedding of virus) (Figure 1, Panel A).

Figure 1. Panel A. Overview of methods and process flow of the proof-of-concept study. Classification methods are applied to 2 cohorts (Table 1), each with two distinct clinical phenotypes: (**I**) Individuals with BC, stratified by oncogenic drivers (TP53 vs PIK3CA), and (ii) HRV-infected patients (symptomatic vs asymptomatic). Each subject provides 2 samples under different conditions: (i) BC—within-subject comparison of cancerous tissue vs. unaffected margins, and (ii) HRV—within-subject comparison before versus during infection. Six classification experiments are conducted on each cohort's extracted transcriptomes, evaluating 3 complementary classification strategies for microcohorts: (i) MLOps-driven robustness (Panel B), (ii) transcriptome dynamics between paired samples (eg, exposed vs unexposed tissue), and (iii) single-subject pathway analytics (N-of-1; details in Figure 2). Panel B. RF classifier pipeline of the BC dataset. The RF classification workflow consists of 5 key steps after extracting an unseen evaluation set: (i) hyperparameter tuning using Weights & Biases MLOp sweep definition, (ii) human-in-the-loop expert heuristics to assess failure patterns and overfitting (YAML‑based sweep configuration: criterion, max_depth, max_features, n_estimators via wandb.sweep function), (iii) iterative model refinement via 300 resampling cycles of 5-fold cross-validation (80% samples in the training set, 20% in the validation set, orchestrated by W&B MLOps (wandb.agent), (iv) MLOps Automated Best Model Selection, and (v) final evaluation on unseen dataset. Panel C. Retroactive feature ablation analysis: feature importance is assessed in both datasets to evaluate the impact of individual features on classification performance. BC: breast cancer; HRV: human rhinovirus; RF: Random Forest;
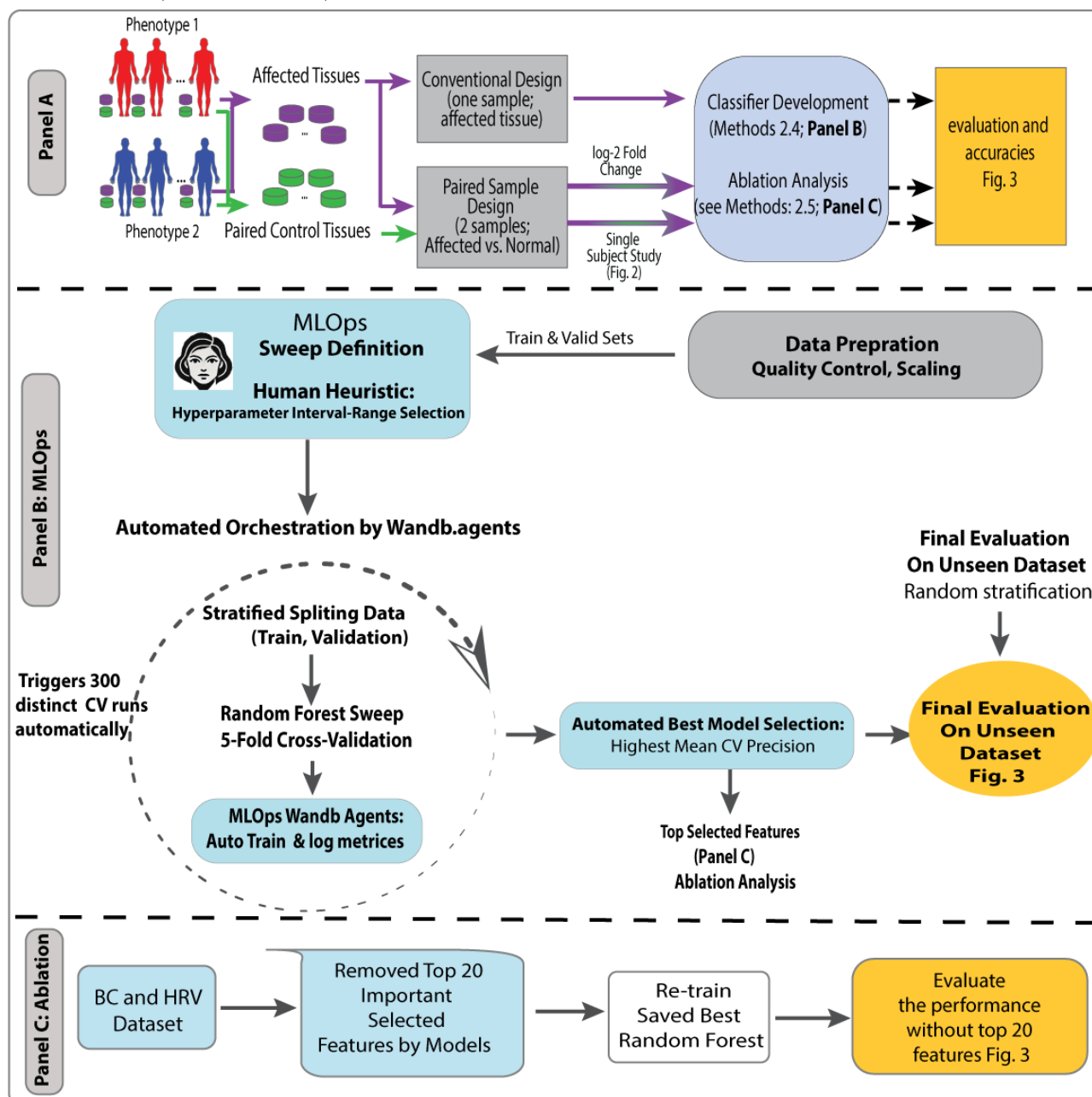
Figure 2. Description of the N-of-1-pathway Wilcoxon analytics in each single subject. We used the "N-of-1-pathways" method [17], which aggregates paired RNA-level signals of each subject into pathway-level effect sizes, conducts a nonparametric Wilcoxon test comparing the pathway-associated mRNAs in each Gene Ontology (GO) Biological Processes [$P < .05$; other thresholds studied elsewhere [2,3,6,16,17] for each subject, enabling downstream classification over a smaller number of human-interpretable GO features. This method identifies significantly altered mRNA sets associated with a pathway between two samples of one subject, yielding 4,442 GO mRNA sets in the BC cohort and 2,332 GO mRNA sets in the HRV cohort. The output consists of ternary matrices indicating response status 1−: negatively regulated, +1: positively regulated, and 0: unaltered GO pathway. For each GO pathway, we compute FC of mRNA expression values between the affected and control tissue of a single individual. A Wilcoxon test is then performed on these values, where the sum of positive ranks (W+) and negative ranks (W−) determine the test statistic W by min (W+,W−). The relative magnitude of W+versus W− indicates whether the pathway is positively or negatively regulated in a significant test (eg, W+>W− indicates a positively regulated pathway; W−>W+ indicates a negatively regulated pathway, and a nonsignificant test indicates an unaltered pathway). HRV scores were refined with a coefficient of variation <31%. FC: fold change; n: number of subjects; P or p: number of features (transcripts); W+: statistically significant Wilcoxon test with up-regulated gene set (pathway score positive); W-l: statistically significant Wilcoxon test with downregulated gene set (pathway score positive); negative): mRNA=messenger RNA; X2: indices of the affected tissue; X1: indices of the control tissue.
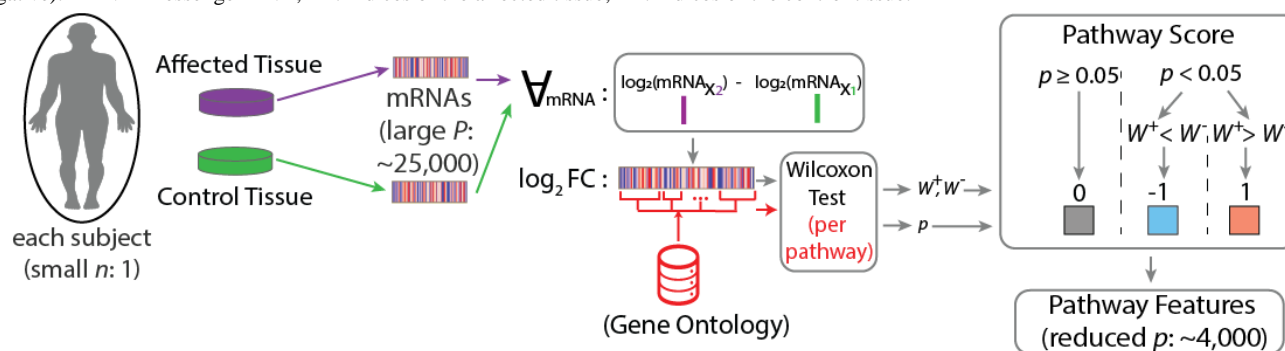
Paired Samples: One Affected Tissue Transcriptome and One Control Tissue Per Subject.

We calculated the fold change by dividing the expression of each mRNA value of the affected tissue by that of the control tissue, in each subject, in each dataset, followed by a $\log_2$ transformation [6]. Single-subject studies (N-of-1-pathways) are described in Figure 1 Panel A and Figure 2.

Model Selection

We evaluated several ML models, including Random Forest (RF), XGBoost, Support Vector Machine (SVM), and Logistic Regression. Random Forest was ultimately chosen due to its robustness, capacity to model nonlinear interactions, and superior predictive performance in identifying symptomatic patients and relevant gene sets. Multimedia Appendix 1 provides a comparative analysis of the ML models, highlighting the factors underlying RF's superior performance.

Classification, Cross-validation, and MLOps

Model robustness was evaluated in both datasets using 5-fold cross-validation. The RF model was integrated into the Weights & Biases (W&B) MLOp framework (W&B v0.17.0, Python 3.11.4) [24] to systematically identify features whose interactions significantly contribute to class differentiation. MLOps facilitated robust experiment tracking, hyperparameter optimization, and model monitoring, applying consistent hyperparameter ranges across the BC dataset (42 samples) and the HRV dataset (19 samples). This setup allowed us to assess MLOps' effectiveness in guiding hyperparameter tuning and model tracking while maintaining human oversight. This study was designed to compare the ability of different combinations of data transformations (single-sample per individual, FC, N-of-1-pathways analytics) to improve performance in small human cohorts (small n<30 individuals) with high feature dimensionality (very large p, transcriptomes=25,000 mRNA features)

In W&B MLOps, the *sweep.yaml* file configured hyperparameter sweeps by defining key parameters, search strategies, optimization metrics, and other relevant settings for systematic model optimization. Python's *StratifiedKFold* strategy ensured class proportion consistency and class imbalance across 5 folds, and this process was repeated across 5 iterations with different folds serving as the validation set, constituting a stratified 5-fold cross-validation unbiased model selection protocol. Accuracy, precision, and recall performance metrics were calculated across cross-validation folds and held-out unseen test sets (Tables 2 and 3). The held-out unseen test partition was sequestered throughout model development and accessed only once, after cross-validation and hyperparameter selection were completed, ensuring that no tuning decisions were informed by test data. To refine hyperparameter ranges, a human expert in the sweep configuration loop revised the best hyperparameter intervals using the *sweep.yaml* configuration. This YAML file specifies the parameters to be tuned, the search strategy, optimization metrics, and other pertinent settings (Figure 1, Panel B). To further evaluate generalizability given the limited cohort sizes, we performed a learning curve analysis and accompanying power calculations; results are provided in Multimedia Appendix 2, which details experimental reproducibility safeguards (eg, immutable YAML configurations, dataset/hyperparameter hashes, deterministic folds, and logging of all trials to MLOps). Methods for tracing RF classifier decisions to biological mechanisms are addressed in Multimedia Appendix 3.

Table . Performance summary of analysis in human rhinovirus (symptomatic vs asymptomatic) Random Forest classifier.[a]

Feature (transcript) transformation design	Single-sample mRNAs	Two-sample (one sample in each condition) mRNAs	
		Fold change	N-of-1 pathways (single-subject studies)
Number of features and samples			
Number of mRNA transcripts	22,279 features	12,496 features	553 (no. of GOs[b])
Training samples	15	15	16
Validation samples	4	4	3
Cross validation (CV) values			
CV accuracy: mean (SD)	.85 (.16)	.95 (.15)	.88 (.14)
CV precision: mean (SD)	.87 (.22)	.97 (.21)	.92 (.14)
CV recall: mean (SD)	.85 (.16)	.95 (.17)	.90 (.16)
CV F1: mean	.86	.91	.96
Selected feature count	266 mRNAs	112 mRNAs	42 GOs
Hyperparameters			
Entropy criterion maximum depth	87	18	42
Maximum features	log2 n-estimators: 148	null n-estimators: 56	sqrt n-estimators: 24

[a]Fold-change model achieves highest CV precision (0.97), while N-of-1 pathway model offers greater stability with the lowest CV (SD 0.14), outperforming single-sample designs across all metrics. Corresponding 90% CIs are provided in Supplement File 5 in Multimedia Appendix 4.

[b]GO: Gene ontology Biological Process gene set.

Table . Performance summary analysis in the breast cancer Random Forest (*PTP53 vs PIK3CA*) classifier[a].

Feature (transcript) transformation design	Single mRNAs	Two sample (one in each condition) mRNAs	
		Fold change	N-of-1 pathways (single-subject studies)
Number of mRNA transcripts	20,502 features	16,384 features	4442 features (no. of GOs[b])
Training sample	27	27	27
Validation sample	6	6	6
Test samples	9	9	9
Cross validation (CV) accuracy: mean (SD)	.72 (.18)	.62 (.15)	.73 (.07)
Unseen test set accuracy			
Test accuracy	.78	.78	.89
Test precision	.78	.86	.90
Test recall	.77	.78	.90
Test F1	.78	.82	.90
Selected feature count	105 mRNAs	97 mRNAs	21 GOs
Entropy criterion maximum depth	115	148	165
Maximum features	sqrt n-estimators: 17	log2 n-estimators: 23	null n-estimators: 8

[a]On the unseen test set, pathway-level features achieved 12% higher accuracy and greater stability compared with fold-change and single-sample classifications. Corresponding 90% CIs are provided in Multimedia Appendix 4.

Feature Importance, Stability, and Top-K Retroactive Feature Ablation

For each dataset and representation (single-sample mRNA, fold-change mRNA, and N-of-1 pathway scores), we trained RF under repeated, stratified 5-fold cross-validation. Within each fit, "feature importance" was computed as a mean decrease in impurity (MDI)—the sample-weighted reduction in node impurity attributable to a feature—and then aggregated across trees, folds, and repeats to yield a global ranking [25]. To assess

the "stability" of per-repeat rankings, we computed (i) Spearman rank correlation (ρ) on the full ordering and (ii) Jaccard overlap of the Top-k feature sets [26-28]. *Top-K* denotes the k highest-ranked features by aggregated MDI computed on the full, unpruned feature space.

We conducted a retroactive feature ablation analysis on both datasets to assess the impact of the top-ranked features identified by our selected classifiers. To harmonize ablations across representations, we prespecified k=20 (two final models selected 21 features, motivating a common k). For retroactive ablation, we removed the top 20 features from the training, validation, and held-out test partitions. We then refit from scratch the previously selected model configuration with its exact, prechosen hyperparameters, without additional tuning or human-in-the-loop changes. The held-out test set, transformed once by dropping the same training-derived top 20 indices, was evaluated a single time. This remove-and-refit procedure estimates the marginal contribution of top-ranked features while minimizing information leakage [29]. This retraining step was conducted to measure the influence of the ablated features on performance metrics such as precision and recall (Figure 1, Panel C). Together, MDI rankings, stability metrics, and ablation results provide post hoc explainability of the model's global feature contributions [30].

Results

In both datasets, RF model robustness was evaluated using 5-fold cross-validation (Methods 2.3‐2.4, Figure 1 Panels A-B). For the 42 individuals BC dataset (23 *TP53* and 19 *PIK3CA*), 80% (27 individuals) was used for training, while the remaining 20% was split into 6 individuals for validation and 9 individuals for testing, ensuring consistent evaluation. Similarly, in the HRV dataset, consisting of 19 individuals (10 symptomatic and 9 asymptomatic), the data were split into 80%

(16 individuals) for training and 20% (3 individuals) for validation. The *StratifiedKFold* approach from the *scikit-learn* Python package was used to maintain consistent class proportions across folds, ensuring validation consistency and reproducibility, and preserve class proportions in every training/validation split (class imbalance results not shown). In MLOp-guided studies (Methods 2.4, Figure 1 Panel B), after testing various hyperparameter interval ranges, a human-in-the-loop (expert) confirmed the following optimal RF hyperparameters: criterion (*gini* or *entropy*), number of estimators (5 to 150), maximum features (sqrt, log2, or None), and tree depth (5 to 200). SVM and XGBoost hyperparameters are not shown as they yielded lower accuracies. As summarized in Table 2 (HRV), Table 3 (BC), and Figure 3, paired-sample–based feature transformation strategies outperformed single-sample approaches across all major evaluation metrics. In the HRV cohort, the fold-change model yielded the highest cross-validation (CV) precision (0.97±0.21 SD) and recall (0.95±0.17 SD), while the N-of-1 pathway-based classifier demonstrated superior stability, achieving lower CV SDs across all metrics, including CV precision. By contrast, the single-sample model achieved a CV precision of only 0.87±0.22 SD. To further assess generalizability given the limited cohort sizes, we performed a learning-curve analysis using N-of-1 pathways as an exemplar; these results are provided in Multimedia Appendix 2. To trace classifier-selected features back to underlying biological processes, we performed heatmap clustering of features and pathway enrichment of transcripts, provided in Multimedia Appendix 3. To address the concern that the superior performance of the N-of-1 pathway method may reflect dimensionality reduction rather than pathway biology, we performed an additional analysis applying comparable feature reduction (~4000 features) to the single-sample and fold-change models; these results are provided in Multimedia Appendix 5.

Figure 3. Paired sample per subject machine learning designs outperform single-sample designs. As illustrated, paired-sample per subject designs, either using $\log_2$ fold change (purple) or single-subject N-of-1-pathway analysis (yellow), achieved higher classification accuracies compared to single-sample per subject designs (green), across both Breast Cancer (BC) and Human Rhinovirus (HRV) microcohorts. However, pathway-based classification surpassed fold-change performance in only one dataset, suggesting that the underlying biology (eg, structure of the information model) of a condition may determine whether pathway-level or gene-level (fold-change) features are more informative for classification. No MLOps indicates a conventional cross-validation run without iterative retraining or sweep-based refinement, serving as the baseline against which the orchestrated MLOp pipeline was compared. Incorporating MLOps (circles and squares) yielded an average accuracy improvement of ~14.5% compared to traditional approaches without iterative retraining (ie, single cross-validation runs). By contrast, classifiers subjected to retroactive top 20 feature ablation (indicated by squares) experienced a performance drop of approximately 25%. MLOps: Machine Learning Operations; mRNA: messenger RNA; Log2: logarithm base 2

mRNA Transformation / Classifier Development	One Transcriptome Per Subject	Two Paired Transcriptomes	
		log2 Fold Change	Single Subject Study (N-of-1 Pathways)
No MLOps	▽ (green)	▽ (purple)	▽ (yellow)
MLOps — Fully Featured	● (green)	● (purple)	● (yellow)
MLOps — Features Ablation	■ (green)	■ (purple)	■ (yellow)
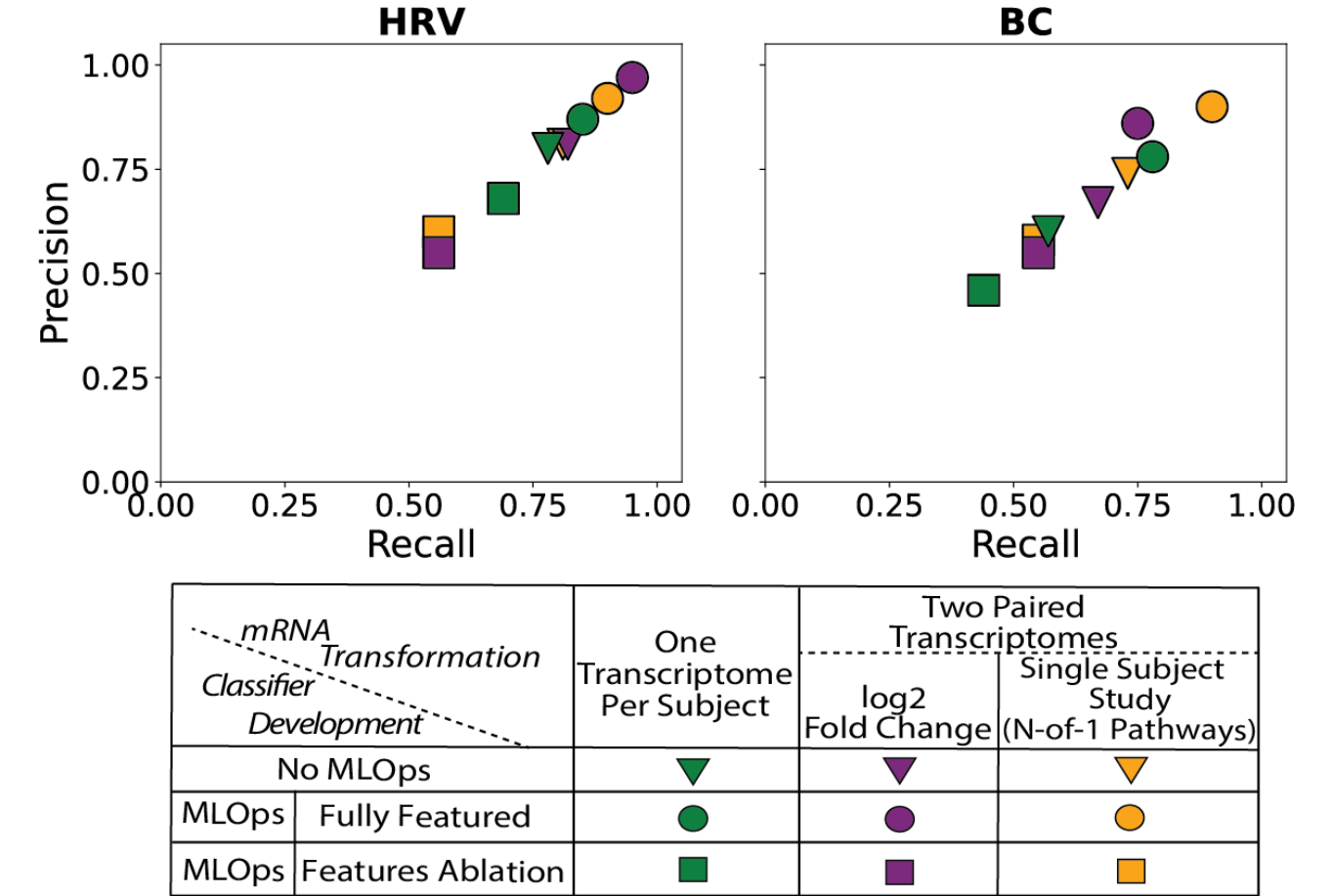
In the BC cohort, the N-of-1 pathway–based model achieved the highest test precision and recall of 0.90, reflecting an approximate 12% absolute improvement over the single-sample classifier (0.78 precision, 0.77 recall). This model also demonstrated greater stability, with a cross-validation SD approximately half that of the single-sample and fold-change approaches. These findings are further illustrated in Figure 3, which compares performance across transformation strategies. Collectively, the results demonstrate the effectiveness of paired-sample transformations—particularly when combined with MLOp-guided optimization—in improving classification accuracy and model stability in micro-cohort settings.

Retroactive feature ablation studies were conducted in breast cancer and HRV datasets (Figure 3; Multimedia Appendix 5). To assess the impact of top-ranked features on model performance, an ablation study was performed by sequentially removing the 20 highest-ranked features identified by the classifiers and retraining the optimal Random Forest model with previously tuned hyperparameters. It consisted of masking these features from the data input and retraining (Methods 2.5; Figure 1 Panel C). This analysis quantified the contribution of these features by evaluating changes in precision and recall, revealing a significant decline in predictive accuracy upon their removal. The results underscore the robustness of the selected features derived through the MLOps-driven pipeline, with performance degradation observed across all feature sets. Of note, most classifiers retained on the order of ~100 features, whereas the final BC model retained only 21 N-of-1 pathway features; nonetheless, the ablation step uniformly removed the top 20 features across all methods to maintain consistency, regardless of the total feature count. In addition, we evaluated models trained using only the top 20 features, which performed substantially better than the ablated models but below the full models (Multimedia Appendix 5), thereby quantifying both the predictive value and the limitations of this small feature subset.

Discussion

Principal Findings and Comparison With Previous Works

Transcriptome classifiers traditionally analyze a single transcriptome per subject, providing a baseline for evaluating

the performance of standard classification methods. In our study, this conventional approach was represented by the single-sample per subject design. Specifically, the BC cohort [11] included 22,279 gene expression values normalized using the trimmed mean of M values (TMM) method [20] from 42 individuals, while the HRV cohort [6] comprised 20,502 Affymetrix GeneChip probe-set expressions normalized using Robust Multiarray Average (RMA) from 19 individuals.

We systematically compared 3 mRNA feature transformation strategies—single-sample, $\log_2$ fold-change (paired design), and N-of-1 pathways (paired design)—across both datasets, using identical hyperparameter sweeps implemented within the W&B MLOp platform (*wandb* v0.17.0, Python 3.11.4). Among the evaluated classifiers (Random Forest, XGBoost, SVM, Logistic Regression), RF was selected for final implementation based on its ability to model nonlinear interactions and superior predictive performance in distinguishing symptomatic individuals and uncovering relevant gene sets (data not shown).

Results consistently demonstrated that paired-sample per designs outperformed single-sample designs, with up to 12% higher precision accuracy observed for the N-of-1 pathway–based approach in BC and 5% in HRV, while recall was increased by 13% and 5%, respectively. However, this performance advantage varied across datasets: while pathway-based classification outperformed fold-change in the BC cohort, fold-change achieved 10% increase in both precision and recall in the HRV dataset. The impact of pathway-level features on classification outcomes is demonstrated by their high importance rankings and the sharp ~25% accuracy drop observed in retroactive ablation, showing that the model not only learns from these features but also relies on them as key decision boundaries. Thus, the consistent finding is that 2-sample transformations outperform single-sample designs, although which representation (fold-change vs pathway) is optimal appears task- and biology-specific. At present, methods to prospectively identify which 2-sample representation will perform best in a given dataset remain undeveloped; however, the differential results here are consistent with biological granularity as oncogene-level classification in breast cancer is inherently pathway mediated, while HRV organism-level symptom classification reflects broader organismal phenotypes. These differences suggest that the underlying disease biology influences whether gene-level or pathway-level features are more informative. In our framework, pathway-level features contribute to classification by encoding coordinated transcriptomic changes into ternary indicators of pathway activation (upregulated, downregulated, or unchanged). Unlike raw expression values or continuous fold-change variables, these ternary ordinal features emphasize significant, coordinated shifts at the pathway level, providing interpretable signals that capture biological mechanisms rather than gene-level noise. These ternary variables act as global indicators of pathway perturbation, enabling the classifier to learn patterns of coordinated biological dysregulation that are not captured by individual transcripts alone. This representation reduces dimensionality by several orders of magnitude, mitigates noise from gene-level variability, and provides features with direct biological meaning.

Distinguishing TP53- from PIK3CA-driven breast cancers is clinically important: TP53 mutations predominate in estrogen receptor–negative tumors and portend poor prognosis, whereas PIK3CA mutations are frequent in estrogen receptor–positive tumors and guide PI3K/mTOR–targeted therapy [21]. Transcriptome-based classifiers that stratify TP53 versus PIK3CA mutations therefore have direct translational value for prognosis and treatment selection. Classifying transcriptome-level signals (~10 m) by oncogenic driver mutations—molecular alterations occurring at the nanometer scale is inherently a proximal task in the biological hierarchy, especially when contrasted with symptom-based classifications for HRV infection, which manifest at the meter scale. Moreover, early-stage primary breast carcinoma remains fundamentally a disease of genetic, genomic, and subcellular pathways. It is therefore more amenable to gene set–based transformations as conventional histology alone cannot reliably distinguish its molecular subtypes without adjunct immunohistochemical or genomic markers. In summary, paired-sample designs consistently improved precision and recall, as hypothesized; however, the optimal transformation method may vary by disease context, with some conditions favoring fold-change models and others better suited to single-subject gene set analysis.

Integrating MLOps into the modeling pipeline led to a ~14.5% improvement in classification accuracy compared with non-MLOps workflows that relied on a single cross-validation run without iterative retraining. This finding underscores the benefit of programmatic, reproducible, and feedback-driven model development. Moreover, our retroactive top feature ablation analysis, which involved retraining classifiers after removing the top 20 features, revealed a ~25% reduction in accuracy, demonstrating the importance of retaining high-contribution features in high-dimensional settings.

In the HRV microcohort, the MLOp-guided fold-change model achieved excellent precision (0.97) and recall (0.95), while single-sample designs were more susceptible to overfitting and noise due to higher dimensionality. By contrast, the N-of-1-pathway approach proved more effective in the BC cohort, which is characterized by heterogeneous tumor biology; this model achieved test precision and recall of 0.90. Conversely, MLOp-guided fold-change analysis in BC yielded lower precision (0.86) and recall (0.75), highlighting the relative strength of pathway-informed features for modeling complex biological variation.

Collectively, these results highlight how paired-sample designs—particularly when paired with MLOps—yield more accurate and interpretable models, especially in small cohort scenarios. Furthermore, our study demonstrates how expert-guided decisions about feature transformations (eg, fold-change vs pathways), integrated with programmatic MLOp workflows, can lead to substantial performance gains. The combination of human-in-the-loop oversight and automated optimization (as shown in Tables 2–3 and Multimedia Appendix 1) offers a pragmatic framework for building biologically grounded classifiers in data-limited settings.

Few studies have systematically addressed classifier development requirements in very small cohorts. Our previous work demonstrated feasibility in a prospective cohort [6] without comparative evaluations against conventional methods or MLOp integration. Transfer learning has shown promise in classifying cell types in single-cell RNA sequencing [31] and transcriptomic datasets derived from large human cohorts [32], but these methods have not yet been applied specifically to small human cohorts for clinical predictive analytics.

Several limitations must be noted: (i) alternative machine learning models (SVM, Logistic Regression, XGBoost) consistently underperformed relative to RF, and results were omitted for brevity. Future research should explore fusion deep learning and transfer learning approaches. (ii) Our conclusions are based on limited datasets, necessitating additional transcriptomic data or simulation studies to robustly assess generalizability. (iii) Despite efforts to control overfitting, inherent constraints persist due to small sample sizes, emphasizing the need to develop microcohorts through subsampling larger paired-sample datasets in future studies; though such datasets are uncommon.

Conclusions

Most of the approximately 65,000 known human diseases remain inadequately treated due to their rarity and the consequent scarcity of comprehensive studies. The low prevalence of these diseases severely limits conventional transcriptomic approaches as bulk RNA sequencing (bRNAseq) typically requires larger cohorts for effective classifier development. Emerging technologies such as spatial RNA sequencing and single-cell RNA sequencing present promising alternatives suitable for smaller cohort studies; however, these methods currently incur approximately 20 times higher costs per sample and capture around 5 times fewer mRNA transcripts. As these technologies become more affordable and achieve improved transcriptomic coverage, novel analytical methodologies tailored for small cohorts are expected to evolve. Additionally, transfer learning techniques, already successfully applied to large-scale transcriptomic datasets, offer considerable potential for small-cohort classification. However, standardized frameworks for applying transfer learning specifically to paired-sample designs are not yet established, highlighting an important area for future research.

This study systematically evaluated multiple complementary approaches designed to enhance the statistical power of bulk mRNA-based classification within microcohorts. Our results demonstrate that the integration of these approaches improves precision and recall by approximately 12.5% ‐ 14.5% compared to traditional single-sample methodologies. Specifically, we propose strategies that include the following: (i) leveraging paired comparisons of affected and control tissues within individual subjects, and (ii) using MLOp-guided analytical workflows combined with expert-in-the-loop oversight to ensure robustness, transparency, and reproducibility. This paired-sample methodology has been shown to improve classifier development in large cohorts [6,33-38], and here we show that, in very small cohorts, it also facilitated classifier development at both the individual mRNA level (via fold-change analysis) and the biologically interpretable knowledge-anchored pathway level (through N-of-1-pathway-based analyses leveraging Gene Ontology gene sets). Our results indicate that both paired-sample representations can outperform single-sample approaches, with fold-change or pathway-based features proving more effective depending on the underlying biological context. The adoption of MLOps practices optimized hyperparameter tuning and model deployment, as well as mitigated overtraining, while expert oversight ensured the biological validity of the results. Collectively, these strategies effectively address the challenges posed by high feature dimensionality and limited sample sizes, thereby laying the groundwork for advancing personalized therapeutic interventions in rare disease contexts.

Although this investigation primarily targeted a specific transcriptomic scale, optimal classifiers for clinical prediction are likely to incorporate comprehensive data across diverse biological scales (metabolome, genome, proteome, methylome, etc) jointly with real-world evidence and clinical dimensions. Future research endeavors should integrate transcriptomic data with multiomics approaches, medical imaging, and patient-centric outcomes to further enhance predictive accuracy and personalized medicine capabilities.

Disclaimer

Large language models (LLMs) were used solely for grammatical editing and improving manuscript clarity. Portions of the breast cancer results (a subset of Table 3 and of the BC panel of Figure 3) were previously presented as a poster titled "Pathway-Anchored Dimension Reduction and MLOps Enable Robust Micro-Cohort Genome-by-Environment Classifier Development" at the American Medical Informatics Association (AMIA) Informatics Summit, March 9–13, 2025, Pittsburgh, PA, USA. Additionally, only the subset of findings pertaining to Table 3 and the subpanel " BC" of Figure 3 were published in the proceedings of the 23rd International Conference on Artificial Intelligence in Medicine (AIME 2025), June 23–26, 2025, Pavia, Italy. In summary, the manuscript text, all Supplementary Files (Files 1–6), Figures 1 and 2, the HRV panel of Figure 3, and Tables 1 and 2 are original to this submission. A subset of the breast-cancer (BC) results previously appeared in abbreviated form in the AIME 2025 poster proceedings (33), limited to portions of Table 3 (BC performance metrics) and the BC subpanel of Figure 3. The present work adds an independent HRV cohort and expands the analytic scope (learning curves/power, feature- stability and ablation,

and MLOps-guided workflows), such that Table 3 and Figure 3 (BC panel) here differ materially in composition, scope, and interpretation and are presented as original contributions.

Authors' Contributions

All authors contributed to the analysis and interpretation of the results. Study design was performed by MS, NP, and YAL. Machine learning analyses were conducted by MS. N-of-1 pathway transformation was completed by LSW. Figure contributions: Figure 1 was contributed by MS; Figure 2 was contributed by by MS, LSW, NP, and YAL; Figure 3 was contributed by MS, LSW, and YAL. Table contributions: Table 1 was contributed by NP and YAL; Tables 2 and 3 were contributed by MS and YAL.

All authors have seen and approved the manuscript and confirm that it has not been accepted or published elsewhere in its entirety.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Comparative analysis of the ML model and rationale for choosing RF.

[PDF File, 190 KB - bioinform_v6i1e80735_app1.pdf]

Multimedia Appendix 2

Learning curve and power analyses for generalizability in microcohort transcriptomics.

[PDF File, 227 KB - bioinform_v6i1e80735_app2.pdf]

Multimedia Appendix 3

Traceability of Random Forest–selected features to biological pathways and interpretive cautions.

[PDF File, 581 KB - bioinform_v6i1e80735_app3.pdf]

Multimedia Appendix 4

Cross-validation accuracies with CIs.

[PDF File, 129 KB - bioinform_v6i1e80735_app4.pdf]

Multimedia Appendix 5

Transcript-level feature reduction controls to evaluate dimensionality effects.

[PDF File, 42 KB - bioinform_v6i1e80735_app5.pdf]

References

1. Nguengang Wakap S, Lambert DM, Olry A, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. Eur J Hum Genet 2020 Feb;28(2):165-173. [doi: 10.1038/s41431-019-0508-0] [Medline: 31527858]
2. Zaim SR, Li Q, Schissler AG, Lussier YA. Emergence of pathway-level composite biomarkers from converging gene set signals of heterogeneous transcriptomic responses. Pac Symp Biocomput 2018;23:484-495. [Medline: 29218907]
3. Rachid Zaim S, Kenost C, Berghout J, Vitali F, Zhang HH, Lussier YA. Evaluating single-subject study methods for personal transcriptomic interpretations to advance precision medicine. BMC Med Genomics 2019 Jul 11;12(Suppl 5):96. [doi: 10.1186/s12920-019-0513-8] [Medline: 31296218]
4. Schissler AG, Gardeux V, Li Q, et al. Dynamic changes of RNA-sequencing expression for precision medicine: N-of-1-pathways Mahalanobis distance within pathways of single subjects predicts breast cancer survival. Bioinformatics 2015 Jun 15;31(12):i293-i302. [doi: 10.1093/bioinformatics/btv253] [Medline: 26072495]
5. Li Q, Schissler AG, Gardeux V, et al. N-of-1-pathways MixEnrich: advancing precision medicine via single-subject analysis in discovering dynamic changes of transcriptomes. BMC Med Genomics 2017 May 24;10(Suppl 1):27. [doi: 10.1186/s12920-017-0263-4] [Medline: 28589853]
6. Gardeux V, Achour I, Li J, et al. "N-of-1-pathways" unveils personal deregulated mechanisms from a single pair of RNA-Seq samples: towards precision medicine. J Am Med Inform Assoc 2014;21(6):1015-1025. [doi: 10.1136/amiajnl-2013-002519] [Medline: 25301808]
7. Li Q, Schissler AG, Gardeux V, et al. kMEn: Analyzing noisy and bidirectional transcriptional pathway responses in single subjects. J Biomed Inform 2017 Feb;66:32-41. [doi: 10.1016/j.jbi.2016.12.009] [Medline: 28007582]
8. Schork NJ. Personalized medicine: Time for one-person trials. Nature New Biol 2015 Apr 30;520(7549):609-611. [doi: 10.1038/520609a] [Medline: 25925459]
9. Alla S, Adari SK. What is MLOps? In: Beginning MLOps with MLFlow: Deploy Models in AWS SageMaker, Google Cloud, and Microsoft Azure: Berkeley, CA: Apress; 2020:79-124. [doi: 10.1007/978-1-4842-6549-9_3]

10. Kreuzberger D, Kühl N, Hirschl S. Machine learning operations (MLOps): overview, definition, and architecture. IEEE Access 2023;11:31866-31879. [doi: 10.1109/ACCESS.2023.3262138]

11. Rajagopal A, Ayanian S, Ryu AJ, et al. Machine learning operations in health care: a scoping review. Mayo Clin Proc Digit Health 2024 Sep;2(3):421-437. [doi: 10.1016/j.mcpdig.2024.06.009] [Medline: 40206123]

12. Singla A. Machine learning operations (MLOps): challenges and strategies. JKLST 2023;2(3):333-340. [doi: 10.60087/jklst.vol2.n3.p340]

13. Rella BPR. MLOps and DataOps integration for scalable machine learning deployment. Int J Multidiscip Res 2022:1-3 [FREE Full text]

14. Liang P, Song B, Zhan X, Chen Z, Yuan J. Automating the training and deployment of models in mlops by integrating systems with machine learning. arXiv. Preprint posted online on May 16, 2024. [doi: 10.48550/arXiv.2405.09819]

15. Zarour M, Alzabut H, Al-Sarayreh KT. MLOps best practices, challenges and maturity models: a systematic literature review. Inf Softw Technol 2025 Jul;183:107733. [doi: 10.1016/j.infsof.2025.107733]

16. Aberasturi D, Pouladi N, Zaim SR, et al. 'Single-subject studies'-derived analyses unveil altered biomechanisms between very small cohorts: implications for rare diseases. Bioinformatics 2021 Jul 12;37(Suppl_1):i67-i75. [doi: 10.1093/bioinformatics/btab290] [Medline: 34252934]

17. Gardeux V, Berghout J, Achour I, et al. A genome-by-environment interaction classifier for precision medicine: personal transcriptome response to rhinovirus identifies children prone to asthma exacerbations. J Am Med Inform Assoc 2017 Nov 1;24(6):1116-1126. [doi: 10.1093/jamia/ocx069] [Medline: 29016970]

18. Zaas AK, Chen M, Varkey J, et al. Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans. Cell Host Microbe 2009 Sep 17;6(3):207-217. [doi: 10.1016/j.chom.2009.07.006] [Medline: 19664979]

19. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature New Biol 2012 Oct;490(7418):61-70. [doi: 10.1038/nature11412]

20. Ciriello G, Gatza ML, Beck AH, et al. Comprehensive molecular portraits of invasive lobular breast cancer. Cell 2015 Oct 8;163(2):506-519. [doi: 10.1016/j.cell.2015.09.033] [Medline: 26451490]

21. Kim JY, Lee E, Park K, et al. Clinical implications of genomic profiles in metastatic breast cancer with a focus on TP53 and PIK3CA, the most frequently mutated genes. Oncotarget 2017 Apr 25;8(17):27997-28007. [doi: 10.18632/oncotarget.15881] [Medline: 28427202]

22. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. Nat Genet 2000 May;25(1):25-29. [doi: 10.1038/75556] [Medline: 10802651]

23. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 2010;11(3):20196867. [doi: 10.1186/gb-2010-11-3-r25] [Medline: 20196867]

24. Biewald L. Experiment tracking with weights and biases. 2020. URL: https://www.wandb.com [accessed 2025-10-02]

25. Louppe G, Wehenkel L, Sutera A, Geurts P. Understanding variable importances in forests of randomized trees. 2013 Presented at: Adv Neural Inf Process Syst p. 431-439 URL: https://papers.nips.cc/paper_files/paper/2013/hash/e3796ae838835da0b6f6ea37bcf8bcb7-Abstract.html [accessed 2025-10-02]

26. Spearman C. The proof and measurement of association between two things. Am J Psychol 1904 Jan;15(1):72. [doi: 10.2307/1412159]

27. Jaccard P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bull Soc Vaudoise Sci Nat 1901;37:547-579 [FREE Full text]

28. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157-1182 [FREE Full text]

29. Nogueira S, Sechidis K, Brown G. On the stability of feature selection algorithms. J Mach Learn 2018;18(174):1-54 [FREE Full text]

30. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. arXiv. Preprint posted online on May 22, 2017. [doi: 10.48550/arXiv.1705.07874]

31. Park Y, Hauschild AC, Heider D. Transfer learning compensates limited data, batch effects and technological heterogeneity in single-cell sequencing. NAR Genom Bioinform 2021 Dec;3(4):lqab104. [doi: 10.1093/nargab/lqab104] [Medline: 34805988]

32. Gore S, Meche B, Shao D, Ginnett B, Zhou K, Azad RK. DiseaseNet: a transfer learning approach to noncommunicable disease classification. BMC Bioinformatics 2024 Mar 11;25(1):107. [doi: 10.1186/s12859-024-05734-5] [Medline: 38468193]

33. Shabanian M, Pouladi N, Wilson LS, Prosperi M, Lussier YA. Enabling transcriptome classification in micro-cohorts with pathway-anchoring and single-subject studies. Presented at: International Conference on Artificial Intelligence in Medicine p. 367-372. [doi: 10.1007/978-3-031-95841-0_68]

34. Zou X, Liu Y, Wang M, et al. scCURE identifies cell types responding to immunotherapy and enables outcome prediction. Cell Rep Methods 2023 Nov 20;3(11):100643. [doi: 10.1016/j.crmeth.2023.100643] [Medline: 37989083]

35. Kim J, Kim H, Lee MS, et al. Transcriptomes of the tumor-adjacent normal tissues are more informative than tumors in predicting recurrence in colorectal cancer patients. J Transl Med 2023 Mar 21;21(1):209. [doi: 10.1186/s12967-023-04053-2] [Medline: 36941605]

36. Rodon J, Soria JC, Berger R, et al. Genomic and transcriptomic profiling expands precision cancer medicine: the WINTHER trial. Nat Med 2019 May;25(5):751-758. [doi: 10.1038/s41591-019-0424-4] [Medline: 31011205]

37. Von Hoff DD, Stephenson JJ, Rosen P, et al. Pilot study using molecular profiling of patients' tumors to find potential targets and select treatments for their refractory cancers. JCO 2010 Nov 20;28(33):4877-4883. [doi: 10.1200/JCO.2009.26.5983]

38. Lee SC, Xu X, Chng WJ, et al. Post-treatment tumor gene expression signatures are more predictive of treatment outcomes than baseline signatures in breast cancer. Pharmacogenet Genomics 2009 Nov;19(11):833-842. [doi: 10.1097/FPC.0b013e328330a39f] [Medline: 19809382]

Abbreviations

AIME: Artificial Intelligence in Medicine
AMIA: American Medical Informatics Association
BC: Breast Cancer dataset
bRNAseq: Bulk RNA sequencing
CV: cross-validation
FC: fold change
GO: gene ontology
HRV: human rhinovirus
LLMs: large language models
MDI: mean decrease impurity
ML: Machine Learning
MLOps: Machine Learning Operations
mRNA: Messenger RNA
PBMCs: peripheral blood mononuclear cells
PIK3CA: phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha
RF: Random Forest
RMA: Robust Multi-array Average (normalization method)
RNA-seq: RNA sequencing
SVM: support vector machine
TCGA: The Cancer Genome Atlas
TCGA-BRCA: TCGA Breast Invasive Carcinoma project code
TMM: trimmed mean of M values
TP53: tumor protein p53
W&B: Weights & Biases MLOps
XGBoost: eXtreme Gradient Boosting

Lung Cancer Diagnosis From Computed Tomography Images Using Deep Learning Algorithms With Random Pixel Swap Data Augmentation: Algorithm Development and Validation Study

Ayomide Adeyemi Abe, PhD; Mpumelelo Nyathi, PhD

Department of Medical Physics, Sefako Makgatho Health Science University, Molotlegi St, Zone 1, Garankuwa, Pretoria, South Africa

Corresponding Author:

Ayomide Adeyemi Abe, PhD

Department of Medical Physics, Sefako Makgatho Health Science University, Molotlegi St, Zone 1, Garankuwa, Pretoria, South Africa

Abstract

Background: Deep learning (DL) shows promise for automated lung cancer diagnosis, but limited clinical data can restrict performance. While data augmentation (DA) helps, existing methods struggle with chest computed tomography (CT) scans across diverse DL architectures.

Objective: This study proposes Random Pixel Swap (RPS), a novel DA technique, to enhance diagnostic performance in both convolutional neural networks and transformers for lung cancer diagnosis from CT scan images.

Methods: RPS generates augmented data by randomly swapping pixels within patient CT scan images. We evaluated it on ResNet, MobileNet, Vision Transformer, and Swin Transformer models, using 2 public CT datasets (Iraq-Oncology Teaching Hospital/National Center for Cancer Diseases [IQ-OTH/NCCD] dataset and chest CT scan images dataset), and measured accuracy and area under the receiver operating characteristic curve (AUROC). Statistical significance was assessed via paired t tests.

Results: The RPS outperformed state-of-the-art DA methods (Cutout, Random Erasing, MixUp, and CutMix), achieving 97.56% accuracy and 98.61% AUROC on the IQ-OTH/NCCD dataset and 97.78% accuracy and 99.46% AUROC on the chest CT scan images dataset. While traditional augmentation approaches (flipping and rotation) remained effective, RPS complemented them, surpassing the performance findings in prior studies and demonstrating the potential of artificial intelligence for early lung cancer detection.

Conclusions: The RPS technique enhances convolutional neural network and transformer models, enabling more accurate automated lung cancer detection from CT scan images.

KEYWORDS

Introduction

Background

Lung cancer is a lethal disease characterized by uncontrolled cell growth in the lungs [1]. These malignant cells can proliferate, invade nearby tissues, and metastasize to other parts of the body [2]. The disease progresses through distinct stages, with advanced stages often proving fatal [3]. Lung cancer comprises multiple histological types and subtypes, affecting individuals regardless of gender [4]. Globally, lung cancer remains the leading cause of cancer-related mortality [5]. In 2020 alone, it accounted for 1.8 million deaths, ranking as the 6th leading cause of death worldwide among individuals younger than 70 years [2]. A key contributor to this high mortality is the frequent absence of early symptoms, leading to late-stage diagnosis and poorer outcomes [6]. The 5-year survival rate for lung cancer patients remains low, emphasizing the critical need for early detection [7]. Early diagnosis significantly improves prognosis, reduces long-term treatment costs, expands therapeutic options, and alleviates the burden on caregivers and families [1,8-10]. However, most cases are still detected at advanced stages, drastically limiting survival rates [5]. These challenges underscore lung cancer as a major public health priority.

Computed tomography (CT) is a medical imaging technique that produces high-resolution cross-sectional images of the lungs, providing detailed anatomical information for clinical evaluation [11]. As a noninvasive diagnostic tool, CT imaging has become indispensable for the early detection of lung cancer, offering superior sensitivity compared to conventional radiography [12,13]. However, the interpretation of CT scans presents significant challenges in clinical practice. The process demands considerable expertise from radiologists, as subtle

early-stage malignancies may demonstrate imaging features that escape human detection, potentially leading to diagnostic oversights [14,15]. The subjective nature of image interpretation introduces variability in diagnostic accuracy among practitioners, which can result in false-positive identification of pulmonary nodules. Such errors may prompt unnecessary invasive procedures for confirmation, exposing patients to avoidable risks and health care systems to additional costs [13]. Furthermore, the comprehensive evaluation of CT examinations is particularly demanding, as each study comprises numerous sequential slices, requiring both individual assessment and integrated analysis. This labor-intensive process frequently overwhelms available radiological resources, contributing to diagnostic delays and extended patient waiting periods [15-17]. To address these limitations, computer-assisted diagnostic systems have been developed to augment radiologists' interpretive capabilities [18]. These automated solutions employ advanced algorithms to analyze CT images, enhancing diagnostic accuracy while improving workflow efficiency [19]. By integrating such technological advancements into clinical practice, health care providers can mitigate the current challenges associated with manual CT interpretation, ultimately improving patient outcomes through more timely and reliable diagnoses.

The application of computer algorithms for the automated early diagnosis of lung cancer from CT scan images has evolved considerably. Early approaches used radiomics and machine learning techniques, but recent advancements have established deep learning (DL) as the predominant methodology [20]. Unlike traditional methods that depend on manually engineered features, a process prone to bias and time constraints, DL employs artificial neural networks to autonomously extract sophisticated features through training [21]. Among DL architectures, both convolutional neural networks (CNNs) and Vision Transformers have demonstrated exceptional potential for the early detection of lung cancer [22]. CNNs gained prominence after 2012, while Vision Transformers emerged in 2020 [23], with both now leading innovations in automated CT scan analysis [18,19].

CNNs and transformers offer distinct advantages for medical image analysis. CNNs, with their inductive bias for spatial locality and translation invariance, benefit from a simpler, parameter-efficient architecture rooted in spatial priors, which is highly effective and easier to train on smaller datasets [24,25]. They specialize in extracting local features and understanding spatial relationships between adjacent pixels. In contrast, transformers excel at capturing long-range dependencies across the entire image [26]. Vision Transformers are particularly scalable, maintaining image resolution better than CNNs during processing [27]. Their parallel processing capability also enables faster training times compared to similarly complex CNNs [28], although they typically require larger training datasets to achieve comparable performance [29]. Recent developments have seen the rise of hybrid networks that combine CNN and transformer architectures, successfully integrating both local and global feature extraction to overcome the limitations of standalone approaches [30,31].

Despite their capabilities, DL models face significant data-related challenges. While these architectures proficiently automate nodule detection, classification, and segmentation in CT scans [32], they demand extensive training data to outperform radiologist interpretations [33]. The scarcity of annotated medical CT datasets presents a major constraint [34], as creating such datasets requires time-consuming, expert-driven image labeling [35]. Data augmentation (DA) has emerged as a crucial solution to expand dataset size and diversity [36], enhancing both the quantity and quality of available training samples [37]. However, selecting appropriate DA techniques for chest CT analysis remains challenging due to several factors, including the variable effectiveness of methods across different datasets and domains [38], potential label distortions and crucial information loss caused by certain transformations [39], and current limitations in improving performance for both CNN and transformer architectures [37,40]. To address these challenges, this study proposes the Random Pixel Swap (RPS) augmentation method, specifically designed to enhance the generalization capabilities of both architectural paradigms in lung cancer diagnosis from chest CT scan images.

Related Work

The effectiveness of DA in training large neural networks was first conclusively demonstrated in 2012 [41], sparking the development of numerous innovative techniques [37]. These methods primarily fall into 2 categories: data synthesis and data transformation [36]. Data synthesis techniques generate novel samples that maintain statistical similarity to the original training data, while data transformation techniques create variations by modifying existing training samples. Both approaches effectively increase training dataset size, quality, and diversity, although they differ significantly in implementation. Data synthesis typically requires parameter learning, a process that can prove computationally intensive and often demands substantial training data to achieve optimal results [42]. In contrast, data transformation techniques generally avoid parameter learning and consequently require less computational resources. Traditional data transformation methods include fundamental image manipulations such as flipping, rotation, cropping, translation, and photometric adjustments (modifications to brightness, saturation, contrast, and hue) [36]. More sophisticated approaches like Cutout [43], Random Erasing [44], MixUp [45], and CutMix [46] have subsequently emerged, achieving state-of-the-art performance across various domains. These advanced techniques have been employed in lung cancer diagnosis from CT scan images [47-49].

The following section provides a comprehensive examination of the Cutout, Random Erasing, MixUp, and CutMix techniques, analyzing their limitations in medical imaging applications and contrasting them with the proposed RPS method. This comparative analysis establishes the foundation rationale for developing specialized augmentation approaches optimized for medical image analysis challenges.

Cutout DA Technique

The Cutout technique randomly selects square regions within images and masks their pixel values [43]. While effective for improving model robustness against occlusions in natural

images, this approach presents significant limitations for medical CT scans. The method's potential to eliminate critical diagnostic information (such as cancerous regions) may degrade performance [38]. Additionally, the masking process can inadvertently alter image labels, further limiting effectiveness [39]. Unlike Cutout, our RPS approach avoids information loss. It preserves diagnostic information by replacing masked regions with pixel values that are derived from other areas within the same CT scan while maintaining original labels.

Random Erasing DA Technique

Random Erasing extends Cutout's functionality by supporting both square and rectangular masks of varying sizes [44]. This technique randomly selects image regions for erasure and replaces them with random pixel values. While the variable mask sizes increase dataset diversity compared to Cutout, the method still suffers from information loss and label alteration issues [36,40]. These limitations are particularly problematic for medical imaging, where preserving anatomical content is crucial.

MixUp DA Technique

MixUp generates new samples through linear interpolation of pixel values and labels from 2 distinct images [45]. This approach enhances model generalization by preventing label memorization and improving adversarial robustness. However, the technique's potential to blur important anatomical boundaries and the requirement of careful hyperparameter tuning can create a bottleneck in medical contexts [47,48]. Furthermore, its convergence speed is often suboptimal [47]. RPS addresses these limitations by operating within single patient scans rather than mixing data across patients and employs a single hyperparameter for more efficient training.

CutMix DA Technique

CutMix combines aspects of previous methods by cutting patches from one image and pasting them onto another while proportionally blending labels [46]. Although this approach leverages the benefits from both Cutout and MixUp, the label blending can introduce noise that degrades model performance [50]. For medical CT scans, combining patches from different patients may confuse learning models, particularly when dealing with subtle pathological features [51]. RPS overcomes these challenges by performing pixel swaps exclusively within individual patient scans and preserving original labels without blending. Figure 1 visually contrasts these techniques with the proposed RPS method.

Figure 1. Computed tomography images for various data augmentation techniques. (A) MixUp; (B) CutMix; (C) Cutout; (D) Random Erasing; (E) Random Pixel Swap. The original image is in column 1, while the augmented images are in columns 2, 3, 4, and 5.
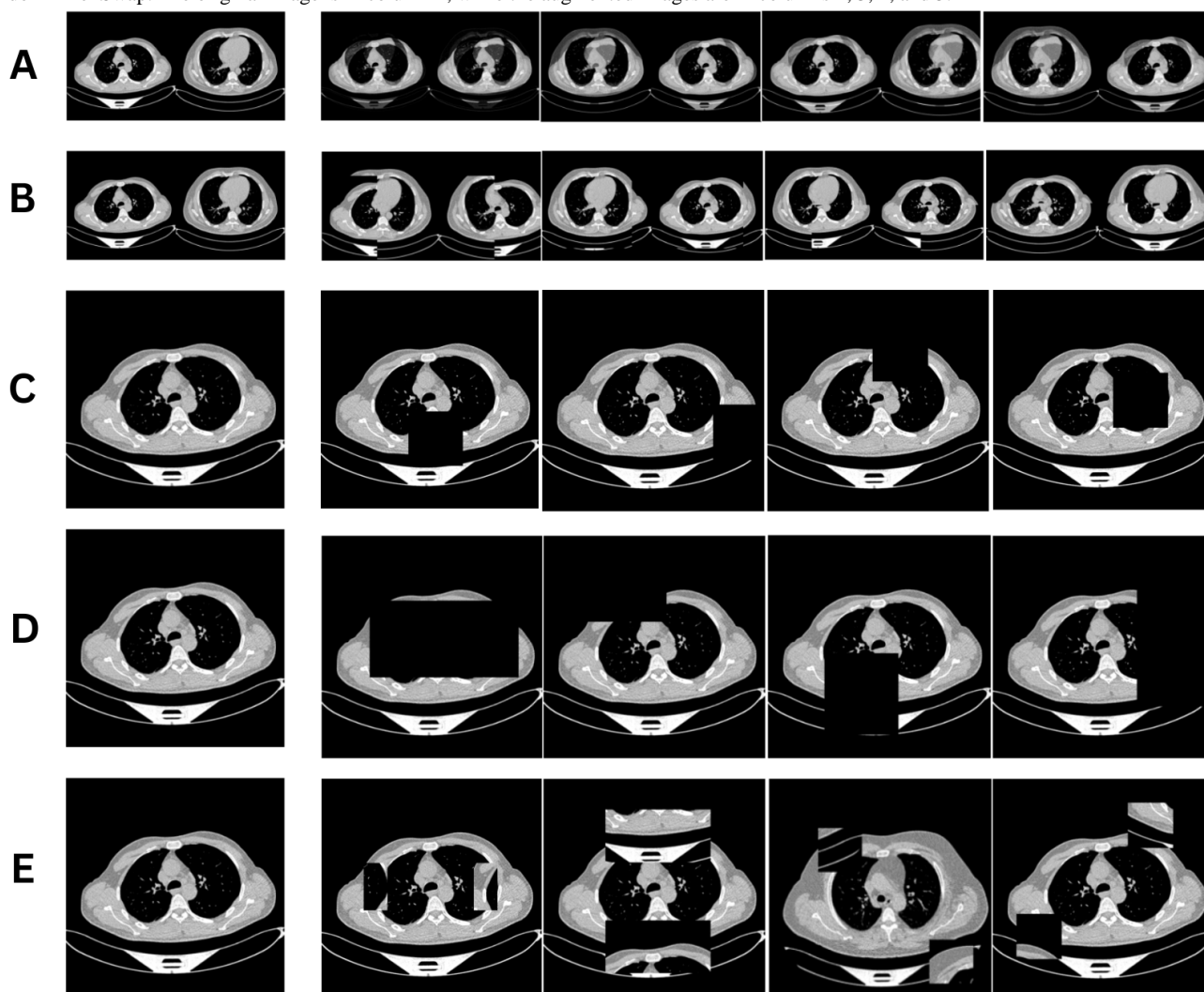
Methods

RPS DA Technique

The RPS technique is a parameter-free DA algorithm that operates with a predefined transformation probability. This method partitions input images into 2 distinct regions that serve as source and target areas for patch selection and swapping operations. The study proposes 4 specific implementation approaches, designated as RPS_H (vertical), RPS_W (horizontal), RPS_U (upper right diagonal), and RPS_D (upper left diagonal) swap configurations, as illustrated in Figure 2. This multidirectional swapping mechanism provides several advantages: it generates diverse transformations within individual patient CT scans while maintaining pathological plausibility, introduces meaningful variability in the training dataset without requiring parameter learning, and preserves all critical diagnostic information by operating exclusively within each scan's original pixel values. The technique's ability to produce multiple distinct transformations from a single image significantly enhances dataset diversity while avoiding the label alteration and information loss issues associated with other augmentation methods.

Figure 2. Four possible swap approaches for the Random Pixel Swap (RPS) data augmentation technique. I is the original image. Areas A_1 and A_2 are the swap regions. RPS_H (vertical), RPS_W (horizontal), RPS_U (upper right diagonal), and RPS_D (upper left diagonal) are the possible swap configurations.
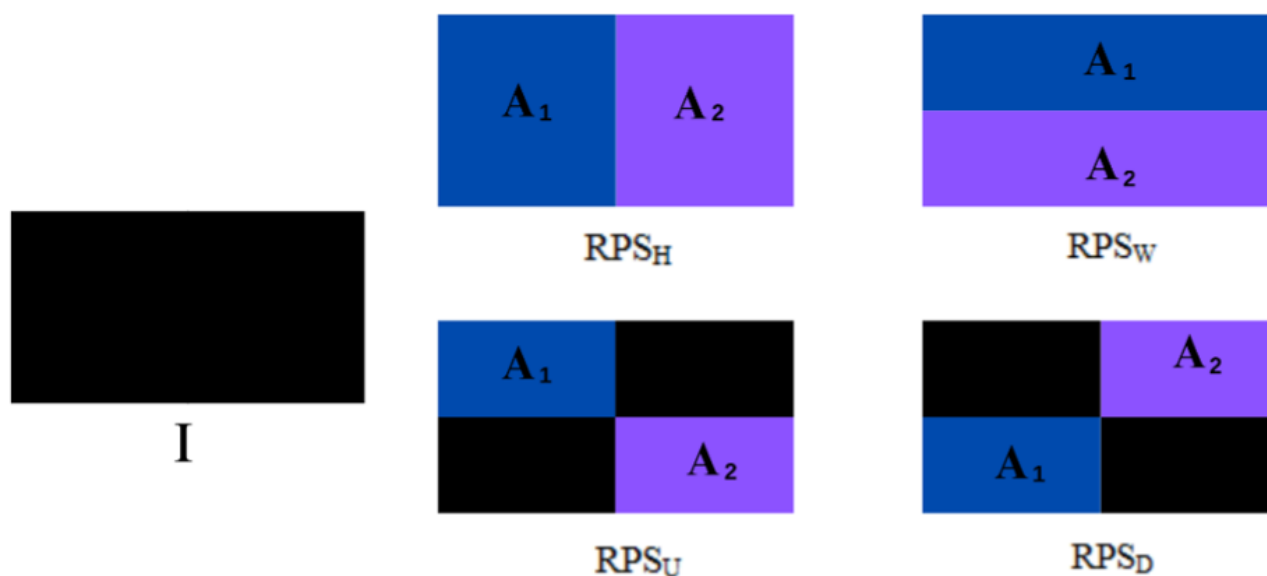
RPS possesses distinct invariant properties compared to other techniques. For an image with N pixels and Li intensity levels, the RPS transformation preserves global intensity, as shown in Equations (1)-(3). The technique employs a controlled, systematic, random patch-based pixel swap, rather than a random point-based pixel swap, ensuring that image content is preserved. This approach generates meaningful variations while maintaining pathological truth, thereby retaining clinical relevance in the context of lung cancer diagnosis.

$$(1) X' = T(X)$$

where T is permutation transform

$$(2) p = \frac{n(i)}{N} = \frac{n(i)'}{N'}; i = 0,1,2,...,L-1$$

$$(3) Ig = \frac{\sum_{i=0}^{L-1} i}{N} = \frac{\sum_{i=0}^{L-1} i'}{N'}$$

where *P* is the probability of a pixel having intensity *i*; $n(i), n(i)'$ is the number of pixels with intensity level $i \in X \wedge X'$, respectively; N, N' is the total number of pixels in $X \wedge X'$, respectively; L is the intensity level; and Ig is the average global intensity.

Implementation of RPS

The RPS technique is implemented by first randomly selecting 2 coordinate points (x_1, x_2) along the x-axis and 2 points (y_1, y_2) along the y-axis within the input image. These coordinates define 2 equal subswap regions: region X bounded by swap area $A1:(x_1, y_1)$ and (x_2, y_2), and As2 bounded by swap area $A2:(x_1, y_1)'$ and $(x_2, y_2)'$. The method incorporates a key hyperparameter called the swap area factor Sf, which ranges from 0.1 to 1.0, to control the extent of augmentation. The actual swap areas Sa1 and Sa2 are derived by scaling the subswap regions using this factor, as specified in Equations (4) and (5). During the augmentation process, the contents of swap area Sa1 are cropped and pasted into swap area Sa2 while simultaneously transferring the contents of swap area Sa2 to swap area Sa1. This bidirectional swapping ensures comprehensive data transformation while preserving all original image information. The complete RPS procedure is formally described in Textbox 1.

$$(4) Sa1 = As1 * Sf$$

$$(5) Sa2 = As2 * Sf$$

Textbox 1. Algorithm 1: Random Pixel Swap data augmentation procedure.

Input: data X; with shape H×W

Output: Augmented data X∗

1: A1∈ 12(H∗W)

2: **Init:** All points P within A1

3: Sf ← Sf   : Sf [0.1, 1.0]

4: **for** Pi,Pj P,do

5: Randomly select Pi , Pj ,

Pi′ ,Pj′ = Pi∗2 , Pj∗2

6: As1 = Area (Pi , Pj)

7: As2 = Area (Pi′ , Pj′)

8: Sa1 = As1 * Sf

9: Sa2 = As2 * Sf

10: X* ←Replace Sa1 with Sa2 in X and Sa2 with Sa1 in X

11: **end for**

12: **return** X∗

Swap Area Factor

The swap area factor Sf is a crucial parameter in the RPS technique, representing the ratio between the subswap region and the total swap area as described in Equation (6). This factor plays a vital role in the augmentation process for two key reasons: (1) it allows customization for different DL architectures that may benefit from varying swap region sizes, and (2) it helps maintain clinical relevance by limiting distortion of diagnostically important anatomical features. The study proposes two distinct implementations of this parameter: (1) single-value swap area factor (SVSF), which applies a fixed value throughout the augmentation process, and (2) multivalue swap area factor (MVSF), which uses multiple values to generate more diverse swap areas. In both implementations, the swap area factor operates within a defined range of 0.1 to 1.0, providing controlled flexibility for different medical imaging scenarios.

$$(6) Sf = AsSa$$

Experimental Validation of the RPS Technique

We conducted comprehensive experiments to validate the effectiveness of the proposed RPS technique in enhancing DL model performance across both CNN and transformer architectures. For our evaluation, we selected 4 established models: ResNet-34 [52], MobileNetV3 (small variant) [53], Vision Transformer (base-16) [23], and Swin Transformer (tiny version) [29], all initialized with preactivated weights. These architectures were chosen based on three key criteria: (1) public availability for reproducible benchmarking, (2) widespread adoption in methodological comparisons [29,48], and (3) efficient training characteristics due to their relatively fewer trainable parameters compared to larger variants.

Our experimental design incorporated three key comparisons: (1) models trained without any augmentation, (2) models trained with RPS augmentation, and (3) models trained with 4 state-of-the-art DA techniques (Cutout [43], Random Erasing [44], MixUp [45], and CutMix [46]). These comparison techniques were selected because they represent current best practices in parameter-free augmentation methods that share conceptual similarities with RPS [48]. We evaluated all models using two key metrics: (1) classification accuracy and (2) area under the receiver operating characteristic curve (AUROC), providing a comprehensive assessment of both overall performance and diagnostic discrimination capability.

Experimental Setup and Implementation

All experiments were conducted using Python 3.12.2 (Python Software Foundation) and PyTorch 2.2.2+ cu118 (PyTorch Foundation) within Jupyter Notebook 7.0.8 (IPython Project), running on an NVIDIA Quadro RTX 3000 GPU (Nvidia Corporation). We adopted the AdamW optimizer with a cross-entropy loss function, using a batch size of 16. The StepLR scheduler was configured with a step size of 10 and a gamma value of 0.5 [52]. Models were trained for 50 epochs, as additional training resulted in overfitting and performance degradation. After evaluating various learning rates, we selected 1×10^{4} as it yielded optimal results. Image normalization was applied with mean and SD values of 0.5 to enhance training stability and accelerate convergence [53].

For RPS implementation, we used a swap area factor of 1.0 with an augmentation probability of 1.0 for all experiments. CNN models processed images at 512×512 and 224×224 resolutions, while transformer architectures used 224×224 resolution due to the Vision Transformer's input size limitations. Although the Swin Transformer supports 512×512 inputs, we maintained a consistent 224×224 resolution across all transformer experiments for fair comparison. All experiments were conducted with a random seed of 42 after verifying consistent performance patterns across 3 different seeds.

Statistical Analysis

To evaluate our hypothesis that an effective DA technique should perform consistently across both CNN and transformer architectures, we treated each technique as an independent variable and considered model performance as the dependent variable. We used paired sample *t* tests [54] to assess significant differences between techniques, considering *P* values <.05 as statistically significant.

For comprehensive technique comparison, we implemented a ranking system based on cumulative scores C (Equations (7) and (8)), where higher scores received lower rank numbers R. This approach enabled holistic performance benchmarking across all models and architectures.

$$(7) \quad C = \sum_{model=1}^{n} model(A + AUROC)$$

$$(8) \quad R_1, R_2, R_3, \ldots, R_{m+1} = C_1, C_2, C_3, \ldots, C_{m+1} \lor C_1 > C_2 > C_3, \ldots, > C_{m+1}$$

where **C** is cumulative score, **R** is rank, **m** is the total number of data augmentation techniques, **n** is the total number of selected models, **A** is accuracy, and AUROC is the area under the receiver operating characteristic curve.

Iraq-Oncology Teaching Hospital/National Center for Cancer Diseases Dataset

The Iraq-Oncology Teaching Hospital/National Center for Cancer Diseases (IQ-OTH/NCCD) dataset contains 1097 JPEG CT images collected from 110 patients [35]. These images were obtained using a SOMATOM Siemens scanner (Siemens Healthineers) and encompass a diverse range of demographic characteristics. The dataset is organized into 3 categories: normal scans, benign tumor scans, and malignant tumor scans. Specifically, it includes 15 cases of benign tumors, totaling 120 images; 40 cases of malignant tumors, totaling 416 images; and 55 cases of normal findings, totaling 561 images. Each image has a resolution of 512×512 pixels. We divided the images in a ratio of 7:3 for training and testing.

Chest CT Scan Images Dataset

The chest CT scan images dataset contains 1000 lung CT scans from patients diagnosed with 3 different types of lung cancers, as well as scans from healthy individuals, all in JPG format [55]. The lung cancer types included in the dataset are adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. The images are organized into training, testing, and validation sets for each lung cancer category.

Ethical Considerations

Ethics approval was obtained from the Sefako Makgatho University Research Committee (ethics reference number: SMUREC/M/12/2022:PG).

Results

Average Training Time Overhead

To evaluate the computational impact of the RPS technique, we measured training duration for 4 architectures (ResNet-34, MobileNetV3 [small], Vision Transformer [base-16], and Swin Transformer [tiny]) with and without RPS implementation. The training time overhead was calculated as the difference between augmented and nonaugmented training times. Experiments were conducted on both the IQ-OTH/NCCD and chest CT scan datasets using 224×224 image resolution, with results averaged across 3 independent runs for reliability.

Our analysis included a comparative assessment of 4 established DA techniques: Cutout, Random Erasing, MixUp, and CutMix. Results demonstrated that while RPS increased training times across all models compared to nonaugmented training, this increase was not statistically significant (*P*=.07). Similarly, comparisons between RPS and other DA techniques revealed no statistically significant differences in computational overhead (Cutout: *P*=.06; Random Erasing: *P*=.17; MixUp: *P*=.49; CutMix: *P*=.16). Among all evaluated methods, RPS showed the highest training time overhead, followed sequentially by MixUp, CutMix, Random Erasing, and Cutout. Complete results are presented in Figure 3.

Figure 3. Average training time overhead of 5 data augmentation techniques across 4 deep learning models.
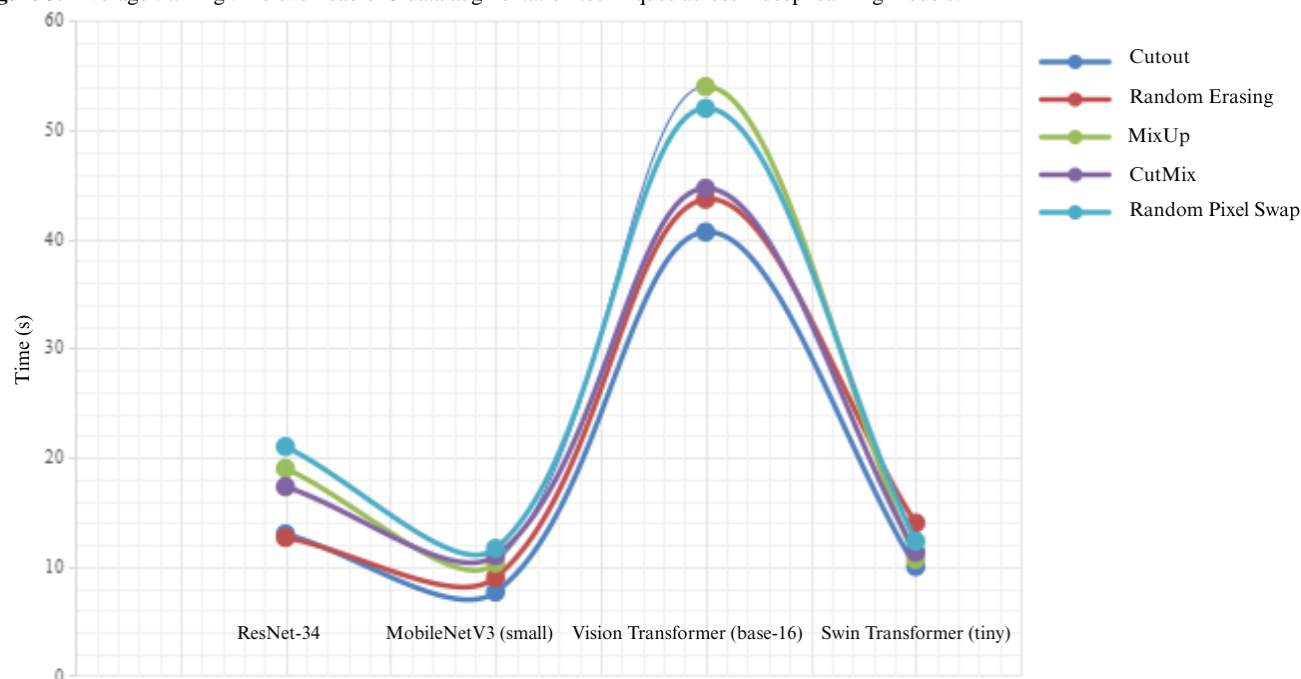
Performance Comparison of RPS With State-of-the-Art DA Techniques for Lung Cancer Detection

To evaluate pulmonary nodule detection in chest CT scan images, the selected CNN and transformer models (ResNet-34, MobileNetV3 [small], Vision Transformer [base-16], and Swin Transformer [tiny]) were trained on the IQ-OTH/NCCD dataset to classify the scan images as normal or containing benign or malignant pulmonary nodules. Experimental results demonstrated that RPS significantly enhanced performance across all 4 architectures (P=.008). The MobileNetV3 model achieved particular success when combined with RPS using 512×512 image resolution, reaching a peak classification accuracy of 94.21%, representing a 1.22% accuracy improvement and 0.86% AUROC increase over the baseline model.

At 224×224 image resolution, our comprehensive comparison of RPS against the 4 established DA methods (Cutout: P=.03; Random Erasing: P=.008; MixUp: P=.02; CutMix: P=.02) revealed consistent superiority of the RPS technique (P<.05). For ResNet-34, RPS exceeded CutMix (the best alternative) by 2.44% and Random Erasing (the least effective) by 5.49% in accuracy. MobileNetV3 showed a 0.3% improvement over Cutout (best alternative) and 1.83% over MixUp (least effective) in accuracy. Transformer architectures demonstrated even more pronounced benefits: Vision Transformer with RPS outperformed Random Erasing by 1.52% and MixUp by 16.77%, while Swin Transformer showed a 1.53% improvement over MixUp and 4.57% over Cutout in accuracy. Across all architectures, performance ranking was as follows: (1) RPS (best technique), (2) Random Erasing, (3) CutMix, (4) MixUp, and (5) Cutout. The detailed results are presented in Table 1.

Table . Classification results of the IQ-OTH/NCCD[a] dataset using preactivated deep learning models with various data augmentation techniques (224×224 image resolution).

Data augmentation	Rank[b]	ResNet-34		MobileNetV3 (small)		Vision Transformer (base-16)		Swin Transformer (tiny)	
		Accuracy, %	AUROC[c], %	Accuracy, %	AUROC, %	Accuracy, %	AUROC, %	Accuracy, %	AUROC, %
Base model[d]	6	85.98	83.39	86.59	93.16	57.62	64.88	85.67	89.06
Cutout[d]	5	85.67	86.11	89.33	92.95	57.01	63.45	85.37	93.68
Random Erasing[d]	2	82.62	91.23	88.72	90.10	71.65	75.41[e]	86.89	91.42
MixUp[d]	4	84.45	91.05	87.80	86.57	56.40	68.55	88.41	92.51
CutMix[d,f]	3	85.67	88.34	88.41	93.02	68.90	69.68	88.41	92.05
Random Pixel Swap[f]	1	88.11[e]	93.70[e]	89.63[e]	93.80[e]	73.17[e]	74.64	89.94[e]	94.79[e]

[a]IQ-OTH/NCCD: Iraq-Oncology Teaching Hospital/National Center for Cancer Diseases.

[b]Rank represents the overall rating for each technique, with "1" indicating the best technique across all models.

[c]AUROC: area under the receiver operating characteristic curve.

[d]Significant difference between an augmentation technique and the Random Pixel Swap technique across all models.

[e]Highest value in the column.

[f]Significant difference between training using an augmentation technique and the base model across all models.

At 512×512 image resolution, ResNet-34 exhibited nuanced performance differences between augmentation techniques: while CutMix achieved a marginal 0.31% higher accuracy than RPS, RPS demonstrated significantly superior diagnostic capability with a 5.31% improvement in AUROC. Furthermore, RPS outperformed the least effective technique (Random Erasing) by 2.13% in accuracy and 3.17% in AUROC. For MobileNetV3, RPS dominated all comparative techniques in both accuracy and AUROC, except for a 1.23% AUROC advantage by CutMix. Specifically, RPS exceeded Cutout (the best alternative technique) by 0.61% and surpassed MixUp (the least effective) by 4.58% in accuracy. Across all evaluated methods, the overall performance ranking was as follows: (1) RPS (best technique), (2) Cutout, (3) CutMix, (4) MixUp, and (5) Random Erasing. The detailed results are presented in Table 2.

Table . Classification results of the IQ-OTH/NCCD[a] dataset using preactivated deep learning models with various data augmentation techniques (512×512 image resolution).

Data augmentation	Rank[b]	ResNet-34		MobileNetV3 (small)	
		Accuracy, %	AUROC[c], %	Accuracy, %	AUROC, %
Base model[d]	6	88.72	78.51	92.99	94.81
Cutout[d]	2	90.24	93.25	93.60	95.42
Random Erasing[d]	5	89.94	91.85	90.85	92.19
MixUp[d]	4	89.94	96.13[e]	89.63	95.18
CutMix	3	92.38[e]	89.71	92.68	96.90[e]
Random Pixel Swap	1	92.07	95.02	94.21[e]	95.67

[a]IQ-OTH/NCCD: Iraq-Oncology Teaching Hospital/National Center for Cancer Diseases.

[b]Rank represents the overall rating for each technique, with "1" indicating the best technique across all models.

[c]AUROC: area under the receiver operating characteristic curve.

[d]Significant difference between an augmentation technique and the Random Pixel Swap technique across all models.

[e]Highest value in the column.

Performance Comparison of RPS With State-of-the-Art DA Techniques for Lung Cancer

Classification From CT Scan Images Using DL Architectures

We evaluated the effectiveness of the RPS technique for lung cancer classification using the chest CT scan images dataset

across multiple DL architectures. The experimental results demonstrated that RPS significantly enhanced classification performance for all architectures (P=.008). RPS combined with ResNet-34 at 512×512 image resolution achieved optimal performance, reaching 97.78% accuracy and 99.46% AUROC.

At 224×224 image resolution, RPS consistently outperformed competing techniques across most models (Cutout: P=.001; Random Erasing: P=.02; MixUp: P=.047; CutMix: P=.18). For ResNet-34, RPS exceeded CutMix (the best alternative) by 0.64% and Random Erasing (the least effective) by 5.08% in accuracy. MobileNetV3 showed even greater improvements over other methods, with RPS surpassing CutMix by 3.49% and MixUp by 9.21% in accuracy. For the implementation with Vision Transformer, RPS surpassed Random Erasing (the best alternative) by 1.91% and MixUp (the least effective) by 18.85% in accuracy. While CutMix showed a 2.22% accuracy advantage over RPS for the Swin Transformer, RPS maintained superior performance against all other techniques, exceeding Cutout by 7.3% (the least effective). Across all architectures, the overall performance ranking was as follows: (1) RPS (best technique), (2) CutMix, (3) Random Erasing, (4) Cutout, and (5) MixUp. The detailed results are presented in Table 3.

Table . Classification results of the chest CT[a] scan images dataset using preactivated deep learning models with various data augmentation techniques (224×224 image resolution).

Data augmentation	Rank[b]	ResNet-34		MobileNetV3 (small)		Vision Transformer (base-16)		Swin Transformer (tiny)	
		Accuracy, %	AUROC[c], %	Accuracy, %	AUROC, %	Accuracy, %	AUROC, %	Accuracy, %	AUROC, %
Base model[d]	5	93.33	99.00	87.30	97.09	82.86	95.84	84.76	96.92
Cutout[d,e]	4	93.02	98.94	85.71	97.62	80.63	94.35	84.13	96.05
Random Erasing[d]	3	90.48	98.54	88.89	97.45	84.76	96.72[f]	88.25	97.28
MixUp[d]	6	91.43	98.57	83.49	96.85	67.82	86.97	90.79	97.87
CutMix	2	94.92	98.69	89.21	97.80	76.82	92.60	93.65[f]	98.74[f]
Random Pixel Swap[e]	1	95.56[f]	99.15[f]	92.70[f]	98.02[f]	86.67[f]	96.32	91.43	98.45

[a]CT: computed tomography.

[b]Rank represents the overall rating for each technique, with "1" indicating the best technique across all models.

[c]AUROC: area under the receiver operating characteristic curve.

[d]Significant difference between an augmentation technique and the Random Pixel Swap technique across all models.

[e]Significant difference between training using an augmentation technique and the base model across all models.

[f]Highest value in the column.

At 512×512 image resolution, the RPS technique demonstrated superior performance compared to all evaluated DA methods (Cutout: P=.13; Random Erasing: P=.27; MixUp: P=.13; CutMix: P=.31). For ResNet-34, RPS matched the accuracy of the top-performing alternative (CutMix) while achieving a 0.21% improvement in AUROC. Furthermore, RPS showed significant gains over the least effective technique (MixUp), with a 7.74% accuracy performance advantage. The MobileNetV3 architecture exhibited even more pronounced benefits, where RPS outperformed CutMix (the best alternative) by 2.23% and surpassed MixUp by 4.45% in accuracy. Across all techniques, the performance ranking was as follows: (1) RPS (best technique), (2) CutMix, (3) Cutout, (4) Random Erasing, and (5) MixUp. The detailed results are presented in Table 4.

Table . Classification results of the chest CT[a] scan images dataset using preactivated deep learning models with various data augmentation techniques (512×512 image resolution).

Data augmentation	Rank[b]	ResNet-34		MobileNetV3 (small)	
		Accuracy, %	AUROC[c], %	Accuracy, %	AUROC, %
Base model	5	96.83	99.25	93.02	98.27
Cutout	3	96.51	99.35	94.60	98.39
Random Erasing	4	96.83	99.42	93.65	98.82[d]
MixUp	6	92.38	98.64	92.38	98.51
CutMix	2	97.78[d]	99.25	94.60	98.61
Random Pixel Swap	1	97.78[d]	99.46[d]	96.83[d]	98.75

[a]CT: computed tomography.

[b]Rank represents the overall rating for each technique, with "1" indicating the best technique across all models.

[c]AUROC: area under the receiver operating characteristic curve.

[d]Highest value in the column.

Performance Analysis of Swap Area Factors for Lung Cancer Diagnosis

The swap area factor serves as a critical hyperparameter in RPS implementation. We systematically evaluated its influence using both SVSF and MVSF configurations across the 0.1 to 1.0 range on the IQ-OTH/NCCD dataset. MVSF provides over 100 possible combinations of lower and upper bounds (eg, 0.1‐0.5 and 0.4‐0.8); however, our experimental configurations maintained a fixed lower bound of 0.1. Experimental results revealed distinct optimal configurations for each architecture. For SVSF implementations, ResNet-34, Vision Transformer, and Swin Transformer achieved peak performance at 1.0, while MobileNetV3 performed best at 0.9. For MVSF implementations, ResNet-34 showed optimal results within 0.1‐0.9, MobileNetV3 performed best at 0.1‐0.7, Vision Transformer excelled at 0.1‐0.3, and Swin Transformer achieved peak performance at 0.1‐0.5.

Comparative analysis demonstrated that SVSF generally outperformed MVSF configurations for a fixed 0.1 lower bound across most architectures, with the notable exception of ResNet-34. For this model, MVSF (0.1‐0.9) surpassed SVSF (1.0) by 0.61% in accuracy and 1.08% in AUROC. The most effective overall configuration combined MobileNetV3 with RPS using an SVSF of 0.9, achieving 94.51% accuracy and 95.77% AUROC. The detailed results are presented in Table 5.

Table . Analysis of the IQ-OTH/NCCD[a] dataset using different deep learning architectures and Random Pixel Swap data augmentation with single-value and multivalue swap area factors (224×224 image resolution).

Swap factor	ResNet-34		MobileNetV3 (small)		Vision Transformer (base-16)		Swin Transformer (tiny)	
	Accuracy, %	AUROC[b], %	Accuracy, %	AUROC, %	Accuracy, %	AUROC, %	Accuracy, %	AUROC, %
Single value								
0.1	89.02	92.26	93.90	95.10	64.02	74.97[c]	86.28	92.02
0.2	91.16	94.62	92.99	95.04	60.98	63.20	83.23	93.59
0.3	90.85	93.48	92.07	95.38	64.02	67.95	89.02	91.63
0.4	89.63	92.66	92.99	95.24	69.21	72.59	84.76	92.81
0.5	90.55	90.96	92.68	95.01	68.60	72.47	87.80	84.13
0.6	90.55	94.76	92.99	95.24	69.21	72.36	84.76	89.63
0.7	91.46	95.23	92.68	95.12	67.99	66.48	83.54	90.65
0.8	89.63	92.22	93.60	95.60	67.99	69.83	89.94[c]	95.95[c]
0.9	90.85	94.42	94.51[c]	95.77[c]	71.65	72.79	83.84	90.99
1.0	92.07[c]	95.02[c]	94.21	95.67	73.17[c]	74.64	89.94[c]	94.79
Multivalue								
0.1 - 0.2	90.55	93.80	93.90[c]	94.83	66.16	69.14	85.98	94.29
0.1 - 0.3	89.33	90.53	93.29	95.18	72.56[c]	77.93[c]	84.76	91.84
0.1 - 0.4	89.33	90.18	93.60	95.03	60.98	73.15	87.50	94.30
0.1 - 0.5	91.16	95.93[c]	93.60	94.84	59.15	68.80	88.11[c]	94.94[c]
0.1 - 0.6	90.55	93.26	92.38	94.60	62.20	59.86	87.20	93.45
0.1 - 0.7	89.94	90.99	93.90[c]	95.33	61.28	70.73	88.11[c]	92.53
0.1 - 0.8	87.80	93.13	93.29	95.00	66.77	76.81	86.28	85.93
0.1 - 0.9	92.68[c]	95.29	93.60	95.29	68.29	64.98	86.59	92.69
0.1 - 1.0	89.02	92.53	93.60	95.71[c]	62.80	69.05	83.54	94.35

[a]IQ-OTH/NCCD: Iraq-Oncology Teaching Hospital/National Center for Cancer Diseases.

[b]AUROC: area under the receiver operating characteristic curve.

[c]Highest value in the column.

Our evaluation of the chest CT scan images dataset using different swap area factor configurations revealed architecture-specific optimal settings. SVSF demonstrated superior performance at 1.0 for both ResNet-34 and MobileNetV3, while Vision Transformer achieved peak accuracy with an SVSF of 0.1. For Swin Transformer, MVSF configurations between 0.1 and 0.6 yielded optimal results. Among all tested combinations, ResNet-34 paired with RPS using an SVSF of 1.0 delivered the highest classification performance, reaching 97.78% accuracy and 99.46% AUROC. The detailed results are presented in Table 6.

Table . Analysis of the chest CT[a] scan images dataset using different deep learning architectures and Random Pixel Swap data augmentation with single-value and multivalue swap area factors (224×224 image resolution).

Swap factor	ResNet-34		MobileNetV3 (small)		Vision Transformer (base-16)		Swin Transformer (tiny)	
	Accuracy, %	AUROC[b], %	Accuracy, %	AUROC, %	Accuracy, %	AUROC, %	Accuracy, %	AUROC, %
Single value								
0.1	96.19	99.13	94.60	98.55	86.67[c]	96.32	94.29[c]	98.72
0.2	97.46	99.27	94.60	98.60	81.27	94.37	92.06	98.78[c]
0.3	96.19	99.22	95.24	98.61	78.73	93.98	93.65	98.65
0.4	97.46	99.20	95.24	98.65	82.54	96.10	92.06	98.46
0.5	97.14	99.41	94.92	98.74	85.08	96.31	91.43	98.38
0.6	97.14	99.30	95.56	98.79[c]	85.40	96.86[c]	91.75	97.85
0.7	97.14	99.19	95.56	98.69	83.81	95.48	93.65	98.65
0.8	97.14	99.38	95.87	98.75	81.59	95.48	91.43	97.95
0.9	96.83	99.35	95.87	98.62	81.27	94.25	88.89	97.80
1.0	97.78[c]	99.46[c]	96.83[c]	98.75	75.56	91.62	91.43	98.45
Multivalue								
0.1 - 0.2	97.14	99.27	94.92	98.59	75.56	92.85	93.65	98.51
0.1 - 0.3	96.51	99.30	93.97	98.62	84.13	95.86	91.43	98.20
0.1 - 0.4	96.51	99.00	94.60	98.55	76.83	92.77	93.65	98.73
0.1 - 0.5	97.46	99.28	94.29	98.63	80.32	94.69	92.38	98.64
0.1 - 0.6	96.51	99.37	95.24	98.59	82.54	95.63	96.19[c]	98.90[c]
0.1 - 0.7	97.78[c]	99.39	94.92	98.65	86.03[c]	96.84[c]	94.29	98.90[c]
0.1 - 0.8	97.46	99.25	93.97	98.62	81.90	94.88	93.02	98.83
0.1 - 0.9	97.48	99.32	94.92	98.70	81.90	94.99	93.65	98.86
0.1 - 1.0	97.46	99.41[c]	95.87[c]	98.75[c]	82.22	95.21	93.33	98.72

[a]CT: computed tomography.

[b]AUROC: area under the receiver operating characteristic curve.

[c]Highest value in the column.

RPS With Lung Region of Interest Segmentation

Prior studies have demonstrated that segmenting lung regions of interest (ROIs) can significantly improve the diagnostic performance of DL models [33,56]. To evaluate the effectiveness of the RPS technique when applied to segmented images, we conducted experiments using the selected models (ResNet-34, MobileNetV3 [small], Vision Transformer [base-16], and Swin Transformer [tiny]). Our investigation used both the IQ-OTH/NCCD dataset and chest CT scan images dataset at 224×224 resolution.

The segmentation process involved multiple steps. We first applied a threshold algorithm to generate a lung mask, followed by dilation and hole-filling operations to ensure comprehensive coverage of pulmonary structures. The final lung ROI was extracted by cropping surrounding pixels along the mask boundaries. The complete procedure is illustrated in Figure 4. For comparative analysis, we evaluated model performance under three conditions: (1) training without augmentation, (2) training with RPS, and (3) training with established augmentation techniques (Cutout, Random Erasing, MixUp, and CutMix). This comprehensive evaluation framework allowed us to assess the relative benefits of RPS when applied to segmented lung images.

XSL•FO

RenderX

Figure 4. Lung segmentation procedure.
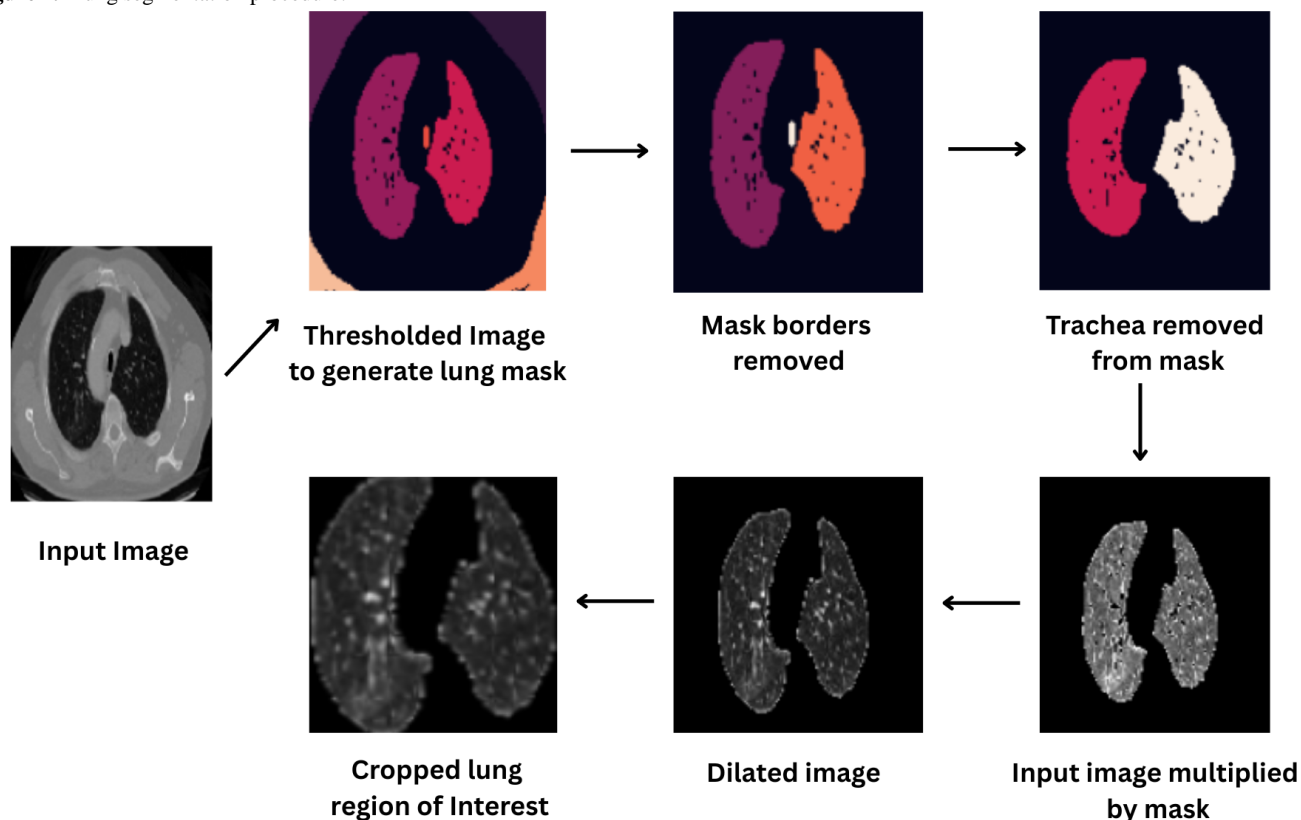
Our experiments with the IQ-OTH/NCCD dataset demonstrated that the RPS technique significantly improved performance across all evaluated models (P=.04) and most techniques (Cutout: P=.049; Random Erasing: P=.004; MixUp: P=.04; CutMix: P=.06). The most notable results were achieved by ResNet-34 with RPS, reaching 97.56% accuracy and 98.61% AUROC. While RPS outperformed all competing techniques for MobileNetV3 and Swin Transformer, CutMix showed superior performance for Vision Transformer, exceeding RPS by 1.52% in accuracy and 0.67% in AUROC. The overall performance ranking across techniques was as follows: (1) RPS (best technique), (2) CutMix, (3) Random Erasing, (4) Cutout, and (5) MixUp.

For the chest CT scan images dataset, the RPS technique consistently improved performance across models (P=.06) and most techniques (Cutout: P=.01; Random Erasing: P=.009; MixUp: P=.01; CutMix: P=.38). The highest performance was again achieved by ResNet-34 with RPS (95.51% accuracy and 98.86% AUROC). While RPS showed superior results for MobileNetV3 and Swin Transformer, CutMix performed better for Vision Transformer (3.21% higher accuracy and 0.6% higher AUROC). The comprehensive performance ranking was similar to that for the IQ-OTH/NCCD dataset and was as follows: (1) RPS, (2) CutMix, (3) Cutout, (4) Random Erasing, and (5) MixUp. The detailed results are presented in Table 7.

Table . Classification results of the IQ-OTH/NCCD[a] and chest CT[b] scan images datasets using preactivated deep learning models with various data augmentation techniques and segmentation of the lung region of interest (224×224 image resolution).

Data augmentation	Rank[c]	ResNet-34		MobileNetV3 (small)		Vision Transformer (base-16)		Swin Transformer (tiny)	
		Accuracy, %	AUROC[d], %	Accuracy, %	AUROC, %	Accuracy, %	AUROC, %	Accuracy, %	AUROC, %
IQ-OTH/NCCD dataset									
Base model[e]	5	96.65	99.13	95.43	97.28	89.94	96.51	93.60	98.21
Cutout[e]	4	96.04	98.86	95.43	96.33	92.38	96.20	93.90	97.80
Random Erasing[e]	3	95.73	97.45	96.65[f]	97.29	91.46	96.27	94.82[f]	98.00
MixUp[e,g]	6	95.87	99.19[f]	91.77	97.11	91.77	96.27	93.29	97.52
CutMix	2	96.65	98.86	94.51	96.39	93.90[f]	97.64[f]	93.29	97.52
Random Pixel Swap[g]	1	97.56[f]	98.61	96.65[f]	98.00[f]	92.38	96.97	94.82[f]	98.12[f]
Chest CT scan images dataset									
Base model	2	95.19	99.03	87.82	96.83[f]	82.69	95.48	90.71	98.11
Cutout[e]	4	94.55	98.85	88.14	97.66	80.77	93.86	88.14	97.32
Random Erasing[e,g]	5	94.55	98.75	86.54	96.52	79.81	89.72	86.86	97.16
MixUp[e,g]	6	94.55	98.77	82.05	95.33	78.85	93.29	85.90	97.10
CutMix	3	95.19	99.05[f]	86.54	96.89	86.86[f]	96.43[f]	87.82	96.73
Random Pixel Swap	1	95.51[f]	98.86	90.71[f]	97.51	83.65	95.83	91.35[f]	98.36[f]

[a]IQ-OTH/NCCD: Iraq-Oncology Teaching Hospital/National Center for Cancer Diseases.

[b]CT: computed tomography.

[c]Rank represents the overall rating for each technique, with "1" indicating the best technique across all models.

[d]AUROC: area under the receiver operating characteristic curve.

[e]Significant difference between an augmentation technique and the Random Pixel Swap technique across all models.

[f]Highest value in the column.

[g]Significant difference between training using an augmentation technique and the base model across all models.

Performance Analysis of the Combination of RPS With Traditional DA Techniques for Lung Cancer Diagnosis

Traditionally, DA techniques, including image flipping and rotation, are widely employed in medical image analysis with DL [44]. To evaluate the potential benefits of combining these methods with the RPS technique, we conducted a systematic comparison. First, we trained selected models (ResNet-34, MobileNetV3 [small], Vision Transformer [base-16], and Swin Transformer [tiny]) using individual traditional techniques: horizontal flipping, vertical flipping, and random rotation (±90°). Subsequently, we trained the models using combinations of each traditional technique with RPS.

Our experiments revealed that the combination of RPS with traditional techniques generally enhanced model performance compared to using traditional methods alone. However, when a traditional technique failed to improve baseline performance, its combination with RPS did not surpass RPS alone. For the IQ-OTH/NCCD dataset, using RPS alone surpassed the individual traditional techniques (horizontal flipping: $P=.63$; vertical flipping: $P=.22$; rotation: $P=.93$). RPS with rotation achieved peak performance for ResNet-34 and Vision Transformer (base-16), improving upon rotation alone by 2.14% and 2.75% in accuracy, respectively. RPS with vertical flipping performed the best for MobileNetV3 (small), exceeding vertical flipping alone by 0.61% in accuracy. However, RPS alone showed superior results for Swin Transformer (tiny).

Similarly, for the chest CT scan images dataset, using RPS alone surpassed the individual traditional techniques (horizontal flipping: $P=.01$; vertical flipping: $P=.03$; rotation: $P=.04$). RPS with rotation demonstrated the strongest overall performance, improving upon rotation by 0.95% in accuracy. RPS with horizontal flipping achieved optimal results for Vision Transformer (base-16), surpassing horizontal flipping alone by 5.71% in accuracy. However, RPS alone outperformed all

combinations for MobileNetV3 (small) and Swin Transformer   (tiny). The detailed results are presented in Table 8.

Table . Classification results of the IQ-OTH/NCCD[a] and chest CT[b] scan images datasets using preactivated deep learning models when 3 traditional data augmentation techniques are combined with the Random Pixel Swap data augmentation technique (224×224 image resolution).

Data augmentation	Rank[c]	ResNet-34		MobileNetV3 (small)		Vision Transformer (base-16)		Swin Transformer (tiny)	
		Accuracy, %	AUROC[d], %	Accuracy, %	AUROC, %	Accuracy, %	AUROC, %	Accuracy, %	AUROC, %
IQ-OTH/NCCD dataset									
Base model[e]	8	85.98	83.39	86.59	93.16	57.62	64.88	85.67	89.06
Horizontal flip	5	82.01	86.70	88.11	93.36	73.78	86.20	87.50	91.86
Horizontal flip with Random Pixel Swap	2	87.20	90.43	88.41	91.21	78.36	89.31[f]	88.72	92.12
Vertical flip[g]	7	87.50	89.82	90.55	93.89[f]	62.50	76.54	89.02	92.41
Vertical flip with Random Pixel Swap[e,g]	6	88.11	89.54	91.16[f]	93.28	68.29	73.94	88.72	93.10
Rotation[g]	4	87.80	90.30	89.63	91.53	75.91	80.70	88.41	92.67
Rotation with Random Pixel Swap[g]	1	89.94[f]	90.28	89.02	91.75	78.66[f]	87.16	89.02	93.10
Random Pixel Swap[g]	3	88.11	93.70[f]	89.63	93.80	73.17	74.64	89.94[f]	94.79[f]
Chest CT scan images dataset									
Base model[e]	8	93.33	99.00	87.30	97.09	82.86	95.84	84.76	96.92
Horizontal flip[e]	7	91.43	98.56	87.62	97.37	82.86	95.83	91.75	98.33
Horizontal flip with Random Pixel Swap[g]	4	93.97	98.96	89.21	97.48	88.57[f]	97.63[f]	92.70	98.09
Vertical flip[e,g]	6	93.33	98.86	83.81	96.83	84.72	96.10	92.06	98.58
Vertical flip with Random Pixel Swap[e,g]	5	93.97	98.81	86.67	97.13	84.76	96.23	91.43	98.04
Rotation[e,g]	3	95.24	99.22	90.16	97.58	84.57	94.95	95.87	99.03
Rotation with Random Pixel Swap[g]	1	96.19[f]	99.24[f]	91.75	97.73	85.23	95.10	96.19	98.99
Random Pixel Swap[g]	2	95.56	99.15	92.70[f]	98.02[f]	86.67	96.32	96.19[f]	98.90[f]

[a]IQ-OTH/NCCD: Iraq-Oncology Teaching Hospital/National Center for Cancer Diseases.

XSL·FO

RenderX

[b]CT: computed tomography.

[c]Rank represents the overall rating for each technique, with "1" indicating the best technique across all models.

[d]AUROC: area under the receiver operating characteristic curve.

[e]Significant difference between an augmentation technique and the Random Pixel Swap technique across all models.

[f]Highest value in the column.

[g]Significant difference between training using an augmentation technique and the base model across all models.

Validation Results of the Generalization Capabilities of the RPS Technique

Enhancing the generalization ability of DL models to unseen data represents a critical objective of DA [46]. To evaluate the RPS technique's capacity to improve model generalization, we conducted experiments using the selected models (ResNet-18, MobileNetV3 [small], Vision Transformer [base-16], and Swin Transformer [tiny]). Models were trained on the IQ-OTH/NCCD dataset and validated on the chest CT scan images dataset (distinct collections acquired using different imaging equipment, protocols, time periods, and geographical locations). All models performed binary classification (cancerous vs normal) of CT images.

Our comparative analysis included the base models, RPS implementation, and selected standard DA techniques. The results demonstrated RPS's superior performance across all architectures (Cutout: P=.05; Random Erasing: P=.054; MixUp: P=.04; CutMix: P=.03), with an exception for the Vision Transformer implementation. Random Erasing showed a marginal 0.8% accuracy advantage over RPS. However, RPS maintained a significant 9.28% improvement in AUROC over Random Erasing. Furthermore, the cumulative ranking was as follows: (1) RPS (best technique), (2) Cutout, (3) CutMix, (4) Random Erasing, and (5) MixUp. The detailed results are presented in Table 9.

Table . Validation results of the generalization capabilities of different data augmentation techniques for lung cancer diagnosis using deep learning (224×224 image resolution).

Data augmentation	Rank[a]	ResNet-34		MobileNetV3 (small)		Vision Transformer (base-16)		Swin Transformer (tiny)	
		Accuracy, %	AUROC[b], %	Accuracy, %	AUROC, %	Accuracy, %	AUROC, %	Accuracy, %	AUROC, %
Base model[c]	3	82.53	84.33	91.24	90.03	79.29	63.60	92.22	95.22[d]
Cutout[c]	2	82.65	97.29	92.09	90.70	81.80	63.49	92.22	85.77
Random Erasing[e]	5	88.71	78.66	91.45	89.66	82.65[d]	58.92	91.96	85.96
MixUp[c]	6	83.80	95.17	91.58	90.36	80.74	52.24	91.58	79.73
CutMix[c]	4	84.57	94.36	90.69	80.26	81.12	66.90	92.09	86.09
Random Pixel Swap[e]	1	90.69[d]	97.48[d]	92.35[d]	93.30[d]	81.85	68.20[d]	92.35[d]	95.04

[a]Rank represents the overall rating for each technique, with "1" indicating the best technique across all models.

[b]AUROC: area under the receiver operating characteristic curve.

[c]Significant difference between an augmentation technique and the Random Pixel Swap technique across all models.

[d]Highest value in the column.

[e]Significant difference between training using an augmentation technique and the base model across all models.

Comparison With Prior Work

Our experimental results demonstrated improvements over the results of previous studies using both the IQ-OTH/NCCD and chest CT scan images datasets. For the IQ-OTH/NCCD dataset, our approach achieved a 7.67% performance improvement over a machine learning technique in the study by Kareem et al [57], a 4.76% improvement over an ensemble of VGG-16, ResNet-50, InceptionV3, and EfficientNetB7 models in the study by Solyman et al [58], and a 2.13% enhancement over an ensemble of 3 custom CNNs in the study by Abe et al [59]. Similarly, for the chest CT scan images dataset, our method showed a 5.78% improvement over a 3-layer custom CNN in the study by Mamun et al [60] and a 2.22% improvement over a 5-layer CNN with a custom Mavage Pooling layer in the study by Abe et al [47]. The comparative results are detailed in Table 10.

Table . Comparison of our study results with the results of previous studies on the analysis of the IQ-OTH/NCCD[a] and chest CT[b] scan images datasets.

Dataset and study	Accuracy, %	Number of classes
IQ-OTH/NCCD		
Kareem et al [57]	89.89	3
Solyman et al [58]	92.80	3
Abe et al [59]	95.43	3
Our study	97.56	3
Chest CT scan images		
Mamun et al [60]	92.00	4
Abe et al [47]	95.56	4
Our study	97.78	4

[a]IQ-OTH/NCCD: Iraq-Oncology Teaching Hospital/National Center for Cancer Diseases.

[b]CT: computed tomography.

Discussion

Principal Findings

The experimental results of the study demonstrated that the RPS DA technique significantly enhanced the diagnostic performance of both CNN and transformer architectures for lung cancer diagnosis from CT scan images. Our comprehensive evaluation demonstrated that RPS consistently outperformed 4 established augmentation methods (CutMix, Random Erasing, MixUp, and Cutout) across multiple performance metrics and diverse experimental conditions. The superior efficacy of RPS stems from its unique capacity to preserve critical anatomical content while generating clinically meaningful variations through controlled intraimage pixel swapping. This characteristic makes RPS particularly valuable for medical imaging applications where maintaining content integrity is essential for an accurate diagnosis.

For CNN architectures, specifically ResNet-34, RPS yielded remarkable performance improvements. ResNet-34 achieved peak accuracies of 97.56% for the IQ-OTH/NCCD dataset and 97.78% for the chest CT scan images dataset, with corresponding AUROC scores of 98.61% and 99.46%, respectively, at 512×512 image resolution. The technique's effectiveness with MobileNetV3 (96.65% accuracy and 98.0% AUROC for the IQ-OTH/NCCD dataset; 96.83% accuracy and 98.75% AUROC for the chest CT scan images dataset) is particularly notable given this model's lightweight architecture, suggesting RPS's potential for deployment in resource-constrained clinical settings where efficient models are often preferred [56]. The study results represent a substantial advancement over conventional augmentation approaches, as RPS effectively addresses the inherent limitation of CNNs in capturing global relationships by creating localized variations that enhance feature learning while preserving diagnostically relevant image features.

The transformer-based architectures (Vision Transformer and Swin Transformer) showed particularly notable improvements when augmented with RPS. While transformer models conventionally demand large-scale training datasets to achieve peak performance, RPS effectively compensated for data limitations by generating variations that preserved the overall image content for proper attention mechanism functioning. For the Vision Transformer, RPS augmentation significantly enhanced performance, reaching 92.38% accuracy and 96.93% AUROC on the IQ-OTH/NCCD dataset and 86.67% accuracy and 96.32% AUROC on the chest CT scan images dataset. The Swin Transformer demonstrated robust performance gains, achieving 94.82% accuracy and 98.12% AUROC on the IQ-OTH/NCCD dataset and 96.19% accuracy and 98.90% AUROC on the chest CT scan images dataset when enhanced with RPS. The study results showed that RPS enables transformer models to develop more robust and clinically relevant feature representations, even with limited training data.

Our comparative analysis revealed RPS's consistent dominance across evaluation metrics and experimental conditions. While CutMix showed marginal advantages in specific scenarios (notably a 0.31% accuracy improvement with ResNet-34 at 512×512 image resolution), RPS maintained substantially better AUROC scores (5.31% higher in the same comparison), indicating more reliable diagnostic discrimination capability. This performance pattern held true across both the IQ-OTH/NCCD and chest CT scan images datasets, with RPS consistently ranking the highest in our comprehensive evaluation framework. Importantly, while conventional augmentation techniques sometimes degraded model performance in certain scenarios [38,40], RPS demonstrated universal performance enhancement across all tested conditions. Three fundamental characteristics explain RPS's exceptional effectiveness. The first characteristic is anatomical content preservation. Unlike methods that erase or mix image regions, RPS maintains all original diagnostic information while creating realistic variations through a controlled, systematic, random patch-based pixel swap within carefully defined ROIs. This approach preserves the clinical relevance of training samples while providing valuable data diversity. The second characteristic is architecture agnostic adaptability. The technique's parameter-free implementation and tunable swap area factor enable optimal performance across diverse model architectures without requiring architecture-specific adjustments. This flexibility makes RPS particularly valuable for medical imaging research,

where multiple architectures may be explored. The third characteristic is clinical pathological relevance. By restricting pixel swaps to anatomically plausible regions within lung tissue (especially when combined with ROI segmentation), RPS enhances the learning of pathological features that may appear anywhere in the pulmonary anatomy, a crucial capability given the unpredictable spatial distribution of malignant nodules in many cancer cases [61].

Validation experiments using independently acquired datasets with different scanning protocols and equipment configurations demonstrated RPS's superior generalization capabilities. The technique achieved these results while adding minimal computational overhead (statistically insignificant increases in training time, $P>.05$), making it practical for real-world clinical implementation. Furthermore, RPS showed excellent compatibility with conventional augmentation methods, providing additional performance gains when combined with rotation and flipping operations, which suggests easy integration into existing medical image processing pipelines.

These findings offer significant implications for the development of computer-aided diagnosis systems. RPS directly addresses two fundamental challenges in medical AI: (1) the scarcity of annotated medical imaging data and (2) the limited generalizability of many models across different clinical settings [23]. By consistently outperforming current state-of-the-art techniques while maintaining computational efficiency, RPS emerges as a versatile solution suitable for both research investigations and clinical deployment. Additionally, the technique's effectiveness suggests promising applications in educational settings for training radiologists, where realistic image variations could enhance learning without requiring additional patient scans.

Conclusions

The findings of this study demonstrate that RPS is a robust and versatile DA technique that significantly enhances the performance of both CNN and transformer architectures for lung cancer diagnosis from CT scan images. By preserving anatomical content while introducing meaningful variability, RPS outperforms existing augmentation methods across multiple metrics and datasets, achieving improved accuracy and AUROC scores. Its computational efficiency, adaptability to diverse architectures, and ability to improve generalization make it particularly valuable for medical imaging applications where data scarcity and model reliability are critical challenges. RPS not only advances the technical frontier of DA but also holds immediate promise for improving computer-aided diagnosis systems in clinical practice. Future work will explore its extension to other medical imaging modalities (magnetic resonance, ultrasound, and x-ray imaging) and extension to 3D applications.

Data Availability

The data supporting the findings of this study are available upon reasonable request from the corresponding author.

The Iraq-Oncology Teaching Hospital/National Center for Cancer Diseases (IQ-OTH/NCCD) dataset is available on Mendeley Data [35]. The chest computed tomography (CT) scan images dataset is available on Kaggle [55]. The code is available on GitHub [62].

Authors' Contributions

Conceptualization: AAA
Data curation: AAA
Formal analysis: AAA
Funding acquisition: AAA
Investigation: AAA
Methodology: AAA
Project administration: NM
Resources: AAA
Software: AAA
Supervision: NM
Validation: AAA
Visualization: AAA
Writing – original draft: AAA
Writing – review & editing: NM
All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

References

1. Yoon SM, Shaikh T, Hallman M. Therapeutic management options for stage III non-small cell lung cancer. World J Clin Oncol 2017 Feb 10;8(1):1-20. [doi: 10.5306/wjco.v8.i1.1] [Medline: 28246582]
2. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA A Cancer J Clinicians 2018 Nov;68(6):394-424. [doi: 10.3322/caac.21492]
3. Schabath MB, Cote ML. Cancer progress and priorities: lung cancer. Cancer Epidemiol Biomarkers Prev 2019 Oct;28(10):1563-1579. [doi: 10.1158/1055-9965.EPI-19-0221] [Medline: 31575553]
4. Travis WD, Brambilla E, Noguchi M, et al. International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma. J Thorac Oncol 2011 Feb;6(2):244-285. [doi: 10.1097/JTO.0b013e318206a221] [Medline: 21252716]
5. Shin HJ, Kim MS, Kho BG, et al. Delayed diagnosis of lung cancer due to misdiagnosis as worsening of sarcoidosis: a case report. BMC Pulm Med 2020 Mar 21;20(1):71. [doi: 10.1186/s12890-020-1105-2] [Medline: 32199453]
6. Ruano-Raviña A, Provencio M, Calvo de Juan V, et al. Lung cancer symptoms at diagnosis: results of a nationwide registry study. ESMO Open 2020 Nov;5(6):e001021. [doi: 10.1136/esmoopen-2020-001021] [Medline: 33214227]
7. Del Ciello A, Franchi P, Contegiacomo A, Cicchetti G, Bonomo L, Larici AR. Missed lung cancer: when, where, and why? Diagn Interv Radiol 2017;23(2):118-126. [doi: 10.5152/dir.2016.16187] [Medline: 28206951]
8. Blandin Knight S, Crosbie PA, Balata H, Chudziak J, Hussell T, Dive C. Progress and prospects of early detection in lung cancer. Open Biol 2017 Sep;7(9):170070. [doi: 10.1098/rsob.170070] [Medline: 28878044]
9. Meza R, Jeon J, Toumazis I, et al. Evaluation of the benefits and harms of lung cancer screening with low-dose computed tomography: modeling study for the US preventive services task force. JAMA 2021 Mar 9;325(10):988-997. [doi: 10.1001/jama.2021.1077] [Medline: 33687469]
10. Fitzgerald RC, Antoniou AC, Fruk L, Rosenfeld N. The future of early cancer detection. Nat Med 2022 Apr;28(4):666-677. [doi: 10.1038/s41591-022-01746-x] [Medline: 35440720]
11. Abujudeh HH, Boland GW, Kaewlai R, et al. Abdominal and pelvic computed tomography (CT) interpretation: discrepancy rates among experienced radiologists. Eur Radiol 2010 Aug;20(8):1952-1957. [doi: 10.1007/s00330-010-1763-1] [Medline: 20336300]
12. Hoffman RM, Atallah RP, Struble RD, Badgett RG. Lung cancer screening with low-dose CT: a meta-analysis. J Gen Intern Med 2020 Oct;35(10):3015-3025. [doi: 10.1007/s11606-020-05951-7] [Medline: 32583338]
13. Bonney A, Malouf R, Marchal C, et al. Impact of low-dose computed tomography (LDCT) screening on lung cancer-related mortality. Cochrane Database Syst Rev 2022 Aug 3;8(8):CD013829. [doi: 10.1002/14651858.CD013829.pub2] [Medline: 35921047]
14. Swensen SJ, Jett JR, Hartman TE, et al. Lung cancer screening with CT: Mayo Clinic experience. Radiology 2003 Mar;226(3):756-761. [doi: 10.1148/radiol.2263020036] [Medline: 12601181]
15. Silvestri GA, Goldman L, Tanner NT, et al. Outcomes from more than 1 million people screened for lung cancer with low-dose CT imaging. Chest 2023 Jul;164(1):241-251. [doi: 10.1016/j.chest.2023.02.003] [Medline: 36773935]
16. Krupinski EA, Berbaum KS, Caldwell RT, Schartz KM, Kim J. Long radiology workdays reduce detection and accommodation accuracy. J Am Coll Radiol 2010 Sep;7(9):698-704. [doi: 10.1016/j.jacr.2010.03.004] [Medline: 20816631]
17. Jacobsen MM, Silverstein SC, Quinn M, et al. Timeliness of access to lung cancer diagnosis and treatment: a scoping literature review. Lung Cancer (Auckl) 2017 Oct;112:156-164. [doi: 10.1016/j.lungcan.2017.08.011] [Medline: 29191588]
18. Sathyakumar K, Munoz M, Singh J, Hussain N, Babu BA. Automated lung cancer detection using artificial intelligence (AI) deep convolutional neural networks: a narrative literature review. Cureus 2020 Aug 25;12(8):e10017. [doi: 10.7759/cureus.10017] [Medline: 32989411]
19. Huang S, Yang J, Shen N, Xu Q, Zhao Q. Artificial intelligence in lung cancer diagnosis and prognosis: current application and future perspective. Semin Cancer Biol 2023 Feb;89:30-37. [doi: 10.1016/j.semcancer.2023.01.006] [Medline: 36682439]
20. Marentakis P, Karaiskos P, Kouloulias V, et al. Lung cancer histology classification from CT images based on radiomics and deep learning models. Med Biol Eng Comput 2021 Jan;59(1):215-226. [doi: 10.1007/s11517-020-02302-w] [Medline: 33411267]
21. Buduma N, Buduma N, Papa J. Fundamentals of Deep Learning, 2nd edition: O'Reilly Media, Inc; 2022.
22. Forte GC, Altmayer S, Silva RF, et al. Deep learning algorithms for diagnosis of lung cancer: a systematic review and meta-analysis. Cancers (Basel) 2022 Aug 9;14(16):3856. [doi: 10.3390/cancers14163856] [Medline: 36010850]
23. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale. arXiv. 2020. URL: https://arxiv.org/abs/2010.11929 [accessed 2025-08-22]
24. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE 1998;86(11):2278-2324. [doi: 10.1109/5.726791]

25. Lu K, Xu Y, Yang Y. Comparison of the potential between transformer and CNN in image classification. Presented at: 2nd International Conference on Machine Learning and Computer Application; Dec 17-19, 2021; Shenyang, China URL: https://ieeexplore.ieee.org/document/9736894 [accessed 2025-08-22]

26. Wang H, Liu Z, Ai T. Long-range dependencies learning based on non-local 1D-convolutional neural network for rolling bearing fault diagnosis. J Dyn Monit Diagn 2022 Apr 12(3):148-159. [doi: 10.37965/jdmd.2022.53]

27. Zhai X, Kolesnikov A, Houlsby N, Beyer L. Scaling vision transformers. Presented at: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Jun 18-24, 2022; New Orleans, LA. [doi: 10.1109/CVPR52688.2022.01179]

28. Matsoukas C, Haslum JF, Söderberg M, Smith K. Is it time to replace CNNs with transformers for medical images? arXiv. 2021. URL: https://arxiv.org/abs/2108.09038 [accessed 2025-08-22]

29. Liu Z, Lin Y, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows. Presented at: 2021 IEEE/CVF International Conference on Computer Vision (ICCV); Oct 10-17, 2021; Montreal, QC, Canada. [doi: 10.1109/ICCV48922.2021.00986]

30. Liu D, Liu F, Tie Y, Qi L, Wang F. Res-trans networks for lung nodule classification. Int J Comput Assist Radiol Surg 2022 Jun;17(6):1059-1068. [doi: 10.1007/s11548-022-02576-5] [Medline: 35290646]

31. Nejad RR, Hooshmand S. HViT4Lung: hybrid vision transformers augmented by transfer learning to enhance lung cancer diagnosis. Presented at: 5th International Conference on Bio-engineering for Smart Technologies (BioSMART); Jun 7-9, 2023; Paris, France. [doi: 10.1109/BioSMART58455.2023.10162074]

32. Atmakuru A, Chakraborty S, Faust O, et al. Deep learning in radiology for lung cancer diagnostics: a systematic review of classification, segmentation, and predictive modeling techniques. Expert Syst Appl 2024 Dec;255:124665. [doi: 10.1016/j.eswa.2024.124665]

33. Santos C, Papa JP. Avoiding overfitting: a survey on regularization methods for convolutional neural networks. ACM Comput Surv 2022 Jan 31;54(10s):1-25. [doi: 10.1145/3510413]

34. Razzak MI, Naz S, Zaib A. Deep learning for medical image processing: overview, challenges and the future. In: Dey N, Ashour A, Borra S, editors. Classification in BioApps Lecture Notes in Computational Vision and Biomechanics: Springer; 2018:323-350. [doi: 10.1007/978-3-319-65981-7_12]

35. Alyasriy H. The IQ-OTHNCCD lung cancer dataset. Mendeley Data. 2020. URL: https://data.mendeley.com/datasets/bhmdr45bh2/1 [accessed 2025-08-22]

36. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data 2019 Dec;6(1):1-48. [doi: 10.1186/s40537-019-0197-0]

37. Chlap P, Min H, Vandenberg N, Dowling J, Holloway L, Haworth A. A review of medical image data augmentation techniques for deep learning applications. J Med Imaging Radiat Oncol 2021 Aug;65(5):545-563. [doi: 10.1111/1754-9485.13261] [Medline: 34145766]

38. Lin CH, Kaushik C, Dyer EL, Muthukumar V. The good, the bad and the ugly sides of data augmentation: an implicit spectral regularization perspective. J Mach Learn Res 2024;25:1-85 [FREE Full text]

39. Maharana K, Mondal S, Nemade B. A review: data pre-processing and data augmentation techniques. Global Transitions Proceedings 2022 Jun;3(1):91-99. [doi: 10.1016/j.gltp.2022.04.020]

40. Balestriero R, Bottou L, LeCun Y. The effects of regularization and data augmentation are class dependent. Presented at: 36th International Conference on Neural Information Processing Systems; Nov 28 to Dec 9, 2022; New Orleans, LA. [doi: 10.5555/3600270.3603015]

41. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2012;60(6):84-90. [doi: 10.1145/306538]

42. Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. Presented at: 28th International Conference on Neural Information Processing Systems; Dec 8-13, 2014; Montreal, Canada. [doi: 10.5555/2969033.2969125]

43. DeVries T, Taylor GW. Improved regularization of convolutional neural networks with cutout. arXiv. 2017. URL: https://arxiv.org/abs/1708.04552 [accessed 2025-08-22]

44. Zhong Z, Zheng L, Kang G, Li S, Yang Y. Random erasing data augmentation. Proc AAAI Conf Artif Intell 2020 Apr 3;34(7):13001-13008. [doi: 10.1609/aaai.v34i07.7000]

45. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. Mixup: beyond empirical risk minimization. arXiv. 2017. URL: https://arxiv.org/abs/1710.09412 [accessed 2025-08-22]

46. Yun S, Han D, Chun S, Oh SJ, Yoo Y, Choe J. CutMix: regularization strategy to train strong classifiers with localizable features. Presented at: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); Oct 27 to Nov 2, 2019; Seoul, Korea (South). [doi: 10.1109/ICCV.2019.00612]

47. Abe A, Nyathi M, Okunade A. Lung cancer diagnosis from computed tomography scans using convolutional neural network architecture with Mavage pooling technique. AIMS Med Sci 2025;12(1):13-27. [doi: 10.3934/medsci.2025002]

48. Jiang Y, Manem VSK. Data augmented lung cancer prediction framework using the nested case control NLST cohort. Front Oncol 2025 Feb 25;15:1492758. [doi: 10.3389/fonc.2025.1492758]

49. Jin X, Zhu H, Li S, et al. A survey on mixup augmentations and beyond. arXiv. 2024. URL: https://arxiv.org/abs/2409.05202 [accessed 2025-08-22]

50. He J, Liu B, Yang X. Non-local patch mixup for unsupervised domain adaptation. Presented at: 2022 IEEE International Conference on Data Mining (ICDM); Nov 28 to Dec 1, 2022; Orlando, FL. [doi: 10.1109/ICDM54844.2022.00116]

51. Oh J, Yun C. Provable benefit of Cutout and CutMix for feature learning. arXiv. 2024. URL: https://arxiv.org/abs/2410.23672 [accessed 2025-08-22]

52. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Jun 27-30, 2016; Las Vegas, NV. [doi: 10.1109/CVPR.2016.90]

53. Howard A, Sandler M, Chen B, et al. Searching for mobilenetv3. Presented at: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); Oct 27 to Nov 2, 2019; Seoul, Korea (South). [doi: 10.1109/ICCV.2019.00140]

54. Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library. Presented at: 33rd International Conference on Neural Information Processing Systems; Dec 8-14, 2019; Vancouver, BC, Canada. [doi: 10.5555/3454287.3455008]

55. Chest CT-scan images dataset. Kaggle. URL: https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images [accessed 2025-08-22]

56. Setio AAA, Traverso A, de Bel T, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. Med Image Anal 2017 Dec;42:1-13. [doi: 10.1016/j.media.2017.06.015] [Medline: 28732268]

57. Kareem HF, AL-Huseiny MS, Y. Mohsen F, A. Khalil E, S. Hassan Z. Evaluation of SVM performance in the detection of lung cancer in marked CT scan dataset. Indones J Electr Eng Comput Sci 2021 Mar;21(3):1731. [doi: 10.11591/ijeecs.v21.i3.pp1731-1738]

58. Solyman S, Schwenker F. Lung tumor detection and recognition using deep convolutional neural networks. In: Girma Debelee T, Ibenthal A, Schwenker F, editors. Pan-African Conference on Artificial Intelligence 2023:79-91. [doi: 10.1007/978-3-031-31327-1_5]

59. Abe AA, Nyathi M, Okunade AA, Pilloy W, Kgole B, Nyakale N. A robust deep learning algorithm for lung cancer detection from computed tomography images. Intelligence-Based Medicine 2025;11:100203. [doi: 10.1016/j.ibmed.2025.100203]

60. Mamun M, Mahmud MI, Meherin M, Abdelgawad A. LCDctCNN: lung cancer diagnosis of CT scan images using CNN based model. Presented at: 10th International Conference on Signal Processing and Integrated Networks (SPIN); Mar 23-24, 2023; Noida, India. [doi: 10.1109/SPIN57001.2023.10116075]

61. Ali H, Mohsen F, Shah Z. Improving diagnosis and prognosis of lung cancer using vision transformers: a scoping review. BMC Med Imaging 2023 Sep 15;23(1):129. [doi: 10.1186/s12880-023-01098-z] [Medline: 37715137]

62. Random pixel swap (RPS) data augmentation technique code. GitHub. URL: https://github.com/Saintcodded/Random-Pixel-Swap-RPS-Data-Augmentation-Technique [accessed 2025-08-22]

Abbreviations

AUROC: area under the receiver operating characteristic curve
CNN: convolutional neural network
CT: computed tomography
DA: data augmentation
DL: deep learning
IQ-OTH/NCCD: Iraq-Oncology Teaching Hospital/National Center for Cancer Diseases
MVSF: multivalue swap area factor
ROI: region of interest
RPS: Random Pixel Swap
SVSF: single-value swap area factor

XSL•FO
RenderX

Using Natural Language Processing to Identify Symptomatic Adverse Events in Pediatric Oncology: Tutorial for Clinician Researchers

Clifton P Thornton[1,2*], PhD; Maryam Daniali[3*], PhD; Lei Wang[3,4*], PhD; Spandana Makeneni[3*], PhD; Allison Barz Leahy[5,6*], MD

[1]Center for Pediatric Nursing Research & Evidence-Based Practice, Children's Hospital of Philadelphia, 734 Schuylkill Avenue, Philadelphia, PA, United States

[2]School of Nursing, University of Pennsylvania, Philadelphia, PA, United States

[3]Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, PA, United States

[4]College of Computing and Informatics, Drexel University, Philadelphia, PA, United States

[5]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

[6]Cancer Center, Children's Hospital of Philadelphia, Philadelphia, PA, United States

[*]all authors contributed equally

Corresponding Author:
Clifton P Thornton, PhD
Center for Pediatric Nursing Research & Evidence-Based Practice, Children's Hospital of Philadelphia, 734 Schuylkill Avenue, Philadelphia, PA, United States

Abstract

Artificial intelligence (AI) is poised to become an integral component in health care research and delivery, promising to address complex challenges with unprecedented efficiency and precision. However, many clinicians lack training and experience with AI, and for those who wish to incorporate AI into research and practice, the path forward remains unclear. Technical barriers, institutional constraints, and lack of familiarity with computer and data science frequently stall progress. In this tutorial, we present a transparent account of our experiences as a newly established interdisciplinary team of clinical oncology researchers and data scientists working to develop a natural language processing model to identify symptomatic adverse events during pediatric cancer therapy. We outline the key steps for clinicians to consider as they explore the utility of AI in their inquiry and practice, including building a digital laboratory, curating a large clinical dataset, and developing early-stage AI models. We emphasize the invaluable role of institutional support, including financial and logistical resources, and dedicated and innovative computer and data scientists as equal partners in the research team. Our account highlights both facilitators and barriers encountered spanning financial support, learning curves inherent with interdisciplinary collaboration, and constraints of time and personnel. Through this narrative tutorial, we intend to demystify the process of AI research and equip clinicians with actionable steps to initiate new ventures in oncology research. As AI continues to reshape the research and practice landscapes, sharing insights from past successes and challenges will be essential to informing a clear path forward.

Introduction

The development of sophisticated machine learning, deep learning, natural language processing (NLP), and large language models has showcased artificial intelligence's (AI's) potential to accelerate advances in health care research and clinical practice [1-3]. However, growing clinician interest in employing AI as a research tool is often met with challenges in understanding its nuances and applications. The proper and safe use of AI requires in-depth knowledge of computer science, big data analytics, and specialized data science and biostatistical approaches – skills that clinicians typically do not possess. Conversely, computer and data scientists with expertise in AI who wish to contribute to clinical advances must develop familiarity with a clinical specialty and acquire a deep understanding of the intricacies of care delivery, research, and biomedical needs. As a result, the effective use of AI in health care environments necessitates collaborative integration between computer science and health care disciplines, bringing together expertise from these disparate fields [4-6].

Although clinicians are increasingly eager to incorporate AI into their research efforts, many face uncertainty on how to

begin or establish effective collaborations with computer and data scientists. Using the initial phase of our pilot AI work as an exemplar, we outline strategies for leveraging AI and NLP in pediatric cancer inquiry, focusing on the process of building a team blending AI and clinical oncology research. Our transparent account details the formation of an interdisciplinary team bridging clinical oncology and data science, highlights challenges encountered, and shares lessons learned. The purpose of this descriptive tutorial is to make AI approachable for clinical researchers who are motivated to address complex clinical questions but may lack technical expertise. Key challenges for teams to consider are explicitly identified within this study. We aim to equip clinician readers with an introductory framework for initiating AI-driven research projects, while emphasizing the logistic, financial, and personnel resources essential for success.

The Clinical Problem and Need for an AI-Based Solution

Cancer-directed therapy is inherently toxic, causing a host of adverse events that are burdensome, costly, dangerous, and sometimes life-threatening [7-9]. When toxicities become severe, future therapy doses are reduced, delayed, or omitted, which potentially compromises long-term survival [10]. Because of these deleterious effects, research focused on early detection has been prioritized, so that prompt and effective interventions can be designed to mitigate toxicity and improve clinical outcomes [11-13].

Therapy-related toxicities are broadly categorized into nonsymptomatic and symptomatic adverse events. Nonsymptomatic adverse events are objective and easy to identify, quantify, and analyze because they are readily detectable through structured data like laboratory values or diagnostic imaging. These clean and structured data allow researchers to stratify patient cohorts, correlate symptoms with biomarkers and treatment factors, and derive actionable insights.

In contrast, symptomatic adverse events are subjective and must be elicited, interpreted, or individually assessed by clinicians [14,15]. Furthermore, these events are typically captured in unstructured, free-text clinical notes which constrains systematic identification and analysis, making data extraction labor-intensive, time-consuming, and prone to inconsistencies [7-9,16,17]. Not surprisingly, the data are often unreliable [18,19], with significant negative repercussions on subsequent analyses. The inability to reliably study symptomatic adverse events is particularly concerning because they are among the most common therapy-related toxicities and frequently lead to treatment interruptions.

AI is a promising method for the reliable extraction and analysis of symptomatic adverse events from electronic medical records (EMRs). In fact, NLP technology has already had preliminary success in identifying their presence within unstructured, free-text clinical notes [20-24].

In pediatric oncology, 5 symptomatic adverse events associated with chemotherapy stand out due to their prevalence and serious sequelae—nausea, vomiting, constipation, neuropathy, and mucositis. Herein, we describe our interdisciplinary approach for assessing the ability of an NLP algorithm to identify these adverse events in pediatric oncology patient records. The initial phase of this work, serving as the exemplar for this tutorial, is to evaluate the degree to which existing NLP models can identify symptomatic adverse events in pediatric cancer therapy.

Infrastructure, Personnel, and Funding

AI-based health care research necessitates substantial data and computer science support. Optimally, this support is institutional, with health care enterprises investing in employing, contracting, or collaborating with skilled data scientists dedicated to advancing clinical inquiry. Collaboration between these technology experts and clinician researchers, along with departmental backing to support clinical inquiry and innovation, as well as the necessary data infrastructure, is essential to cultivating advancements in this emerging domain [25].

Our institution houses a Data Science and Biostatistics Unit (DSBU), a centralized service unit that comprises a robust mix of 30 PhD- and master-level biostatisticians and data scientists who work with principal investigators to address research questions via data consultation, study design, methodology expertise, data preparation, data analyses, and manuscripts development. The DSBU is housed within the Department of Biomedical and Health Informatics, which provides an academic home and service base for all research informatics activities at the institution, including the development and deployment of intellectual, technical, and educational resources in biomedical computing.

Through an enterprise-level strategic initiative, our institute developed a next-generation suite of tools and services, Arcus, that provides a digital laboratory environment for investigators and project staff to securely store, access, and process electronic patient data. The Arcus program is staffed by archivists, librarians, information analysts, cloud computing engineers, programmers, statisticians, and privacy experts. Data are managed through the oversight of the Institutional Review Board (IRB), and access is governed by multiple institutional policies. Arcus security configuration and controls are based on the HIPAA (Health Insurance Portability and Accountability Act) Security Rule.

For this work, project team members from DSBU and Arcus included 3 PhD-prepared data scientists and a data integration manager. Initial services to set up the project were provided at no cost through the internal consultation mechanisms. As the project developed and expanded, pilot funding was secured through internal grant mechanisms and preliminary data were used to secure external grant funding. A data science supervisor available through the center provided guidance in approaching an AI-based research project.

The project team's clinical experts were 2 oncology clinicians and researchers who served as coprincipal investigators—a PhD-prepared scientist and nurse practitioner under the Center for Pediatric Nursing Research and Evidence-Based Practice and Cancer Center and an attending pediatric oncologist in the Division of Pediatric Oncology and School of Medicine.
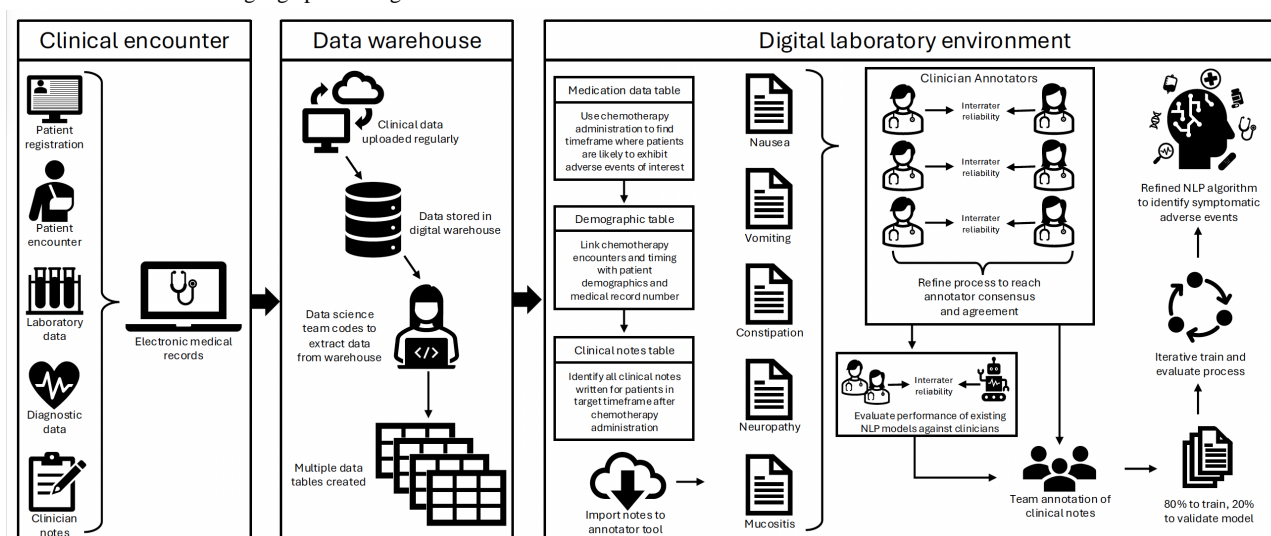
Building the Digital Laboratory

The clinician researchers consulted with the data science team extensively to determine necessary data elements and ensure feasibility. An IRB application was submitted, the research was determined to meet exemption criteria, and a HIPAA waiver was authorized (IRB 24‐021922).

Activities relating to building the digital laboratory, including data flow and processing, are outlined in Figure 1. Inclusion criteria were set to any patient aged younger than 25 years who received treatment for cancer at our institution within the previous 10 years. We used *International Classification of Diseases, Ninth Revision* (*ICD-9*) or *International Classification of Diseases, Tenth Revision* (*ICD-10*) diagnosis codes, Current Procedural Terminology codes for cancer-directed therapies in conjunction with institutional cancer registry data to identify those who received cancer treatment ("Clinical Encounter" in Figure 1). Eligible patients were assigned a unique identifier and added to the digital laboratory. Importantly, each unique identifier retained a link to the patient's electronic health record (EHR) medical record number to ensure reliable linking of patients with relevant clinical data. Necessary EHR data elements (eg, chemotherapy administration records, clinical notes, and laboratory values) were identified via joint clinical and data science team meetings and were then imported from the data warehouse into the digital environment ("Data Warehouse" in Figure 1).

Figure 1. Research project progression and data flow from clinical encounters to the data warehouse, and manipulation within the digital laboratory environment. NLP: natural language processing.

Although careful planning to meet aims is necessary for all research projects, big data and AI-based research involves the additional step of evaluating the accessibility and reliability of data. A key challenge in building a digital lab is the extensive refinement of data that is required because digital storage of medical data differs from digital data display (the way data appears to the clinician in the EHR). Within the data warehouse, clinical notes are sorted and stored based on their version status as templated, signed, addended, or modified – with each note potentially possessing multiple versions. But in the clinical setting, the only note displayed for staff is the most recent version. Therefore, to ensure data matched the clinical documentation, the data science team wrote complex code that selected the most recent version, irrespective of its assigned status. This was essential because there are millions of source notes for this work and importing multiple versions of each is not feasible due to time, data storage, and computational processing limitations.

Chemotherapy agents were identified using medication classification codes created for the purpose of this work and then integrated with patient medication administration records to identify the specific administration time and dose. This vital step underscores the need for a skilled data scientist or analyst to be an integral member of the research team. Laboratory values, easily extracted from the source EHR data warehouse, also were imported to assist clinical researchers with interpretation of data, as needed.

After 14 months of collaborative effort, all data were imported to the laboratory ("Digital laboratory environment" in Figure 1) which included data on 18,408 patients, encompassing 4.8 million clinical notes and over 450 million medication dose administrations. From this point forward, all research activities were performed in the digital laboratory environment. It should be noted that due to the massive size of EHR data files and the sheer number of individual variables, discrete data elements are imported to the digital laboratory in the form of tables in a relational database. For example, the medication administration table comprised dozens of datapoints for each of the hundreds of millions of doses administered within our patient cohort. Similarly, the demographic information table contained dozens of variables and associated metadata for each patient. The clinical notes table not only included the full note text, but also other metadata that provided information about the notes themselves (eg, timestamps, subtype, and author type).

Identifying the relevant and necessary data elements from these tables and joining them in relational databases required the expertise of a PhD-level data scientist with fluency in programming and querying in SQL and R languages. The

clinician scientists provided direction for selecting elements but did not have the skill to perform the tasks. Once relational tables and databases were created, the team could jointly verify data integrity through face validity of items (eg, chemotherapy agents matched oncologic diagnoses for patients). The team also reviewed randomly selected medical records of patients in the database to ensure correct elements and values were identified and joined appropriately in the newly created data tables.

During the process of building the digital laboratory, unexpected challenges arose from the complexity of the structures of EHR data and the differences between digital data storage and display that complicated data pulling and importing approaches. The complexity of identifying and pulling these data was also underestimated by the data science team, and the process took much longer than expected, by a scale of about a year. Clinician researchers taking initial steps to AI-based methods should account for time required to learn new skills and take additional time to clean and validate data. However, the accessibility to data scientist and technology expert knowledge, skills, and time coupled with the infrastructure provided by institutional investments and external grant funding made the project both feasible and possible.

Identifying Notes of Interest

Training and evaluation of the NLP model is an iterative process requiring labeled data. For this project, the labeled data are annotations of text, wherein a clinician reads through clinical notes and tags sections that indicate the absence, presence, and severity of the adverse event of interest. A typical allocation of 80% of annotated notes for model training and 20% for validation was used. The necessary number of labeled notes varies considerably depending on the complexity of the task (ie, difficulty of being able to identify the adverse event of interest) and the selected NLP methodology. For these reasons, it is not possible to a priori estimate the minimum number of notes required to adequately train and validate the model. Thus, we used an incremental annotation process starting with a minimum sample size for a limited population similar to previous work [26]. For clinician researchers accustomed to a priori–determined sample sizes, this was difficult to conceptualize and resulted in downstream challenges in time management and resource allocation for the project. Adopting a qualitative research mindset – where recruitment is ongoing until data saturation is achieved – is helpful when conceptualizing sample size for a project like this, despite being a technique not used frequently in quantitative methodologies.

The process of identifying notes for annotation required several months and the expertise of a PhD-prepared data scientist skilled in coding and data analysis. Our goal was to identify notes with a high likelihood of containing documentation related to the adverse events of interest to facilitate faster model training. As such, clinical researchers identified key scenarios and exposures associated with nausea, vomiting, constipation, neuropathy, and mucositis. This process involved specifying chemotherapy agents, dosages, and the typical time frames within which these toxicities manifest following administration. The schema used for identifying notes meeting these criteria is outlined in Figure 2.

Figure 2. Schema for identifying clinical notes to annotate for natural language processing training. BSA: body surface area; *ICD-10*: *International Classification of Diseases, Tenth Revision.*
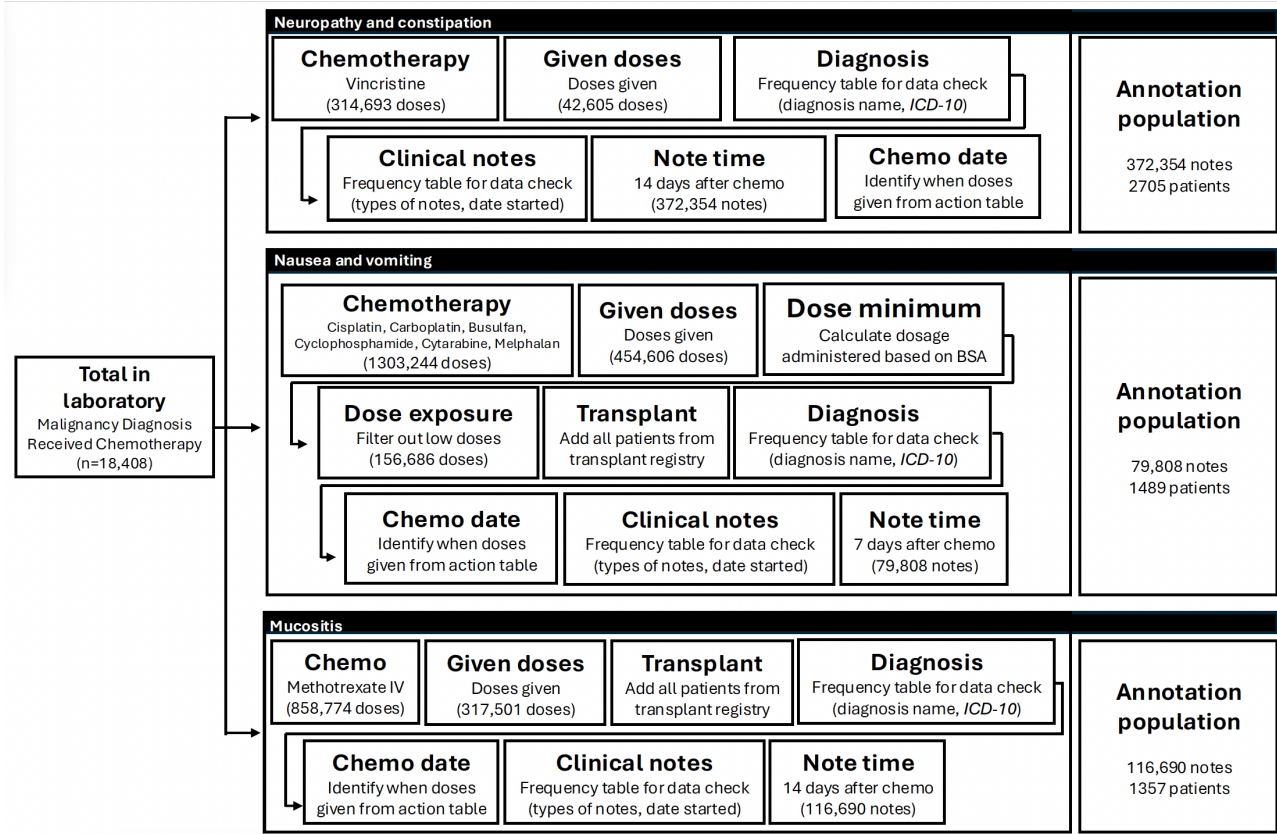
To identify clinical notes most likely to document constipation or neuropathy, we identified instances of vincristine administration. For nausea and vomiting, highly emetogenic chemotherapy agents were identified [27]. Given that emetogenicity depends on dosage, body surface area was calculated using the most recent height and weight measurements, and doses below the emetogenic threshold were excluded. Patients undergoing conditioning chemotherapy for stem cell transplantation, which is universally highly emetogenic, were included based on an institutional transplant registry. To identify documentation of mucositis, we focused on intravenous methotrexate administrations as well as stem cell transplantation.

Chemotherapy doses were identified from the medication table, and a frequency table of administration events, including action and date and time, was reviewed to ensure proper documentation (eg, marked as "given" in the medical records). Determination of how administered medications are recorded in the data warehouse required consultation with an informaticist, since multiple actions (eg "missed," "late," "withheld," "administered," and "given") are assigned to medications in the dataset with ambiguous meanings. Cross-referencing with data visible in the EHR was required to ensure that the devised algorithm and decisions were made. As before, the clinical researchers learned that data stored in the data warehouse is far more complex than that which is displayed in the EHR. Identifying administered medications within the medication administration record in the "visible" EHR, for example, is far more straightforward, but incredibly labor-intensive.

Patient identifiers were cross-referenced with demographic and diagnosis tables, followed by the generation of a frequency table of oncologic diagnoses and associated *ICD-9* and *ICD-10* codes for each target symptom. Clinical researchers reviewed these tables for errors or incongruences to establish face validity, ensuring that chemotherapy agents matched the diagnoses. After validating these data, we cross-referenced the clinical notes table using patient identifiers to extract notes written within 14 days of chemotherapy administration for neuropathy, constipation, and mucositis; and 7 days for nausea or vomiting, in accordance with expected clinical timelines.

Initial review suggested that certain note types—such as history and physicals, progress notes, nursing notes, and discharge summaries—were most likely to contain relevant data. However, inconsistencies in data labeling posed challenges; for instance, "progress notes" were used for documentation by multiple specialties, adding noise to the dataset. After careful review, notes authored by clinical nutritionists, pharmacists, social workers, case managers, speech and language pathologists, occupational therapists, and physical therapists were excluded. Only the most recent version of each note (signed, addended, or modified) as determined by date of note initiation and note status was retained.

Key challenges to identifying relevant note types, versions, and authors arose from the time-intensive nature of extensive data extraction and manual review required. Clinical staff encountered challenges in understanding how medical record data were stored within the data warehouse, particularly regarding labeling of note versions and determining when patients received medications. Overcoming this challenge highlights the importance of properly understanding the metadata that accompanies variables of interest, and the parallel importance of including all metadata in the digital laboratory. As before, the team learned that the vocabulary typically used in the clinical environment does not match that used in informatics. For example, in clinical practice, "administered" or "given" are used synonymously to indicate that a patient has received a medication. However, these had different meanings in the data warehouse, so understanding how data are labeled and not making assumptions is vital. Validating the data by reviewing constructed tables and comparing them to patient medical records is necessary to ensure the integrity of the data. These are both nuanced and time-consuming steps that should be considered as expected components of all big data or AI-based research projects.

Real-time collaboration with a dedicated data scientist enabled efficient extraction and validation of large datasets. The integration of this expertise allowed for immediate adjustments based on clinical input, ensuring that the final dataset was both comprehensive and focused and underscored the importance of interdisciplinary collaboration and iterative problem-solving.

Annotation and Validation

An annotation guide was created by the clinician researchers to standardize the annotation process and ensure consistency in identifying and grading adverse events. The guide aimed to provide clear instructions for clinical abstractors and facilitate uniform application of the National Cancer Institute's Common Terminology Criteria for Adverse Events (CTCAE) [28] to patient records.

The guide was created iteratively, beginning with an initial draft used by clinician researchers during joint annotation sessions. Common challenges encountered during annotation were documented, and adjudication decisions were included to ensure consistency. Common data extraction elements that required discussion among clinicians were included in the guide to define consensus between researchers and to provide consistency to annotators. The guide accounts for nuances of clinical documentation such as shorthand abbreviations, terminology variations, and physical exam findings. To initiate the annotation process, 100 notes, representing an intersection of chemotherapy exposures associated with all the target adverse events, were uploaded to the annotation tool. Clinicians independently annotated 30 notes, comparing results to assess alignment that facilitated refinement of the annotation guide before independently completing the remaining 70 notes. Annotation overlap and agreement were systematically evaluated, with areas of disagreement manually adjudicated and further revisions made to the guide.

A second and third batch of 100 notes was then annotated independently and annotator agreement calculated after each round. Annotator agreement was evaluated by interrater reliability calculated by tag agreement at the symptom level (constipation, mucositis, nausea, neuropathy, and vomiting) and at the symptom degree level (eg, CTCAE severity level).

Weighted Cohen kappa quantified the level of agreement to provide a measure of agreement accounting for the likelihood of agreement occurring by chance. Manual adjudication after each round was then undertaken, followed by revision of the annotation guide. Discrepancies were explored to identify opportunities for improvement and additional nuances in clinical documentation.

Unexpectedly, initial low agreement between abstractors highlighted challenges in applying CTCAE criteria to retrospective medical records. This partially stemmed from the format of notes in the annotator tool. Because they were removed from the EMR system, there was an inability to incorporate contextual data typically used by clinicians to make severity assessments. Administration of as-needed medication, for example, was not always apparent in free-text clinical notes. Such ambiguities are inherent to retrospective reviews and reflect broader limitations in applying clinical grading systems to medical record data, but the iterative approach facilitated the creation of a detailed annotation guide and established a reliable methodology for future annotation efforts. The complexity of these clinical scenarios underscores the need for expert clinicians to remain closely involved with annotations when training AI models.

This study used a modified version of an open-source NLP pipeline, clinical text analysis and knowledge extraction system (cTAKES) [29], as a baseline for comparison against clinician annotations and our novel AI-based model in phenotyping constipation, mucositis, nausea, neuropathy, and vomiting. While cTAKES offers a valuable NLP solution for clinical text, its default configuration is computationally intensive and unsuitable for large-scale datasets. Our existing pipeline addressed this limitation by implementing a distributed processing pipeline capable of handling millions of clinical notes. It also further enhanced cTAKES by incorporating the human phenotype ontology to improve entity recognition and improving the negation annotator to refine accuracy in identifying negated findings [30,31]. This modified cTAKES pipeline served as a baseline for evaluating the performance of our novel transformer models.

With the revised annotation guide and further adjudication between annotators, F_1-scores could be assessed between our baseline NLP model and the clinician annotators. The F_1-score accounts for both sensitivity and recall of an NLP model. The existing off-the-shelf NLP model (cTAKES) was unable to reliably identify symptomatic adverse events of interest for pediatric oncology patients based on interrater reliability, Cohen kappa, and F_1-score analyses. This is clinically problematic, as reliable identification would be necessary for clinical work and to use this model for research purposes. Furthermore, the model is unable to identify symptom severity, further highlighting a need for the development of a fit-for-purpose novel NLP model which is proposed as stage 2 of this study.

Barriers and Lessons Learned

The first phase of this work provides valuable findings that justify continued research in this area. Our experiences as a newly developed transdisciplinary research team offer insights relevant to other teams that are beginning to integrate AI technologies into clinical research. Table 1 provides a review of our key challenges and the associated implications specific to this work.

Table . Key challenges, impact specific to this project, and facilitators for success in overcoming challenges.

Key challenge and implications	Facilitator
Require substantial data and computer science support	
Clinician scientists and researchers with limited knowledge in computer science and big data methodology	In-house Data Science and Biostatistical Unit with PhD- and master-level biostatisticians and data scientists
Cost associated with collaborative efforts and time of external experts	Free data science consultation for clinical investigators and internal pilot funding that allowed securement of external grants
Platform to manage very large data files and analyze millions of data-points in analyses	Enterprise-level strategic initiative developed a suite of tools and services for large-scale data analyses
Complexity of data structures between electronic health records and data warehouse	
Multiple versions of millions of clinical notes needed to be reviewed to select the correct version	Collaborative effort between PhD-prepared data scientist who coded and executed the tasks and clinicians who validated the output
Chemotherapy agents need to be identified and incorporated to patient selection as part of inclusion criteria	Senior data integration analysts created bespoke labeling system to identify all chemotherapy agents
Clinical data and associated metadata are stored in massive, discrete data tables	PhD-level data scientists with skills in variable identification, database management, and creation of relational databases
Extensive time for database creation and importing of large files to create a workable data model	Flexible timelines and expectations, mutual goals and understanding, and a data model that supports ongoing addition of new data elements
Inconsistent or misunderstood data labeling in the warehouse	
Validate research data to ensure consistency with clinical entry formats	Data extraction from the data warehouse and then validated against medical records by clinician staff
Extensive filtering of data elements to ensure integrity of data used for research purposes	Real-time collaboration between data scientists and clinician team members to refine and validate data filtering
Subjective nature of clinical interpretation of patient scenarios	
Lack of contextual data available for clinical symptom evaluation	Expert clinicians are required to annotate text for model training
Consistent method is needed to identify outcomes of interest to train AI[a] models	Creation of an annotation guide and consistent ontology
Multiple targets for annotation, creating a complicated validation process	Annotation tools and software as standard components of the digital lab environment
Transparent assessment of agreement for decision making between clinicians	Annotation review by expert clinicians to assess performance before model training and evaluation
Bridging distinct scientific domains to enable unified project execution	
Mutual understanding of priorities, feasibility, and methodology between data science and clinical research team members	Open, clear, and respectful communication; time to understand terminology and needs; flexible timelines and ongoing dedication from all research team members

[a]AI: artificial intelligence.

Barriers that slowed progress were primarily related to the inevitable learning curves encountered when embarking on a novel line of inquiry or acquiring a new skill set. The clinical researchers underestimated the time required to develop proficiency in these new methods and the time-intensive nature of interdisciplinary communication. Considerable effort was needed to understand how raw data are stored, transformed, and imported into a digital laboratory. This is noteworthy, not just for planning purposes for other teams, but also in understanding that data labeling and storage is unique to both the individual EMR platforms and the health institutions that use them. This makes the algorithm we have developed for identifying clinical notes specific to our institution and not likely directly transferrable to other sites. However, our methodology and approach can be replicated using institution-specific data elements and metadata, but this will require ongoing time investment.

Key challenges relating to the building of the digital laboratory related to the need for complex coding to identify appropriate clinical notes, the development of novel codes to identify chemotherapy agents, extensive data cleaning and refinement, and time-intensive data validation activities. Variations between how data are presented in the live, front-end version of EMR systems and how they are transformed and stored in the data warehouse created difficulty in translating between these views and ensuring the data accessed were accurate and correct. Logistic challenges related to data acquisition and organization arose from the size of datasets and tables because they included vast amounts of metadata in their raw form and extensive time for the data team to identify appropriate sources for importing. These challenges were overcome by continual partnership between clinical and data science team members and ensuring mutual understanding of needs before each phase of work.

Unfortunately, these unanticipated difficulties extended the project's timeline beyond what was initially anticipated.

Similarly, substantial time was dedicated to ensuring that the data science team comprehended the clinical scenarios underpinning this work. This reflexive exchange was critical for troubleshooting, planning data extraction, and conducting validation activities for model training. As a result, establishing the digital laboratory took significantly longer than anticipated, requiring adjustments to project timelines. Working meetings often focused on aligning terminology and achieving a mutual understanding of project milestones, underscoring the importance of interdisciplinary fluency. Finally, as with many research projects, cost considerations posed challenges. Incentives for clinicians to annotate notes could facilitate a larger group of trained annotators or dedicated research assistants, accelerating the process of achieving an adequate sample size for model training.

As AI becomes increasingly embedded in clinical practice, these models may become core components of clinical and research training programs, underscoring the need for ongoing interdisciplinary collaboration between data scientists and clinicians. These advancements signal an exciting future for AI-driven methodologies in improving patient care and advancing clinical research.

Facilitators and Necessary Infrastructure

Key facilitators to successfully completing the initial phase of our pilot work are also summarized in Table 1, matched to the implications of this project. They mainly included robust data science infrastructure and support in addition to flexibility of time and working toward mutual understandings. The DSBU and Arcus teams supplied critical expertise, technology, and financial resources, which were leveraged to scaffold this research project and are noted essential components of this type of collaborative work [25]. The clinical researchers defined a research question amenable to AI solutions, fostering a synergistic collaboration between the teams. A balance of funding and accessible resources is needed such that a researcher can either have access to the data science personnel or be able to contract with them for research purposes. These resources enabled our team to establish relationships, evaluate feasibility, and begin data harvesting to generate preliminary data that ultimately secured external funding. Once established, ongoing collaboration, shared priorities, and mutual commitment among team members facilitated a unified direction forward and long-term engagement in the project.

Clinician investigators who desire to engage with AI research need to have affiliation with an organization that has embraced and built an environment to support this work. Doing so requires the organization to make significant financial and personnel investments and overcome several hurdles and barriers to build a team that can orchestrate a large AI platform. Organizations must first determine that the clinical or financial benefits from an AI platform outweigh the upfront costs and long-term risks, requiring a long-term investment mindset [32]. Typical approaches involve identifying AI as a potential useful tool for improving the execution of daily operations and, once instituted,

can be used as a research platform. It is therefore primarily integrated to an organization as part of reengineering business processes [33], although there are cases of initiating AI platforms for research purposes as a primary objective. In either case, primary concerns and challenges are typically related to cost, confidentiality and security, data integration and system compatibility, and trustworthiness.

Upfront costs for AI infrastructure are high. Computational resources and power for initial training of algorithms are much higher than later simple execution of the models [34]. Lengthy time to production or to see benefit can deincentivize companies from investments [32], especially when considering that benefits and success are subject to time and other costly factors like computational power [35]. Computational resources, staff, personnel, training, and ongoing maintenance – including data audits, revised learning algorithms, ongoing data management, and updates – further add cost to AI adoption across a multitude of industries [32,34,35]. For these reasons, some smaller pharmaceutical companies, for example, have declined integration because the upfront costs are too high, unlike their larger counterparts, who see significant financial gain from even a small amount of process improvement [33]. However, taking strategic recommendations from end users, ensuring that there are well-defined problems amenable to AI-based solutions, and ensuring clear objectives for its use ensure valued return on investment [32,36].

Beyond cost, confidentiality and data security are of paramount concern, especially in health systems that are subject to stringent privacy laws and ethical considerations [6,34-36]. Safeguarding patient information requires legal counsel, information security personnel, and computer scientists. Similarly, these resources assist with concerns of data integration and system compatibility, ensuring that the AI platform can accept, synthesize, and augment existent data and work synergistically with programs already in use. For health care, this includes the EHR system, radiology software, mobile apps, pharmacy programs, billing systems, and scheduling programs.

Finally, uptake and integration of AI are halted if there is concern about the trustworthiness of the programs or if users – inclusive of clinicians, staff, and patients – have unfavorable views [33]. Known trust issues, algorithmic biases, lack of transparency, and unfairness have deincentivized health systems from adopting AI because it is viewed as an unreliable technology [32,33,37]. Further, health care providers often feel threatened by AI, worried that it will replace their positions. Concern for having AI handle the large, complex tasks of care, they will only perform simple tasks and lose skill over time or have to continuously learn about emerging technologies. Past successes of using AI in health care, however, indicate that it can augment, not replace, care practices. By reconsidering AI as an enabler, health care practices have seen improvements in diagnostics, radiology, analyzing data from wearable technologies, EHR monitoring, use of digital assistants, decision support systems, and breakthroughs in drug discovery, care models, streamlining workflow, and minimizing administrative burdens [32].

Conclusion

Despite these barriers and unexpected challenges, the results of this pilot study emphasize the transformative potential of AI in clinical research. The successful incorporation of AI into clinical workflows can replace the labor-intensive, time-consuming, and often imprecise process of manual data extraction. The model is being trained on clinical notes from a single institution, and since institutions use individualized note templates with templated free text, the NLP model may not be transferrable to other sites. However, future phases of this project can include data imported from diverse clinical sites to refine the model and expand its capability.

NLP, in particular, holds significant promise as a methodological innovation to address the limitations of extracting symptomatic adverse events from medical records. Future use of more lightweight models or integration of a large language model into the health system may further improve research efficiency. The development of a custom workflow that allowed for parallel processing of thousands of clinical notes simultaneously by a relatively small and inexpensive model. By improving research efficiency across health system networks, AI enables the rapid and consistent identification of symptomatic adverse events among patients treated for cancer. Leveraging these large patient cohorts, researchers can better explore the etiology, management, and mitigation of therapy-related toxicities.

Progress in harnessing the potential of AI in clinical research hinges on successful partnerships between clinical and data science researchers. This transparent account of our journey as a newly formed interdisciplinary team integrating AI into oncology research provides a framework, key lessons, and actionable recommendations for clinicians aiming to explore AI applications. Success is contingent on institutional support—both financial and logistical—and the assembly of a team of data and computer scientists with aligned priorities. Regardless of previous research experience, sufficient time must also be allocated to achieve mutual understanding, acquire new skills, build trust, and foster effective working relationships. By sharing our experience, we are hopeful that readers are empowered to take their first steps with greater confidence, mitigate delays we encountered, and chart a more efficient path toward advancing their own AI-driven research endeavors.

Conflicts of Interest

None declared.

References

1. Shao D, Dai Y, Li N, et al. Artificial intelligence in clinical research of cancers. Brief Bioinformatics 2022 Jan 17;23(1). [doi: 10.1093/bib/bbab523]
2. Levine AB, Schlosser C, Grewal J, Coope R, Jones SJM, Yip S. Rise of the machines: advances in deep learning for cancer diagnosis. Trends Cancer 2019 Mar;5(3):157-169. [doi: 10.1016/j.trecan.2019.02.002] [Medline: 30898263]
3. Vamathevan J, Clark D, Czodrowski P, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov 2019 Jun;18(6):463-477. [doi: 10.1038/s41573-019-0024-5] [Medline: 30976107]
4. Crossnohere NL, Elsaid M, Paskett J, Bose-Brill S, Bridges JFP. Guidelines for artificial intelligence in medicine: literature review and content analysis of frameworks. J Med Internet Res 2022 Aug 25;24(8):e36823. [doi: 10.2196/36823] [Medline: 36006692]
5. Bawack RE, Fosso Wamba S, Carillo KDA. A framework for understanding artificial intelligence research: insights from practice. JEIM 2021 Feb 4;34(2):645-678. [doi: 10.1108/JEIM-07-2020-0284]
6. Shah P, Kendall F, Khozin S, et al. Artificial intelligence and machine learning in clinical development: a translational perspective. NPJ Digit Med 2019;2(1):69. [doi: 10.1038/s41746-019-0148-3] [Medline: 31372505]
7. Carlotto A, Hogsett VL, Maiorini EM, Razulis JG, Sonis ST. The economic burden of toxicities associated with cancer treatment: review of the literature and analysis of nausea and vomiting, diarrhoea, oral mucositis and fatigue. Pharmacoeconomics 2013 Sep;31(9):753-766. [doi: 10.1007/s40273-013-0081-2] [Medline: 23963867]
8. Hooke MC, Linder LA. Symptoms in children receiving treatment for cancer-part I: fatigue, sleep disturbance, and nausea/vomiting. J Pediatr Oncol Nurs 2019;36(4):244-261. [doi: 10.1177/1043454219849576] [Medline: 31307321]
9. Tay N, Laakso EL, Schweitzer D, Endersby R, Vetter I, Starobova H. Chemotherapy-induced peripheral neuropathy in children and adolescent cancer patients. Front Mol Biosci 2022;9:1015746. [doi: 10.3389/fmolb.2022.1015746] [Medline: 36310587]
10. Thornton CP, Orgel E. Dose-limiting mucositis: friend or foe? Support Care Cancer 2023 Oct 7;31(10):617. [doi: 10.1007/s00520-023-08101-x] [Medline: 37804322]

XSL•FO

RenderX

11. Berman R, Davies A, Cooksley T, et al. Supportive care: an indispensable component of modern oncology. Clin Oncol (R Coll Radiol) 2020 Nov;32(11):781-788. [doi: 10.1016/j.clon.2020.07.020] [Medline: 32814649]

12. Scott EC, Jewell A. Supportive care needs of people with pancreatic cancer: a literature review. Cancer Nursing Practice 2019 Sep 3;18(5):35-43. [doi: 10.7748/cnp.2019.e1566]

13. Snaman J, McCarthy S, Wiener L, Wolfe J. Pediatric palliative care in oncology. J Clin Oncol 2020 Mar 20;38(9):954-962. [doi: 10.1200/JCO.18.02331] [Medline: 32023163]

14. Basch E, Reeve BB, Mitchell SA, et al. Development of the National Cancer Institute's patient-reported outcomes version of the common terminology criteria for adverse events (PRO-CTCAE). J Natl Cancer Inst 2014 Sep;106(9):dju244. [doi: 10.1093/jnci/dju244] [Medline: 25265940]

15. Trotti A, Colevas AD, Setser A, Basch E. Patient-reported outcomes and the evolution of adverse event reporting in oncology. J Clin Oncol 2007 Nov 10;25(32):5121-5127. [doi: 10.1200/JCO.2007.12.4784] [Medline: 17991931]

16. Merrow M, King N. Optimizing antiemetic therapy for children undergoing chemotherapy. J Pediatr Nurs 2022;66:136-142. [doi: 10.1016/j.pedn.2022.06.006] [Medline: 35759994]

17. Smith EML, Kuisell C, Cho Y, et al. Characteristics and patterns of pediatric chemotherapy-induced peripheral neuropathy: a systematic review. Cancer Treat Res Commun 2021;28:100420. [doi: 10.1016/j.ctarc.2021.100420] [Medline: 34225104]

18. Miller TP, Li Y, Kavcic M, et al. Accuracy of adverse event ascertainment in clinical trials for pediatric acute myeloid leukemia. J Clin Oncol 2016 May 1;34(13):1537-1543. [doi: 10.1200/JCO.2015.65.5860] [Medline: 26884558]

19. Miller TP, Marx MZ, Henchen C, et al. Challenges and barriers to adverse event reporting in clinical trials: a children's oncology group report. J Patient Saf 2022 Apr 1;18(3):e672-e679. [doi: 10.1097/PTS.0000000000000911] [Medline: 34570002]

20. Hong JC, Fairchild AT, Tanksley JP, Palta M, Tenenbaum JD. Natural language processing for abstraction of cancer treatment toxicities: accuracy versus human experts. JAMIA Open 2020 Dec;3(4):513-517. [doi: 10.1093/jamiaopen/ooaa064] [Medline: 33623888]

21. Li A, da Costa WL Jr, Guffey D, et al. Developing and optimizing a computable phenotype for incident venous thromboembolism in a longitudinal cohort of patients with cancer. Res Pract Thromb Haemost 2022 May;6(4):e12733. [doi: 10.1002/rth2.12733] [Medline: 35647478]

22. Mashima Y, Tamura T, Kunikata J, et al. Using natural language processing techniques to detect adverse events from progress notes due to chemotherapy. Cancer Inform 2022;21:11769351221085064. [doi: 10.1177/11769351221085064] [Medline: 35342285]

23. Muñoz AJ, Souto JC, Lecumberri R, et al. Development of a predictive model of venous thromboembolism recurrence in anticoagulated cancer patients using machine learning. Thromb Res 2023 Aug;228:181-188. [doi: 10.1016/j.thromres.2023.06.015] [Medline: 37348318]

24. Zitu MM, Zhang S, Owen DH, Chiang C, Li L. Generalizability of machine learning methods in detecting adverse drug events from clinical narratives in electronic medical records. Front Pharmacol 2023;14:1218679. [doi: 10.3389/fphar.2023.1218679] [Medline: 37502211]

25. Flood EL, Schweig L, Froh EB, et al. The Arcus experience: bridging the data science gap for nurse researchers. Nurs Res 2024;73(5):406-412. [doi: 10.1097/NNR.0000000000000748] [Medline: 38773838]

26. El-khalek HA, Aziz RF, Morgan ES. Identification of construction subcontractor prequalification evaluation criteria and their impact on project success. Alexandria Engineering Journal 2019 Mar;58(1):217-223. [doi: 10.1016/j.aej.2018.11.010]

27. Gupta K, Walton R, Kataria SP. Chemotherapy-induced nausea and vomiting: pathogenesis, recommendations, and new trends. Cancer Treat Res Commun 2021;26:100278. [doi: 10.1016/j.ctarc.2020.100278] [Medline: 33360668]

28. Common terminology criteria for adverse events v50. National Cancer Institute Cancer Therapy Evaluation Program. 2017. URL: https://ctep.cancer.gov/protocoldevelopment/electronic_applications/ctc.htm#ctc_50 [accessed 2023-07-11]

29. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507-513. [doi: 10.1136/jamia.2009.001560] [Medline: 20819853]

30. Thayer J, Pennington JW. Fault-tolerant, distributed, and scalable natural language processing with ctakes. Presented at: AMIA Annual Symposium; Nov 16-20, 2019; Washington, DC, USA.

31. Daniali M, Galer PD, Lewis-Smith D, et al. Enriching representation learning using 53 million patient notes through human phenotype ontology embedding. Artif Intell Med 2023 May;139:102523. [doi: 10.1016/j.artmed.2023.102523] [Medline: 37100502]

32. Esmaeilzadeh P. Challenges and strategies for wide-scale artificial intelligence (AI) deployment in healthcare practices: a perspective for healthcare organizations. Artif Intell Med 2024 May;151:102861. [doi: 10.1016/j.artmed.2024.102861] [Medline: 38555850]

33. Kulkov I. The role of artificial intelligence in business transformation: a case of pharmaceutical companies. Technol Soc 2021 Aug;66:101629. [doi: 10.1016/j.techsoc.2021.101629]

34. Shaw J, Rudzicz F, Jamieson T, Goldfarb A. Artificial intelligence and the implementation challenge. J Med Internet Res 2019 Jul 10;21(7):e13659. [doi: 10.2196/13659] [Medline: 31293245]

35. Martínez-García M, Hernández-Lemus E. Data integration challenges for machine learning in precision medicine. Front Med (Lausanne) 2021;8:784455. [doi: 10.3389/fmed.2021.784455] [Medline: 35145977]

36. Sinha S, Lee YM. Challenges with developing and deploying AI models and applications in industrial systems. Discov Artif Intell 2024;4(1):55. [doi: 10.1007/s44163-024-00151-2]

37. Wubineh BZ, Deriba FG, Woldeyohannis MM. Exploring the opportunities and challenges of implementing artificial intelligence in healthcare: a systematic literature review. Urol Oncol 2024 Mar;42(3):48-56. [doi: 10.1016/j.urolonc.2023.11.019] [Medline: 38101991]

Abbreviations

AI: artificial intelligence
cTAKES: clinical text analysis and knowledge extraction system
CTCAE: Common Terminology Criteria for Adverse Events
DSBU: Data Science and Biostatistics Unit
EHR: electronic health record
EMR: electronic medical record
HIPAA: Health Insurance Portability and Accountability Act
ICD-10: *International Classification of Diseases, Tenth Revision*
ICD-9: *International Classification of Diseases, Ninth Revision*
IRB: Institutional Review Board
NLP: natural language processing

XSL•FO
RenderX

Extracting Knowledge From Scientific Texts on Patient-Derived Cancer Models Using Large Language Models: Algorithm Development and Validation Study

Jiarui Yao[1,2*], PhD; Zinaida Perova[3*], PhD; Tushar Mandloi[3], MSc; Elizabeth Lewis[3], MSc; Helen Parkinson[3], PhD; Guergana Savova[1,2], PhD

[1]Computational Health Informatics Program, Boston Children's Hospital, 401 Park Drive, Boston, MA, United States

[2]Harvard Medical School, Boston, MA, United States

[3]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Jiarui Yao, PhD
Computational Health Informatics Program, Boston Children's Hospital, 401 Park Drive, Boston, MA, United States

Abstract

Background: Patient-derived cancer models (PDCMs) have become essential tools in cancer research and preclinical studies. Consequently, the number of publications on PDCMs has increased significantly over the past decade. Advances in artificial intelligence, particularly in large language models (LLMs), offer promising solutions for extracting knowledge from scientific literature at scale.

Objective: This study aims to investigate LLM-based systems, focusing specifically on prompting techniques for the automated extraction of PDCM-related entities from scientific texts.

Methods: We explore 2 LLM-prompting approaches. The classic method, direct prompting, involves manually designing a prompt. Our direct prompt consists of an instruction, entity-type definitions, gold examples, and a query. In addition, we experiment with a novel and underexplored prompting strategy—soft prompting. Unlike direct prompting, soft prompts are trainable continuous vectors that learn from provided data. We evaluate both prompting approaches across state-of-the-art proprietary and open LLMs.

Results: We manually annotated 100 abstracts of PDCM-relevant papers, focusing on PDCM papers with data deposited in the CancerModels.Org platform. The resulting gold annotations span 15 entity types for a total 3313 entity mentions, which we split across training (2089 entities), development (542 entities) and held-out, eye-off test (682 entities) sets. Evaluation includes the standard metrics of precision or positive predictive value, recall or sensitivity, and F_1-score (harmonic mean of precision and recall) in 2 settings: an exact match setting, where spans of gold and predicted annotations have to match exactly, and an overlapping match setting, where the spans of gold and predicted annotations have to overlap. GPT4-o with direct prompting achieved F_1-scores of 50.48 and 71.36 for exact and overlapping match settings, respectively. In both evaluation settings, LLaMA3 soft prompting improved performance over direct prompting (F_1-score from 7.06 to 46.68 in the exact match setting; and 12.0 to 71.80 in the overlapping evaluation setting). Results with LLaMA3 soft prompting are slightly higher than GPT4-o direct prompting in the overlapping match evaluation setting.

Conclusions: We investigated LLM-prompting techniques for the automatic extraction of PDCM-relevant entities from scientific texts, comparing the traditional direct prompting approach with the emerging soft prompting method. In our experiments, GPT4-o demonstrated strong performance with direct prompting, maintaining competitive results. Meanwhile, soft prompting significantly enhanced the performance of smaller open LLMs. Our findings suggest that training soft prompts on smaller open models can achieve performance levels comparable to those of proprietary very large language models.

KEYWORDS

Introduction

Patient-derived cancer models (PDCMs) are created from a patient's own tumor sample and capture the complexity of human tumors to enable more accurate, personalized drug development and treatment selection. These models, including patient-derived xenografts (PDXs), organoids, and cell lines, allow researchers to test treatments and identify the most effective therapies, and have emerged as indispensable tools in both cancer research and precision medicine. The US National Institutes of Health (NIH) have made significant investments in the generation and characterization of these models, with more than US $3 billion dedicated to active grants referencing PDCMs with a component of their research based on data extracted from the NIH RePORTER [1] for fiscal year 2024 alone. The number of publications using PDCMs continues to increase generating substantial and rich metadata and data that require standardization, harmonization, and integration to maximize the impact of these models and their associated data within the research and clinical communities. CancerModels.Org platform [2] serves as a unified gateway to the largest collection of PDCMs and related data. It empowers researchers and clinicians to discover suitable models for testing research hypotheses, conducting large-scale drug screenings, and advancing precision medicine initiatives. Extraction of PDCM-relevant knowledge and its harmonization within CancerModels.Org is essential to ensure that basic and translational researchers, bioinformaticians, and tool developers have access to PDCM knowledge. While manual curation of publications ensures high accuracy when performed by domain experts, it is time-consuming and labor-intensive. Thus, a more streamlined and efficient knowledge acquisition method is needed to address the growing demand within the scientific community for the timely availability of the PDCM metadata and its associated data.

In parallel, large language models (LLMs) [3-5] often referred to as generative artificial intelligence (AI) systems are trained on vast amounts of data and have demonstrated impressive capabilities in the health care domain [6-8]. Researchers have studied the use of LLMs in addressing various tasks related to health care such as diagnosing conditions [9,10], clinical decision support [11], answering patient questions [12], and medical education [13,14]. It has been shown that LLMs can extract meaningful information from texts [15-17].

In this work, we explore LLM-prompting techniques with the goal of extracting knowledge from PDCM-relevant scholarly publications. We focus on the classic direct prompting [4] and the underexplored soft prompting [18] with state-of-the-art (SOTA) proprietary and open LLMs. Our experimental results provide insights into selecting the optimal prompting methods for specific tasks. The contributions of this paper are:

1. Studying the feasibility of SOTA LLMs as oncology knowledge extractors for PDCM-relevant information from scholarly scientific literature.
2. Creating a manually curated gold dataset spanning 15 entity types for a total 3313 entity mentions from 100 abstracts of PDCM-relevant papers.
3. Researching and comparing, to our knowledge for the first time, direct versus soft prompting techniques for oncology knowledge extraction, specifically PDCM-relevant information from scholarly scientific literature.

Methods

Concepts

We define "knowledge" as entities of interest to researchers working with PDCMs in the cancer research field. For example, the patient's diagnosis provides a reference point to confirm that a PDCM faithfully recapitulates the biology of the original tumor and is essential for ensuring the model's relevance and reliability in studies of cancer progression or treatment response. Thus, "diagnosis" is important to understand the model's characteristics in the context of patient's disease. The patient's age can significantly affect the molecular and genetic characteristics of the tumor. For example, pediatric cancers often have distinct genetic drivers and tumor microenvironments compared to cancers in older adults. In addition, age-related biological factors, such as immune system, metabolism, and hormone levels, influence how a tumor responds to treatments. Thus, knowing the patient's age is imperative for predictive accuracy of the model in preclinical testing and relevance of research findings. Therefore, we selected 15 most commonly used CancerModels.Org data model attributes (Table 1), which include the attributes defined in the minimal information standard for patient-derived xenograft models [19] and the draft minimal information standard for in vitro models [20].

Table . Entity definitions based on the CancerModels.Org data model with examples and interannotator agreement F_1-scores in the exact match setting that requires the spans of the annotators to match exactly.

Entity type	Definition	Example	IAA[a]
diagnosis	Diagnosis at the time of collection of the patient tumor used in the cancer model	TNBC[b]	61.67
age_category	Age category of the patient at the time of tissue sampling	Adult, pediatric	60
genetic_effect	Any form of chromosomal rearrangement or gene-level changes	Missense, amplification	57.67
model_type	Type of patient-derived model	PDX[c], organoid	53.33
molecular_char	Data or assay generated from or performed on the model in this study	RNA sequencing, whole-exome sequencing	54.33
biomarker	Gene, protein or biological molecule identified in or associated with patient's/model's tumor	BRCA1[d], IDH[e], epidermal growth factor receptor 2	61.33
treatment	Treatment received by the patient or tested on the model	Surgery, chemotherapy, PARP-inhibitor	55.67
response_to_treatment	Effect of the treatment on the patient's tumor or model	Progression-free survival, reduced tumor growth	55
sample_type	The type of material used to generate the model or how this material was obtained	Tissue fragment, autopsy	49
tumor_type	Collected tumor type used for generating the model	Primary, recurrent	49.67
cancer_grade	Quantitative or qualitative grade reflecting how quickly the cancer is likely to grow	Grade 1, low-grade	42
cancer_stage	Information about the cancer's extent in the body according to specific type of cancer staging system	TNM[f] system, T0, stage I	59.33
clinical_trial	The type of clinical trial or Clinicaltrials.org identifier	Phase II, prospective randomized clinical trials	60.67
host_strain	The name of the mouse host strain where the tissue sample was engrafted for generating the PDX model	NOD-SCID[g]	61.67
model_id	ID of the patient-derived cancer model generated in this study	PHLC402	100

[a]IAA: interannotator agreement.

[b]TNBC: triple-negative breast cancer.

[c]PDX: patient-derived xenograft.

[d]BRCA1: breast cancer gene 1.

[e]IDH: isocitrate dehydrogenase.

[f]TNM: tumor node metastasis.

[g]NOD-SCID: nonobese diabetic severe combined immunodeficiency.

Corpus

We used 100 abstracts to develop the gold-standard corpus annotated for the 15 entities (Table 1). The abstracts were chosen from publications linked to the PDCMs submitted to CancerModels.Org platform. They were selected to cover all 3 types of models in the resource-PDXs, organoids, and cell lines. The final corpus is available on GitHub (see Data and Code Availability section).

Three annotators (ZP, TM, and EL) independently labeled entities in all 100 abstracts for a total of 40 hours. The annotation quality was tracked through interannotator agreement (IAA), a measure of agreement between each annotation produced by different annotators working on the same dataset. The IAA is an indication of how difficult the task is for humans and it becomes the target for system development. We used pairwise F_1-score as the IAA metric [21] in the exact match setting that

XSL•FO

RenderX

requires the spans of the annotators to match exactly. We computed the agreement between each pair of annotators and averaged across the 3 sets of scores. The final IAA for each entity type is reported in Table 1. The IAA range is 42‐100 indicating moderate agreement. Note that the lowest agreement is for low occurrence entity types, for example, cancer_grade has only 8 instances with 42 IAA. These low-frequency entity types are more likely to be overlooked by the human experts as annotation is a cognitively demanding task. Thus, to ensure a high-quality gold-standard dataset, we overlayed the single

annotations with an adjudication step, where the annotators discussed annotation disagreements and potential missed annotations to come to final joint decisions. The resulting gold dataset spans 15 entity types for a total 3313 entity mentions (refer Table 2 for distributions) was split into training, development, and test sets in the standard 60:20:20 ratio. The train set was used for creating entity extraction algorithms, the development set for refining the algorithms, and the test set for the final evaluation.

Table . Distribution of entity type annotations across training, development, and test sets.

Entity type	Training, n	Development, n	Test, n	Total, n
diagnosis	362	122	114	598
age_category	19	0	0	19
genetic_effect	69	20	33	122
model_type	326	114	110	550
molecular_char	128	37	46	211
biomarker	503	118	163	784
treatment	426	77	130	633
response_to _treatment	99	21	28	148
sample_type	22	8	7	37
tumor_type	61	19	28	108
cancer_grade	6	1	1	8
cancer_stage	7	1	4	12
clinical_trial	35	2	4	41
host_strain	9	0	7	16
model_id	17	2	7	26
Total	2089	542	682	3313

Prompting Methods

Various prompting techniques have been proposed since the emergence of LLMs [22-25]. At a high level, these prompting techniques can be divided into 2 categories, direct prompting [4] and soft prompting [18,24,26] . The main difference between the two methods is the prompt representation, that is whether the prompt consists of human language words or vectors (Figure 1). Direct prompting (or discrete prompting) is the most intuitive and now classic prompting method where users directly interact with LLMs using natural language. For example, a user may ask ChatGPT to "Write a thank you note to an old friend of my parents"; in this case, the text within the quotation marks is a discrete prompt. Soft prompting (or continuous prompting) uses

a machine learning approach to train a sequence of continuous vectors, which are the "virtual tokens" of the prompt. It is worth noting that soft prompting differs from fine tuning. With soft prompting, the LLM parameters are not updated, only the soft prompt parameters are adjusted. In contrast, finetuning requires to update the parameters of the entire LLM, and therefore needs more computation resources. Both prompting techniques have their advantages and disadvantages. Compared to direct prompting, soft prompting does not require the tedious process of manually creating prompts; however, it requires some labeled data to train the prompt. In this work, we explore both direct and soft prompting as we aim to explore the latest developments in LLMs and prompting techniques for the task of extracting PDCM entities from abstracts of academic papers.

Figure 1. Illustration of the 2 prompting methods. In direct prompting, a prompt contains a sequence of words. In soft prompting, a prompt consists of a list of vectors. LLM: large language model.
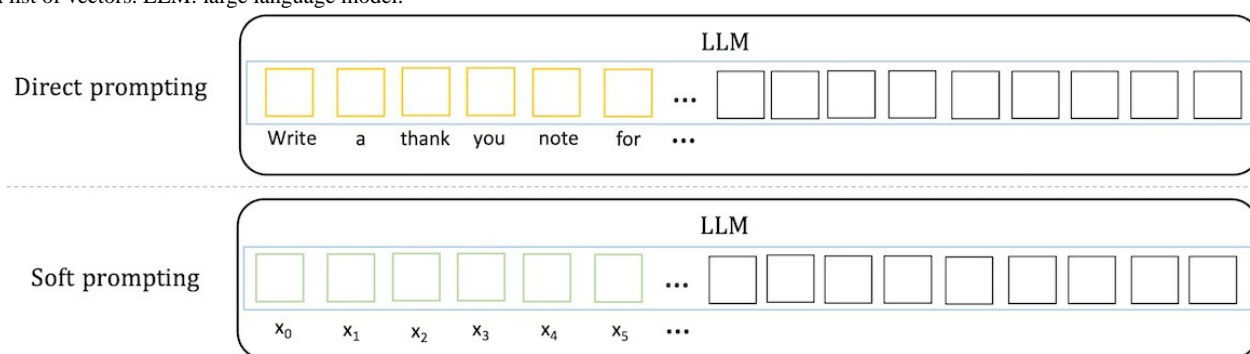
Direct Prompting

When asking LLMs to extract entities such as diagnoses or biomarkers, the most intuitive way is to ask LLMs to output the entities directly. In example 1 below, "ALK" is a biomarker entity. One may expect the model to output *{"biomarker" [ALK]}*. However, we note that the string "ALK" is mentioned multiple times in this example text, therefore it is not clear which "ALK" the model refers to. To get the most precise extraction to facilitate a more fine-grained analysis, we instruct the model to output the offsets of the specific mentions in the text (ie, the spans). For instance, if the model gives us *[(48, 51, "ALK," biomarker), (323, 326, "ALK," biomarker), …]*, we know that from character 48 to character 51, there is a biomarker entity, "ALK." Similarly, we can find another biomarker entity "ALK" at position 323 ‐ 326.

> *Example 1:*
>
> *Oncogenic fusion of anaplastic lymphoma kinase (ALK) with echinoderm microtubule associated protein like 4 protein or other partner genes occurs in 3 to 6% of lung adenocarcinomas. Although fluorescence in situ hybridization (FISH) is the accepted standard for detecting anaplastic lymphoma receptor tyrosine kinase gene (ALK) gene rearrangement that gives rise to new fusion genes, not all ALK FISH-positive patients respond to ALK inhibitor therapies.*

We started our exploration by designing prompts with an explicit instruction to specify the character offsets of each entity along with the entity text and type (eg, *48, 51, "ALK", biomarker*). However, our experiments show that it was challenging for the LLM to output the correct character offsets, a seemingly straightforward task (all the model needs to do is to count the number of characters); however, the complexity of this seemingly straightforward task is likely due to the LLM's way of breaking words outside its vocabulary into so-called word pieces, for example, "organoid" is broken down into 2 word pieces "organ" and "-oid." Considering that LLMs were trained as generative models [3,4], we subsequently cast the entity extraction task as a generation task, where we instructed the model to mark the entities with XML tags. For instance, if the model outputs "Oncogenic fusion of anaplastic lymphoma kinase (<biomarker>ALK</biomarker>) with echinoderm microtubule …," then postprocessing the output with regular expressions would find the exact position of "ALK" in the text. Specifically,

we asked the LLMs to mark the start and end of an entity with <entity_type> and </entity_type> tags, where entity_type is a placeholder for the specific entity type, such as biomarker or treatment (refer Table 1 for the full list).

Soft Prompting

Designing the direct prompts manually could be time-consuming and minor changes in the prompt language could lead to drastic changes in the model performance [24,27]. On the other hand, soft prompting requires some amount of gold data for its training and annotating gold data by domain experts could also be time-consuming. Fortunately, only a small set of labeled data are needed to train soft prompts. As described above, we created a gold dataset, which we used for training and evaluating our soft prompting approach.

There are a few soft prompting methods, the difference usually lies in how the prompt vectors are initialized and learned. Prompt-tuning [18] is a technique that learns the prompt by adding a list of virtual tokens (ie, vectors) in front of the input, where the virtual tokens can be randomly initialized, or drawn from a pretrained word embedding [28] set. Another method is P-tuning [24], which uses small neural networks such as feedforward neural networks [29] (multilayer perceptron) or recurrent neural networks [30] (eg, long-short term memory) as the prompt encoder to learn the prompt. Only the parameters in the prompt encoder are updated during training, while the weights in the LLMs remain frozen. In our experiments, we found P-tuning did not always converge to an optimal solution for our task perhaps due to the random initialization of the vectors rather than using carefully pretrained word embeddings. Therefore, we focused on prompt-tuning in this work. Following Lester et al [18], we initialized the vectors in the prompt with the embeddings of the label words in the entity type set (Table 1).

The standard approach for entity extraction in natural language processing is via token classification [31]. Concretely, a classifier is trained to predict the label for each token in a sentence according to a predefined label set. Additionally, each label is prepended with a B or I prefix to indicate the entity's Beginning or Inside mention, respectively. An example is provided in Figure 2. "Ewing sarcoma" is an entity mention of the diagnosis type. Thus "Ewing" and "sarcoma" are labeled as "Diagnosis," while all other tokens are labeled as "O," meaning they are Outside of an entity. To be more precise, "Ewing" is at the beginning of the diagnosis entity, and "sarcoma" is inside

of the entity, so they are labeled as "B-Diagnosis" and "I-Diagnosis," respectively.

To summarize, we trained a multiclass classifier for the soft-prompting training step. There are 15 entity types in our dataset, therefore there are 15×2+1=31 labels for token classification, with one extra label for "O."

Figure 2. An example of entity extraction as token classification.
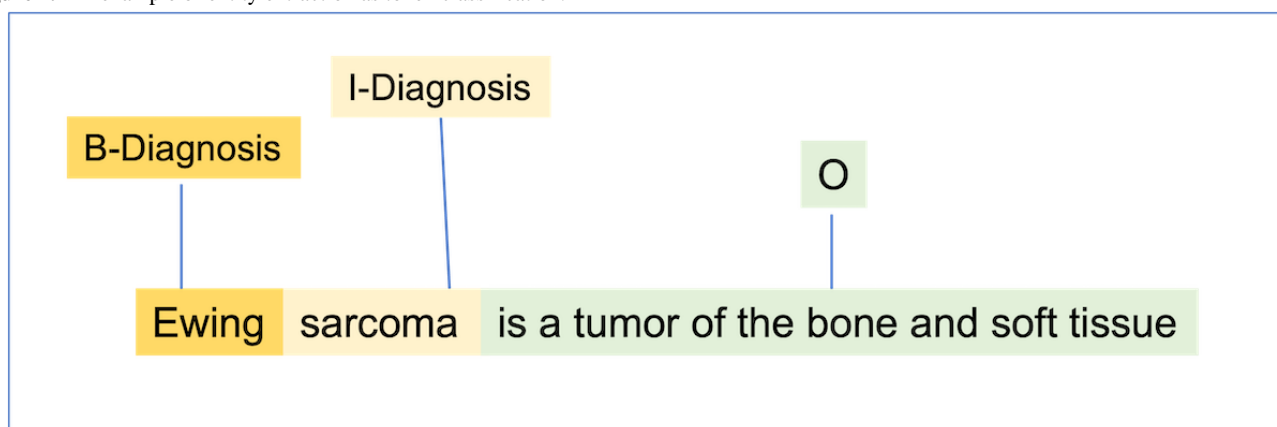
Experimental Set-Up

For efficiency purposes, we used Apache cTAKES [32] to split an abstract into sentences which were then passed to the LLMs to extract entities one sentence at a time. Our direct prompt included the instruction, the definition of each entity type, 5 examples (few-shot in-context learning) and the query (the sentence). The in-context learning [4] is a common practice in LLM prompting and has consistently shown improved results as the examples guide the LLM onto an optimal path [33,34]. Figure 3 presents our prompt template, and examples are in Multimedia Appendix 1.

Figure 3. Prompt template used in direct prompting.
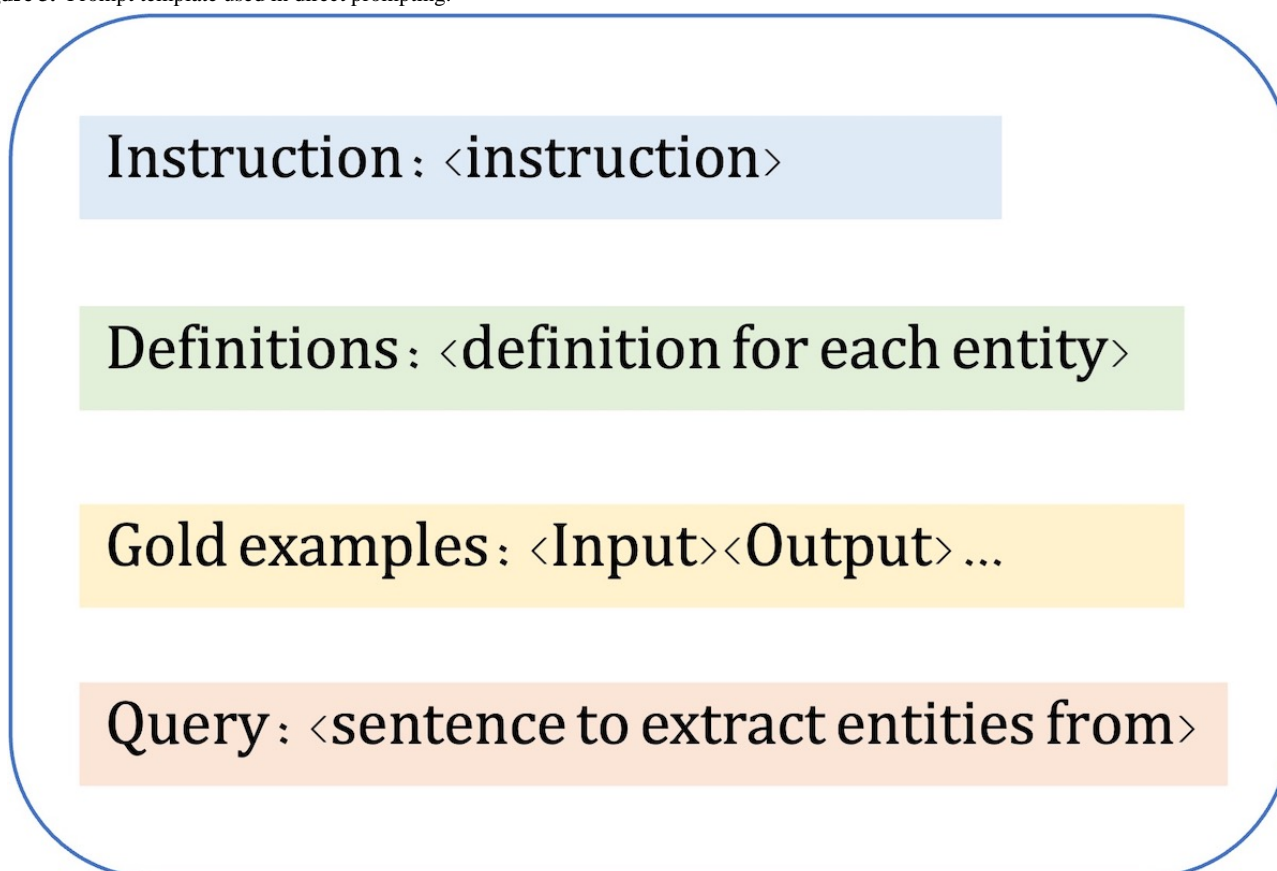
When choosing the LLMs, we used GPT-4o [35], one of the most powerful proprietary LLMs at the time of this study, and SOTA open LLMs from the LLaMA3 family [36], including LLaMA3.1 70B, LLaMA3.1 8B, LLaMA3.2 1B, and LLaMA3.2

3B. We did not use GPT-4o or LLaMA3.1 70B to train the soft prompts due to computational resource limitations; thus, our work here is representative of the computational environment in the vast majority of academic medical centers and research

labs. We set the soft prompt length to 30. We trained the soft prompt on the training set for 50 epochs with a learning rate of 0.001. Hyperparameters were tuned on the development set using the LLaMA3.1 8B model.

We report the evaluation results on the test set in the next section. In addition, we apply 5-fold cross-validation and report the average scores with SDs. For the 5-fold cross-validation, we excluded the 3 abstracts used to sample the gold examples for direct prompting and split the remaining 97 abstracts into 5 folds with a 20:20:20:20:17 ratio. For direct prompting, we ran the model on each fold and reported the average scores. For soft prompting, we set aside one fold as the test set and trained the soft prompts on the remaining 4 folds.

Results

We used the standard evaluation metrics of precision or positive predictive value, recall or sensitivity, and F_1-score (the harmonic mean of precision and recall) with 2 evaluation settings: "exact match" setting requires the span output from the model to exactly match the span of the gold annotation, and "overlapping match" setting allows the model to get partial credit if its extraction overlaps the spans in the gold annotation. For example, the model may extract "patient-derived tumor xenograft (PDX)" as a model_type entity, while the gold annotation is "patient-derived tumor xenograft (PDX) models." Under the "exact match" setting, "patient-derived tumor xenograft (PDX)" is NOT a match to "patient-derived tumor xenograft (PDX) models;" while under the "overlapping match" setting, it is a match since the spans overlap.

Tables 3 and 4 show the evaluation results on the test set. In Table 3, we can see that under the "exact match" setting, GPT-4o direct prompting achieves the highest F_1-score of 50.48. The performances of the LLaMA3 family models drop as the model size decreases, with F_1-score from 38.40 for the 70B model to 6.78 for the 1B model. However, there is a consistent improvement in F_1-scores with soft prompting over direct prompting. For the LLaMA3.2 models, the performance of the 3B model improves significantly, with F_1-score rising from 7.06 to 46.68 F_1-score—more than 8 points higher than the LLaMA3.1-70B model with direct prompting (F_1-score=38.40), despite the substantial difference in model size.

Table . Evaluation results on the test set (exact match) as precision or positive predictive value, recall or sensitivity, and F_1-score (harmonic mean of precision and recall).

Exact match	Precision	Recall	F_1-score
Direct prompting			
GPT-4o	56.09	45.89[a]	50.48[a]
LLaMA3.1-70B	57.27[a]	28.89	38.40
LLaMA3.1-8B	35.80	18.48	24.37
LLaMA3.2-3B	25.23	4.10	7.06
LLaMA3.2-1B	23.48	3.96	6.78
Soft prompting			
LLaMA3.1-8B	47.17	45.75	46.44
LLaMA3.2-3B	47.30[a]	46.09[a]	46.68[a]
LLaMA3.2-1B	46.19	45.01	45.59

[a]These are the best results.

XSL•FO

RenderX

Table . Evaluation results on the test set (overlapping match) as precision or positive predictive value, recall or sensitivity, and F_1-score (harmonic mean of precision and recall).

Overlapping match	Precision	Recall	F_1-score
Direct prompting			
GPT-4o	76.96	66.52[a]	71.36[a]
LLaMA3.1-70B	77.95[a]	43.99	56.24
LLaMA3.1-8B	50.54	27.49	35.61
LLaMA3.2-3B	41.03	7.03	12.00
LLaMA3.2-1B	35.34	6.01	10.28
Soft prompting			
LLaMA3.1-8B	71.19	70.53	70.86
LLaMA3.2-3B	72.05[a]	71.55[a]	71.80[a]
LLaMA3.2-1B	70.38	70.48	70.42

[a]These are the best results.

Similar trends are observed in Table 4 under the "overlapping match" evaluation. GPT4-o shows an F_1-score performance of 71.36, maintaining its position as the top performer for direct prompting. The 3 smaller LLaMA3 models continue to benefit from soft prompting, with the LLaMA3.2 3B model achieving slightly higher score than GPT4-o with direct prompting (F_1-scores of 71.80 vs 71.36).

Tables 5 and 6 present the results with 5-fold cross-validation under "exact match" and "overlapping" match respectively. Once again, our observations indicate that with soft prompting, the smaller LLaMA models attain performance levels comparable to GPT-4o.

Table . Five-fold cross-validation results (exact match) as precision or positive predictive value, recall or sensitivity, and F_1-score (harmonic mean of precision and recall).

Exact match	Precision	Recall	F_1-score
Direct prompting, mean (SD)			
GPT-4o	60.73 (2.69)	49.92 (3.46)	54.75 (2.84)
LLaMA3.1-70B	57.56 (1.53)	31.70 (1.24)	40.87 (1.25)
LLaMA3.1-8B	38.29 (3.29)	20.57 (2.18)	26.75 (2.61)
LLaMA3.2-3B	27.01 (3.20)	5.25 (0.80)	8.80 (1.29)
LLaMA3.2-1B	9.84 (5.98)	0.74 (0.47)	1.38 (0.87)
Soft prompting, mean (SD)			
LLaMA3.1-8B	51.76 (3.09)	50.21 (2.24)	50.94 (2.55)
LLaMA3.2-3B	50.99 (2.43)	49.54 (2.98)	50.24 (2.53)
LLaMA3.2-1B	49.34 (3.47)	49.98 (3.19)	49.13 (3.10)

XSL•FO
RenderX

Table . Five-fold cross-validation results (overlapping match) as precision or positive predictive value, recall or sensitivity, and F_1-score (harmonic mean of precision and recall).

| Overlapping match	Precision	Recall	F_1-score
Direct prompting, mean (SD)			
GPT-4o	77.82 (2.54)	67.52 (2.17)	72.28 (1.88)
LLaMA3.1-70B	78.01 (1.14)	47.77 (0.71)	59.25 (0.81)
LLaMA3.1-8B	52.75 (3.02)	29.78 (2.60)	38.04 (2.84)
LLaMA3.2-3B	42.42 (2.89)	8.64 (1.09)	14.34 (1.64)
LLaMA3.2-1B	22.09 (5.74)	1.67 (0.54)	3.10 (0.99)
Soft prompting, mean (SD)			
LLaMA3.1-8B	73.78 (3.09)	73.77 (1.25)	73.75 (2.06)
LLaMA3.2-3B	73.48 (1.97)	73.51 (1.11)	73.48 (1.31)
LLaMA3.2-1B	71.51 (3.43)	73.25 (2.46)	72.34 (2.63)

Discussion

Principal Findings

Our experiments demonstrate that soft prompting, a relatively underexplored aspect of LLM prompting, can significantly enhance the performance of smaller LLMs. The 3 LLaMA models exhibit comparable performance under soft prompting (an F_1-score of 46 in the exact match setting, and 70 in the overlapping match setting). These results are particularly promising results given the limited training data, consisting of 60 abstracts with 2089 entity mentions. Please note that all F_1-scores mentioned in this section refer to the F_1-scores on the test set.

How much data is needed to train the soft prompt? To answer this question, we trained LLaMA3.2 1B model, the smallest model used in this work, with different amounts of training data.

Figure 4 shows the relation between the proportion of training data and the F_1-scores on the test set (overlapping match). Solid performance was achieved with only 5% of the training data (26 sentences from 3 abstracts). With 25% of the training data (129 sentences from 15 abstracts), the model achieved an F_1-score of 68.21, only 2 points lower than using the entire training set, and only 3 points lower than GPT4-o with direct prompting. Despite the impressive performance of GPT4-o direct prompting, one potential issue is that not all data used in biomedical research can be sent to proprietary models such as GPT or the Gemini family models [8] via public application programming interfaces. That is, for applications using real patient data that require Health Insurance Portability and Accountability Act–compliant platforms, our findings demonstrate that achieving performance comparable to proprietary LLMs such as GPT4-o remains feasible through soft prompting. However, this approach necessitates a tradeoff, requiring a small set of labeled data for optimal effectiveness.

XSL•FO

RenderX

Figure 4. Performance curve of the LLaMA3.2 1B model as the size of training data increases.
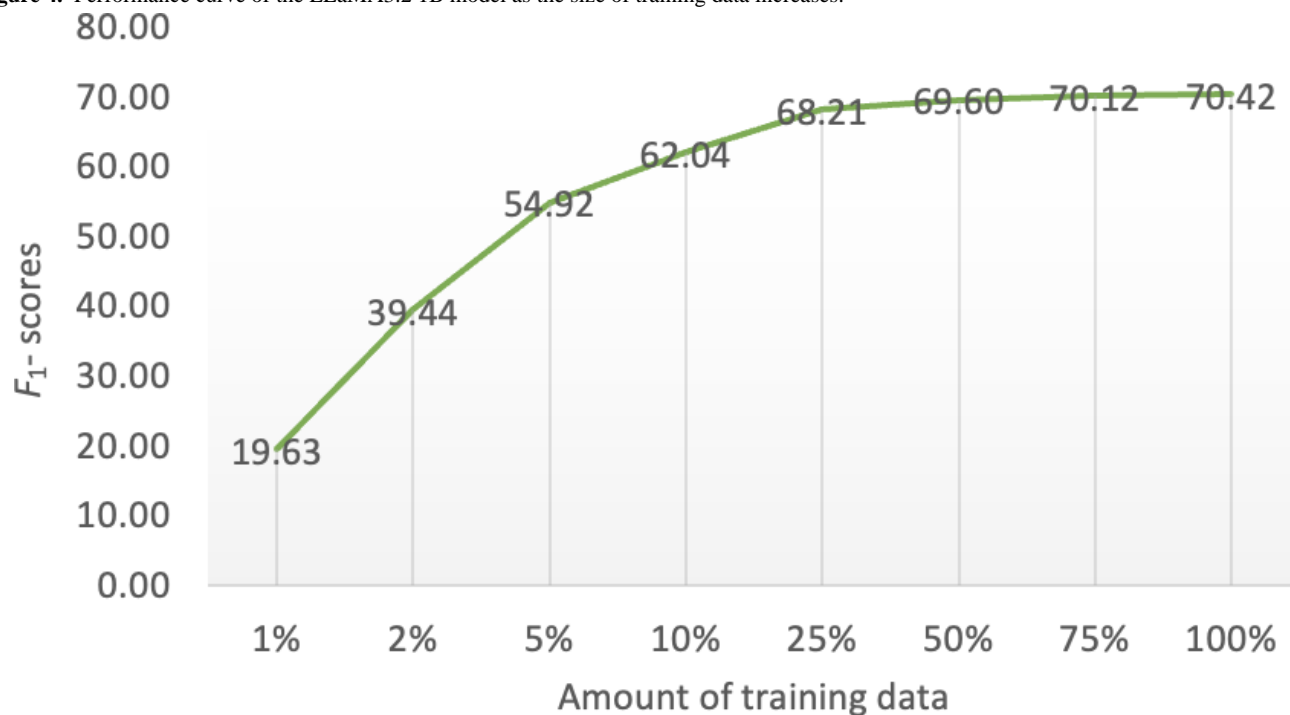
Some entities appear more frequently than other entities in our dataset. For example, diagnosis and treatment mentions are more frequent than mentions of cancer_grade. In Table 7, we present the number of instances of each entity type in our dataset and the corresponding performance of GPT4-o direct prompting. We can see that GPT4-o performs the best for the entity types that have the most instances—diagnosis, model type, and treatment entities. Of these frequent entity types, biomarker is the one with the lowest performance. Our error analysis points to several factors that could have contributed to these results, including ambiguous and inconsistent mentions and contextual dependencies. In this task, we defined a biomarker as "gene, protein or biological molecule identified in or associated with patient's/model's tumor." Thus, biomarker entities can be mentioned using their full names (eg, epidermal growth factor receptor, lnc-RP11-536 K7.3, echinoderm microtubule-associated protein-like 4), standardized gene or protein symbols (*NPM1*, KRAS, PTEN) or abbreviations of metabolites (NADPH, D2HG). Moreover, a biomarker entity (eg, "MEK") often overlaps with a treatment entity (eg, "MEK inhibitor"). The ambiguity in biomarker entity mentions might interfere with the model's ability to recognize them consistently. In addition, biomarker entities are often mentioned as lists (see Example 2) resulting in a different frequency within and across the abstracts and patterns of entity mentions, in comparison with other entities. Overall, ambiguity emerges as the primary source of error. More precise definitions, accompanied by examples illustrating the distinct meanings, might present a solution. Table S2 in Multimedia Appendix 1 provides the breakdown of errors per entity type along with examples.

Table . Evaluation results of GPT4-o with direct prompts for each entity type as precision or positive predictive value, recall or sensitivity, and F_1-score (harmonic mean of precision and recall). Results are overlapping match setting on the test set.

Entity type	Training instances, n	Development instances, n	Test instances, n	Precision	Recall	F_1-score	IAA[a]
diagnosis	362	122	114	92.47	75.44	83.09[b]	61.67
age_category	19	0	0	0.0	0.0	0.0	60
genetic_effect	69	20	33	45.71	47.06	46.38	57.67
model_type	326	114	110	88.07	84.21	86.10[b]	53.33
molecular_char	128	37	46	65.22	63.83	64.52[b]	54.33
biomarker	503	118	163	85.05	55.49	67.16[b]	61.33
treatment	426	77	130	81.74	70.15	75.50[b]	55.67
response_to_treatment	99	21	28	38.64	60.71	47.22	55
sample_type	22	8	7	45.45	71.43	55.56[b]	49
tumor_type	61	19	28	66.67	57.14	61.54[b]	49.67
cancer_grade	6	1	1	50.0	100	66.67[b]	42
cancer_stage	7	1	4	33.33	25.0	28.57	59.33
clinical_trial	35	2	4	80.0	100	88.89[b]	60.67
host_strain	9	0	7	100	28.57	44.44	61.67
model_id	17	2	7	66.67	28.57	40.0	100

[a]IAA: interannotator agreement.

[b]F_1-scores exceeding interannotator agreement.

Example 2:

Genomic alterations involved RB1 (55%), TP53 (46%), PTEN (29%), BRCA2 (29%), and AR (27%), and there was a range of androgen receptor signaling and NEPC marker expression.

The moderate performance of entity types such as genetic_effect, molecular_char, and response_to_treatment, and tumour_type is due to the number of training instances ranging from 61 to 128 as well as the IAA ranging from 49.67 to 57.67. The moderate IAA scores of those entity types underscore the need for refined annotation protocols and modeling strategies that better capture domain-specific knowledge. Furthermore, the lower performance observed for entity types with smaller sample sizes (eg, model_id) highlights the need for enhancing model performance on low-frequency labels. Future research could explore strategies such as data augmentation to improve the model's generalizability for underrepresented entities.

The extraction of PDCM-relevant knowledge is not an easy task for the domain experts as indicated by the IAA (F_1-score below 65 for all entity types except for model_id). In 9 out of 15 entity types, the system performance in an overlapping match setting exceeds the IAA (last two columns of Table 7). This is the case for categories with plentiful training instances (eg, diagnosis, model_type) as well as for categories with fewer training instances (eg, sample_type, cancer_grade). For the exact match setting, in 6 out of 15 entity types, the system performance exceeds the IAA (last two columns in Table 8). Therefore, the LLM could be a viable assistant, with its outputs reviewed by a domain expert to ensure the accuracy of the finalextraction. We believe such human-in-the-loop approaches present a promising direction for future research and application.

Table . Evaluation results of GPT4-o with direct prompts for each entity type as precision or positive predictive value, recall or sensitivity, and F_1-score (harmonic mean of precision and recall). Results are exact match setting on the test set.

Entity type	Training instances, n	Development instances, n	Test instances, n	Precision	Recall	F_1-score	IAA[a]
diagnosis	362	122	114	77.17	62.28	68.93[b]	61.67
age_category	19	0	0	0.0	0.0	0.0	60.0
genetic_effect	69	20	33	25.71	27.27	26.47	57.67
model_type	326	114	110	56.88	56.36	56.62[b]	53.33
molecular_char	128	37	46	54.35	54.35	54.35[b]	54.33
biomarker	503	118	163	46.74	26.38	33.73	61.33
treatment	426	77	130	72.34	52.31	60.71[b]	55.67
response_to _treatment	99	21	28	27.91	42.86	33.80	55
sample_type	22	8	7	45.45	71.43	55.56[b]	49
tumor_type	61	19	28	50.0	39.29	44.0	49.67
cancer_grade	6	1	1	50.0	100	66.67[b]	42
cancer_stage	7	1	4	33.33	25.0	28.57	59.33
clinical_trial	35	2	4	40.0	50.0	44.44	60.67
host_strain	9	0	7	100	14.29	25.0	61.67
model_id	17	2	7	66.67	28.27	40.0	100

[a]IAA: interannotator agreement.

[b]F_1-scores exceeding the interannotator agreement.

We would like to note that the work presented in the paper was done in a computational environment representative of the vast majority of academic medical centers and nonindustry research labs. Although we have access to SOTA Graphics Processing Units, we still found ourselves constrained as to the extent to which we could use very large language models. The larger community needs to address the growing gap in computational resources between big tech and the rest of the research community.

Limitations

As this is a feasibility study, we limited ourselves to the extraction of entity mentions of 15 entity types chosen from attributes in the descriptive standards for PDCMs. While these are recognized by the PDCM and oncology community, they do not cover all knowledge in the PDCM-relevant texts. Some refinement of the entity types will be beneficial to improve prompting results.

We limited our corpus to 100 abstracts from papers associated with PDCMs deposited in CancerModels.Org. We did not assess the abstracts for the presence and equal distribution of all the entities. Thus, there were very few mentions of some entities in the corpus (eg, cancer_stage), negatively affecting our overall F_1-score. We decided not to exclude those entities as these results could guide refinements of future studies. The computational methods discussed here are applicable to other studies requiring the extraction of textual information from scientific papers. Future work could involve extending this method to extract knowledge from the main body of the papers.

Conclusions

This study investigates the potential of LLMs as powerful tools for extracting PDCM-relevant knowledge from scientific literature—an essential task for advancing cancer research and precision medicine. By comparing direct and soft prompting across both proprietary and open LLMs, we provide valuable insights into the most effective strategies for PDCM-relevant knowledge extraction. Our findings indicate that GPT-4o, when used with direct prompting, maintains competitive performance, while soft prompting significantly enhances the effectiveness of smaller LLMs. In conclusion, our results demonstrate that training soft prompts on smaller open models can achieve performance levels comparable to those of proprietary LLMs.

To our knowledge, this is the first study to implement SOTA LLMs prompting for knowledge extraction in the PDCM domain and the first to explore the emerging topic of soft prompting in this context. Our findings demonstrate that LLMs can effectively streamline the extraction of complex cancer model metadata, potentially reducing the burden of manual curation and accelerating the integration of PDCMs into research and clinical workflows. Additionally, this study lays the foundation for future research aimed at optimizing LLMs for large-scale knowledge extraction tasks. Efficiently extracting and harmonizing PDCM-relevant knowledge will ultimately drive progress in cancer research and precision oncology, equipping researchers and clinicians with better tools to improve patient

outcomes. More broadly, our study contributes to the ongoing discourse on the applicability of LLMs, acknowledging that while they offer transformative potential, they are not a universal solution for all tasks.

Acknowledgments

Data Availability

The data and code will be available upon publication in the CancerModels.Org Github repository [37].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Prompts used in direct prompting experiments and detailed error analysis.
[DOCX File, 20 KB - bioinform_v6i1e70706_app1.docx]

References

1. RePORTER. National Institutes of Health. URL: https://reporter.nih.gov/ [accessed 2024-12-16]
2. Perova Z, Martinez M, Mandloi T, et al. PDCM Finder: an open global research platform for patient-derived cancer models. Nucleic Acids Res 2023 Jan 6;51(D1):D1360-D1366. [doi: 10.1093/nar/gkac1021] [Medline: 36399494]
3. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. arXiv. Preprint posted online on Jun 12, 2017. [doi: 10.48550/arXiv.1706.03762]
4. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. arXiv. Preprint posted online on May 28, 2020. [doi: 10.48550/arXiv.2005.14165]
5. OpenAI, Achiam J, Adler S, et al. GPT-4 technical report. arXiv. Preprint posted online on Mar 15, 2023. [doi: 10.48550/arXiv.2303.08774]
6. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023 Mar 30;388(13):1233-1239. [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
7. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large language models in medicine: the potentials and pitfalls: a narrative review. Ann Intern Med 2024 Feb;177(2):210-220. [doi: 10.7326/M23-2772] [Medline: 38285984]
8. Saab K, Tu T, Weng WH, et al. Capabilities of Gemini models in medicine. arXiv. Preprint posted online on Apr 29, 2024. [doi: 10.48550/arXiv.2404.18416]
9. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA 2023 Jul 3;330(1):78-80. [doi: 10.1001/jama.2023.8288] [Medline: 37318797]
10. Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. NPJ Digit Med 2024 Jan 24;7(1):20. [doi: 10.1038/s41746-024-01010-1] [Medline: 38267608]
11. Williams CYK, Miao BY, Kornblith AE, Butte AJ. Evaluating the use of large language models to provide clinical recommendations in the emergency department. Nat Commun 2024 Oct 8;15(1):8236. [doi: 10.1038/s41467-024-52415-1] [Medline: 39379357]
12. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 1;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
13. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ 2024 Nov;58(11):1276-1285. [doi: 10.1111/medu.15402] [Medline: 38639098]
14. Kung TH, Cheatham M, Medenilla A, 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]
15. Perot V, Kang K, Luisier F, et al. LMDX: language model-based document information extraction and localization. In: Ku LW, Martins A, Srikumar V, editors. Findings of the Association for Computational Linguistics ACL 2024: Association for Computational Linguistics; 2024:15140-15168. [doi: 10.18653/v1/2024.findings-acl.899]
16. Arsenyan V, Bughdaryan S, Shaya F, Small KW, Shahnazaryan D. Large language models for biomedical knowledge graph construction: information extraction from EMR notes. In: Demner-Fushman D, Ananiadou S, Miwa M, Roberts K, Tsujii J, editors. Proceedings of the 23rd Workshop on Biomedical Natural Language Processing: Association for Computational Linguistics; 2024:295-317. [doi: 10.18653/v1/2024.bionlp-1.23]

17.   Munnangi M, Feldman S, Wallace B, Amir S, Hope T, Naik A. On-the-fly definition augmentation of LLMs for biomedical NER. In: Duh K, Gomez H, Bethard S, editors. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Association for Computational Linguistics; 2024:3833-3854. [doi: 10.18653/v1/2024.naacl-long.212]

18.   Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning. In: Moens MF, Huang X, Specia L, Yih SW, editors. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Association for Computational Linguistics; 2021:3045-3059. [doi: 10.18653/v1/2021.emnlp-main.243]

19.   Meehan TF, Conte N, Goldstein T, et al. PDX-MI: minimal information for patient-derived tumor xenograft models. Cancer Res 2017 Nov 1;77(21):e62-e66. [doi: 10.1158/0008-5472.CAN-17-0582] [Medline: 29092942]

20.   PDCMFinder/MI-standard-in-vitro-models. GitHub. URL: https://github.com/PDCMFinder/MI-Standard-In-vitro-models [accessed 2024-12-23]

21.   Hripcsak G, Rothschild AS. Agreement, the F-measure, and reliability in information retrieval. J Am Med Inform Assoc 2005;12(3):296-298. [doi: 10.1197/jamia.M1733] [Medline: 15684123]

22.   Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. arXiv. Preprint posted online on Jan 28, 2022. [doi: 10.48550/arXiv.2201.11903]

23.   Wang X, Wei J, Schuurmans D, et al. Self-consistency improves chain of thought reasoning in language models. arXiv. Preprint posted online on Mar 21, 2022. [doi: 10.48550/arXiv.2203.11171]

24.   Liu X, Zheng Y, Du Z, et al. GPT understands, too. arXiv. Preprint posted online on Mar 18, 2021. [doi: 10.48550/arXiv.2103.10385]

25.   Schulhoff S, Ilie M, Balepur N, et al. The prompt report: a systematic survey of prompting techniques. arXiv. Preprint posted online on Jun 6, 2024. [doi: 10.48550/arXiv.2406.06608]

26.   Li XL, Liang P. Prefix-tuning: optimizing continuous prompts for generation. arXiv. Preprint posted online on Jan 1, 2021. [doi: 10.48550/arXiv.2101.00190]

27.   Zhou Y, Muresanu A I, Han Z, Paster K, Pitis S, Chan H. Large language models are human-level prompt engineers. arXiv. Preprint posted online on Nov 3, 2022. [doi: 10.48550/arXiv.2211.01910]

28.   Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. arXiv. Preprint posted online on Oct 16, 2013. [doi: 10.48550/arXiv.1310.4546]

29.   Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev 1958 Nov;65(6):386-408. [doi: 10.1037/h0042519] [Medline: 13602029]

30.   Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov 15;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]

31.   Tjong Kim Sang EF, De Meulder F. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. Presented at: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL; May 31, 2003; Edmonton, Canada. [doi: 10.3115/1119176.1119195]

32.   Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507-513. [doi: 10.1136/jamia.2009.001560] [Medline: 20819853]

33.   von Oswald J, Niklasson E, Randazzo E, et al. Transformers learn in-context by gradient descent. arXiv. Preprint posted online on Dec 15, 2022. [doi: 10.48550/arXiv.2212.07677]

34.   Hendel R, Geva M, Globerson A. In-context learning creates task vectors. In: Bouamor H, Pino J, Bali K, editors. Findings of the Association for Computational Linguistics: EMNLP 2023: Association for Computational Linguistics; 2023:9318-9333. [doi: 10.18653/v1/2023.findings-emnlp.624]

35.   OpenAI, Hurst A, Lerer A, et al. GPT-4o system card. arXiv. Preprint posted online on Oct 25, 2024. [doi: 10.48550/arXiv.2410.21276]

36.   Grattafiori A, Dubey A, Jauhri A, et al. The Llama 3 herd of models. arXiv. Preprint posted online on Jul 31, 2024. [doi: 10.48550/arXiv.2407.21783]

37.   PDCMFinder/prompt-llm. GitHub. URL: https://github.com/PDCMFinder/prompt-llm [accessed 2024-12-23]

Abbreviations

IAA: interannotator agreement
LLM: large language model
PDCM: patient-derived cancer model
PDX: patient-derived xenografts
SOTA: state-of-the-art

XSL•FO
RenderX

<u>Original Paper</u>

A Hybrid Deep Learning–Based Feature Selection Approach for Supporting Early Detection of Long-Term Behavioral Outcomes in Survivors of Cancer: Cross-Sectional Study

Tracy Huang[1*], BA; Chun-Kit Ngan[2*], BA, PhD; Yin Ting Cheung[3], PhD; Madelyn Marcotte[2], BSc; Benjamin Cabrera[4], BSc

[1]Emory University, Atlanta, GA, United States

[2]Worcester Polytechnic Institute, Worcester, MA, United States

[3]Chinese University of Hong Kong, Hong Kong, China (Hong Kong)

[4]Arizona State University, Tempe, AZ, United States

[*]these authors contributed equally

Corresponding Author:
Chun-Kit Ngan, BA, PhD
Worcester Polytechnic Institute
100 Institute Rd
Worcester, MA, 01609
United States
Phone: 1 (508) 831 5000
Email: cngan@wpi.edu

Abstract

Background: The number of survivors of cancer is growing, and they often experience negative long-term behavioral outcomes due to cancer treatments. There is a need for better computational methods to handle and predict these outcomes so that physicians and health care providers can implement preventive treatments.

Objective: This study aimed to create a new feature selection algorithm to improve the performance of machine learning classifiers to predict negative long-term behavioral outcomes in survivors of cancer.

Methods: We devised a hybrid deep learning–based feature selection approach to support early detection of negative long-term behavioral outcomes in survivors of cancer. Within a data-driven, clinical domain–guided framework to select the best set of features among cancer treatments, chronic health conditions, and socioenvironmental factors, we developed a 2-stage feature selection algorithm, that is, a multimetric, majority-voting filter and a deep dropout neural network, to dynamically and automatically select the best set of features for each behavioral outcome. We also conducted an experimental case study on existing study data with 102 survivors of acute lymphoblastic leukemia (aged 15-39 years at evaluation and >5 years postcancer diagnosis) who were treated in a public hospital in Hong Kong. Finally, we designed and implemented radial charts to illustrate the significance of the selected features on each behavioral outcome to support clinical professionals' future treatment and diagnoses.

Results: In this pilot study, we demonstrated that our approach outperforms the traditional statistical and computation methods, including linear and nonlinear feature selectors, for the addressed top-priority behavioral outcomes. Our approach holistically has higher F_1, precision, and recall scores compared to existing feature selection methods. The models in this study select several significant clinical and socioenvironmental variables as risk factors associated with the development of behavioral problems in young survivors of acute lymphoblastic leukemia.

Conclusions: Our novel feature selection algorithm has the potential to improve machine learning classifiers' capability to predict adverse long-term behavioral outcomes in survivors of cancer.

(JMIR Bioinform Biotech 2025;6:e65001) doi:10.2196/65001

Introduction

Background

The number of survivors of cancer is increasing globally. The American Cancer Society recently reported that in 2023, a total of 1,958,310 new cancer cases were projected to occur in the United States [1]. Treatment advances have resulted in a dramatic improvement in the survival rates of most cancers, especially in resource-limited countries and regions. However, this growing population of survivors of cancer may develop a myriad of treatment-related adverse effects that lead to a compromised health status. Studies have also shown that survivors of cancer are more likely than the general population to experience negative long-term behavioral outcomes, such as anxiety, depression, attention problems, and sluggish cognitive tempo, after cancer treatments [2]. Contemporary treatment strategies have led to improved life expectancy after treatment for pediatric cancer, especially in survivors of acute lymphocytic leukemia (ALL) [3]. Given that studies have shown that the promotion of a healthy lifestyle and interventions that reduce physical and mental health burdens can lead to reduction in all-cause and cause-specific mortality, addressing the risk factors of adverse functional outcomes early on is critical [4-6]. Thus, developing an effective approach to identify crucial factors and then detect these negative outcomes in advance is needed so that medical therapists can intervene early and take the appropriate actions and treatments promptly to mitigate adverse effects in survivors of cancer.

Current Approaches for Detecting Adverse Behavioral Outcomes in Survivors of Cancer

Currently, to support the identification of relevant factors and the early detection of adverse behavioral outcomes for survivors of cancer, clinical scientists use various statistical analyses to understand the relationship among those behavioral outcomes, cancer treatments, chronic health conditions, and socioenvironmental factors [7-9]. Specifically, traditional statistical methods (ie, linear regression analysis) are used to extract predictor variables and then model the relationship between the extracted predictor variables and the behavioral outcomes. This analysis assumes that the behavioral outcomes are, for the most part, linearly correlated with those predictor variables. However, this assumption may not always hold in this complex and dynamic problem. Furthermore, the predictors for those behavioral outcomes extracted by statistical methods may have weak prediction accuracy, as modeling human behavioral outcomes is challenging due to its multifactorial nature (ie, many predictors as well as interactions among the predictors affecting the outcome), heterogeneity (ie, differences across individuals), nonlinearity of data, multicollinearity (ie, highly correlated variables), class imbalance (ie, few observations of the outcome of interest), and missing data [10,11]. As a result, this class of linear regressors can only account for a small proportion of variance, with limited usability in a clinical setting. Thus, developing an effective computational methodology that can maximize the use of those data for prognostic and predictive behavioral outcomes is highly desirable.

To address the abovementioned problems, feature selection techniques in machine learning (ML) play an important role. Feature selection techniques can be broadly divided into 4 categories: filter, wrapper, embedded, and hybrid. Filter methods select features based on their statistical significance to the outcome of interest. Unlike other feature selection methods, such as wrapper and embedded methods, filter methods function independent of any ML classifiers. However, filter methods are less accurate than other methods of feature selection, such as wrapper methods. In addition, there is a risk of selecting redundant features when using filter methods that do not consider the correlation between features. Wrapper methods use a greedy search algorithm (ie, an iterative algorithm that makes the locally optimal choice at each step) with a classifier to sequentially add and remove features from the classifier to maximize the specified scoring metrics, that is, precision, recall, and F_1-score. The output is the best subset of features that the algorithm found. While wrapper methods are proficient in achieving high classification accuracy, they are not efficient in computation time or complexity. In addition, there is also a risk of overfitting with wrapper methods, where the classifier is highly trained to generate accurate predictions for the training data only and cannot correctly create generalized predictions for testing data or any novel datasets. Embedded methods use qualities from both filter and wrapper methods to perform feature selection during the construction of the ML classifiers. The baseline embedded methods that are commonly used are least absolute shrinkage and selection operator (Lasso), Ridge, and ElasticNet. However, to effectively use embedded methods, prior knowledge of the feature sets is required. In addition, embedded methods could pose problems when identifying small feature sets. Hybrid methods combine filter and wrapper methods to take advantage of the benefits each method provides, while minimizing their limitations [12]. A filter method first selects a subset of features, which are then input into a wrapper method to further select the best subset of features. As hybrid methods are a combination of filter and wrapper methods, they inherit problems from both—filter methods may exclude important features and wrapper methods are inefficient in computation time.

Goal of This Study

To bridge the abovementioned gaps, we propose a hybrid deep learning–based feature selection approach to support early detection of long-term adverse behavioral outcomes in survivors of cancer. Specifically, our goals are four-fold: (1) devise a data-driven, clinical domain–guided framework to select the best set of features among cancer treatments, chronic health conditions, socioenvironmental factors, and others; (2) develop a 2-stage feature selection algorithm, that is, a multimetric, majority-voting filter and a deep dropout neural network (DDN), to dynamically and automatically select the best set of features for each behavioral outcome; (3) conduct an experimental case study on our existing study data with 102 survivors of ALL (aged 15-39 years at evaluation and >5 years postcancer diagnosis) who were treated in a public hospital in Hong Kong; and (4) design and implement radial charts to illustrate the significance of the selected features on each behavioral outcome to support clinical professionals' future treatment and diagnoses.

In this pilot study, we demonstrate that our approach outperforms the traditional statistical and computation methods, including linear and nonlinear feature selectors, for the addressed top-priority behavioral outcomes.

Methods

Review of Baseline Feature Selection Methods

Overview

Four baseline feature selection methods were used in the experimental studies as a comparison for our novel feature selection algorithm (Textbox 1).

Textbox 1. Summary of the baseline feature selection methods.

Filter

- Correlation-based feature selection (CFS)
- Information gain (IG)
- Maximum relevance minimum redundancy (MRMR)

Wrapper

- Sequential forward selection (SFS)
- Sequential backwards selection (SBS)
- Stepwise selection (SS)

Embedded

- Least absolute shrinkage and selection operator (Lasso)
- Ridge
- ElasticNet

Hybrid

- CFS→SFS
- IG→SFS
- MRMR→SFS
- CFS→SFS
- IG→SFS
- MRMR→SFS
- CFS→SBS
- IG→SBS
- MRMR→SBS
- CFS→SS
- IG→SS
- MRMR→SS

Filter Methods

Filter methods select features based on their statistical significance to the outcome of interest, independent of any ML classifiers. To evaluate the performance of existing filter methods, we use information gain (IG), maximum relevance minimum redundancy (MRMR), and correlation-based feature selection (CFS) [13]. IG is calculated by comparing the entropy of the dataset before and after a transformation. When IG is used for feature selection, it is called mutual information and works by evaluating the IG of each variable in the context of the target. The MRMR algorithm selects the best K features at each iteration that have maximum relevance with respect to the target variable and minimum redundancy with respect to the other features. The CFS algorithm involves splitting the features into subsets based on whether their values are continuous or discrete and can be used to measure the correlation between features and the target outcomes. For continuous data, Pearson correlation can be used, and for discrete data, symmetrical uncertainty can be used. Symmetrical uncertainty is a measure of relevance between features and targets that uses mutual information [14]. When evaluating the performance of the existing filter methods, we selected the top 15 features that had the highest scores for each of the 3 approaches.

Wrapper Methods

For binary classification, wrapper methods use a greedy search algorithm with a classifier to sequentially add and remove features from the classifier to maximize the specified scoring metric, that is, precision, recall, and F_1-score. The output is the best subset of features that the algorithm found. To evaluate existing wrapper methods' performances, we selected 3 commonly implemented wrapper methods: sequential forward selection (SFS), sequential backward selection (SBS), and stepwise selection (SS). SFS starts with an empty subset of features and iteratively adds features if adding them improves the specified score, according to the ML classifier. The selection terminates when a feature subset of the desired size k, where k refers to the number of features expected by the domain experts, is reached. In contrast, SBS starts with a full subset of all the features and iteratively removes features if removing them increases the specified score, according to the classifier. The selection also terminates when a feature subset of the desired size k is reached. SS, also known as bidirectional selection, alternates between forward and backward selection to select the best subset of features. To implement the wrapper selection approaches, we used the support vector machine classifier and used accuracy as the default scoring metric [15]. We also specified that the selection process should terminate when a feature subset of size 15 is reached. For the purpose of the study, we decided a priori that the feature subset should be limited to 15 because if there are too many exploratory factors in the model, the contribution of each factor to the variance may be too small and its clinical significance may be questionable.

Embedded Methods

Embedded methods use qualities from both filter and wrapper methods to perform feature selection during the construction of the ML classifier. The embedded classifiers we used were Lasso, Ridge, and ElasticNet. Lasso regression is a form of linear regression that imposes an L1 regularization penalty to identify the features that minimize the prediction error [16]. Similar to Lasso, Ridge regression is another form of linear regression that uses an L2 penalty instead [17]. ElasticNet regression merges Lasso and Ridge regression using the L1 and L2 regularization penalties [18]. ElasticNet regression can shrink some features to zero, similar to Lasso, while reducing the magnitude of other features, like Ridge. For each evaluated embedded method, we selected the top 15 most relevant features for each behavioral outcome.

Hybrid Methods

Hybrid methods combine filter and wrapper methods to take advantage of the benefits each method provides, while minimizing their limitations [12]. We implemented 9 different hybrid methods using the top 30 features selected from the 3 filter methods (ie, CFS, IG, and MRMR) and inputting them each into the 3 wrapper methods, including SFS, SBS, and SS, to subsequently select the top 15 features.

Data-Driven, Clinical Domain–Guided Framework

In this section, we describe and explain our framework that consisted of 6 main modules (Figure 1). The cancer survivor medical records, including the features, such as biomarkers, chronic health conditions, and socioeconomic factors, were first passed into the data cleaner that "sanitizes" the records with the clinical domain knowledge from our investigators. Note that throughout the framework, our clinical domain experts assisted us with certain processes. In this case study, for example, it consisted of replacing missing values in a patient's record by averaging the existing values of the corresponding feature among all the other patients' records grouped by a specific cancer type, age range, and biological sex. Clinical domain experts also helped us interpret and explain what different variable values mean for us to properly transform them into the correct variables.

Figure 1. Data-driven, clinical domain–guided framework.
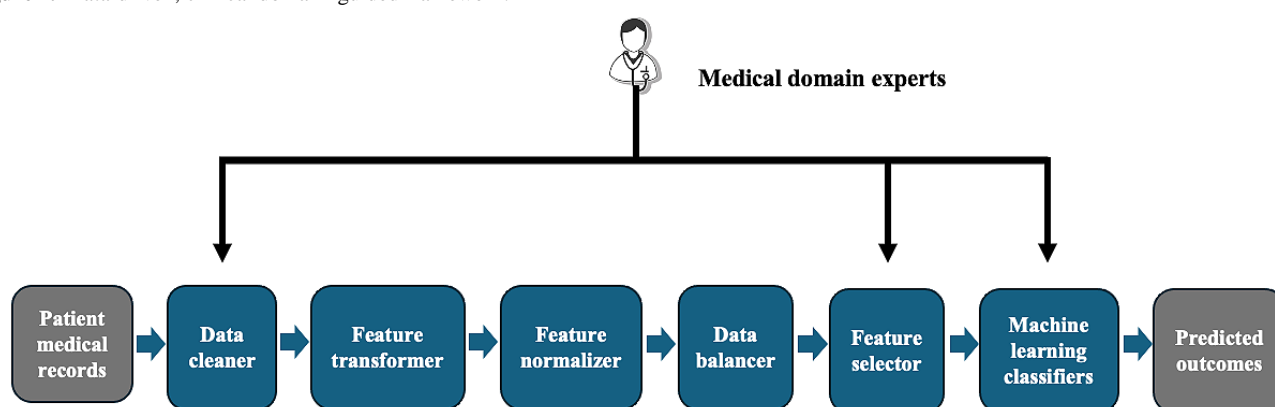
Afterward, the records were passed into the feature transformer, where the one-hot encoding technique was used to transform categorical variables into binary ones [19]. For instance, we transformed the "gender" variable from categorical to binary by replacing "M" and "F" with 1 and 0.

Following feature transformation, the records were normalized by the feature normalizer. The Shapiro-Wilk test, the Kolmogorov-Smirnov test, and the D'Agostino-Pearson test were used to check whether features follow a normal distribution. If 2 out of the 3 tests conclude that a feature follows a normal distribution, it is standardized by removing the mean and scaling to unit variance [20-22]. Otherwise, features are normalized using the minimum-maximum normalization technique so that all features have values between "0" and "1." This eliminates any feature bias, where features with high values are given more importance than features with low values [23].

Once the records are cleaned, transformed, and normalized, they are then passed into the data balancer. At this point, the results differ depending on the behavioral outcome being predicted. The synthetic minority oversampling technique for nominal and continuous (SMOTE-NC) is used to artificially balance the instances where the number of patients having a behavioral outcome of "1" is the minority, which is most often the case as cancer survivor datasets are often imbalanced. The SMOTE-NC technique oversamples the minority class in unbalanced datasets by creating synthetic examples instead of oversampling using replacement. The algorithm involves computing the median of the SD of continuous variables for the minority class and using the median to penalize nominal features that differ between the considered feature vector and its potential nearest neighbors, conducting nearest neighbors computation, and populating the synthetic class [24]. The SMOTE-NC technique is also used to artificially oversample the minority gender so the final datasets can have equal instances of "0" and "1" for the behavioral outcome. We specifically chose the SMOTE-NC technique over the regular synthetic minority oversampling technique because our dataset had a mixture of nominal and continuous features. synthetic minority oversampling technique can only handle datasets with continuous features. The data were then split into 69.6% (71/102) training and 30.4% (31/102) testing data.

Once the survivors of cancer' clinical records passed through all the steps of data preprocessing, they were passed into our hybrid deep learning–based feature selection that was a 2-stage feature selection algorithm, that is, a multimetric, majority-voting filter and a DDN, to dynamically and automatically select the best set of features for each behavioral outcome. Specifically, the first stage was a novel filter method that uses 4 metrics to select the most relevant features for a behavioral outcome and removes any redundant features. The second stage was a DDN that replaces a wrapper method, where it further selects features from the ones selected by the multimetric, majority-voting filter to maximize prediction performance in ML classifiers. Note that our clinical domain experts used their clinical expertise to recommend certain features that should be kept in all the final feature lists due to their clinical importance (ie, gender, current age, and age at diagnosis in our case), if they were not already selected to be in the final feature list by our feature selection approach. Finally, the training data with the final feature list selected from the feature selector with the clinical domain expertise were passed into 3 ML classifiers, including logistic regression, naive Bayes, and k-nearest neighbors, to calculate the precision, recall, and F_1-score for the performance evaluation on the testing data.

2-Stage Feature Selection Algorithm

Our proposed 2-stage feature selection algorithm consisted of 2 sequential stages, including a multimetric, majority-voting filter, and a DDN (Figure 2).

Figure 2. Two-stage feature selection algorithm.
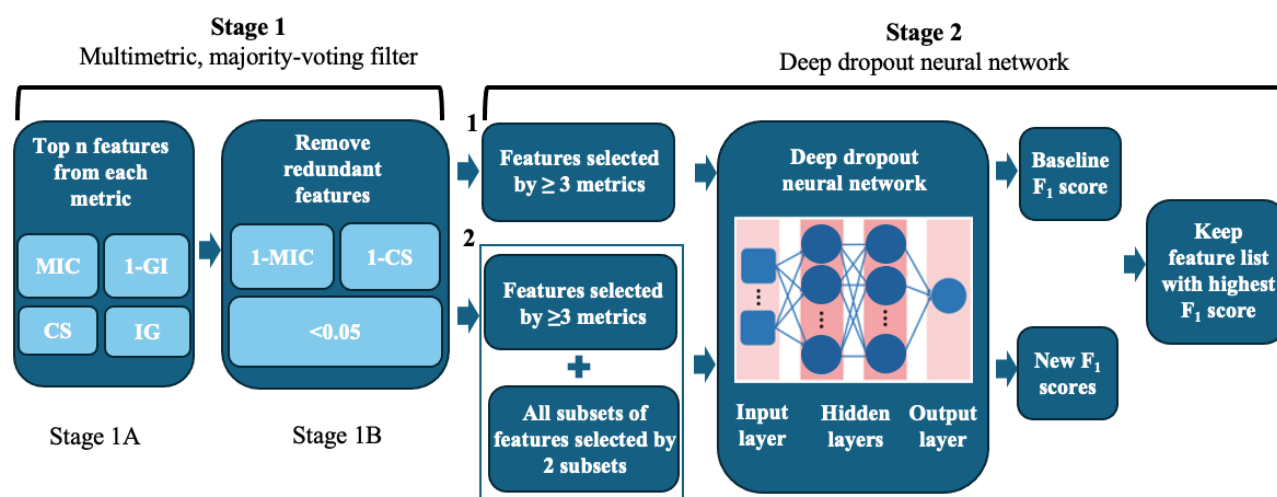
Stage 1: A Multimetric, Majority-Voting Filter

Overview

Our hybrid deep learning–based feature selection methodology specifically addressed the limitations of existing feature selection methods. In the first stage, it removed redundant features, which some existing filter methods do not consider. Specifically, our 3 majority-voting (MV) filter had 2 processing steps in stage 1.

In stage 1A, we used 4 different metrics to select the features that are the most relevant to predict a behavioral outcome. Those metrics include maximal information coefficient (MIC), Gini index (GI), IG, and correlation score (CS) that we calculated between each candidate feature in our preprocessed dataset and the corresponding behavioral outcome of interest. The MIC is a measure of the strength of the linear or nonlinear association between 2 variables X and Y, where $X \in R$ is the input feature and $Y \in R$ is the corresponding behavioral outcome.

The GI represents the amount of probability of a specific feature that is classified incorrectly when selected randomly. Unlike the other 3 metrics, a higher GI score represents lower associations with the behavioral outcome of interest. To make the scale of the correlation strength between X and Y consistent among all the metrics, the metric that we used was 1–GI instead.

That is, for all the 4 metrics, a higher value indicated a higher association with the behavioral outcome of interest.

The IG is a measure of the expected reduction in entropy caused by partitioning the samples according to a specific attribute X.

The CS between X and Y is calculated using the Pearson correlation coefficient, point-biserial correlation, and the ϕ coefficient, based upon the data type of X and Y [25]. When both X and Y are the continuous variables, the Pearson correlation coefficient should be used. When comparing 1 continuous and 1 binary variable, the point-biserial correlation is used [26]. Finally, when comparing 2 binary variables, the ϕ is used. All these measures are values between −1 and 1, with −1 being a perfect negative correlation and 1 being a perfect positive correlation, while 0 represents no correlation. We take the absolute value of each measure so that the CS is always between 0 and 1.

After we calculated the values of all 4 abovementioned metrics between each candidate feature and the behavioral outcome of interest, we ranked the top N features (ie, the number of features expected by the domain experts) for each of the metrics in descending order and stored them in a master list, without repetition. From this master list, we constructed 3 feature lists. The first list contained the features selected by at least 3 metrics, as they are highly likely relevant to predict the behavioral outcome and are then included in the final feature list. The second one contained the features selected by exactly 2 metrics, as they might have been relevant to predict the behavioral outcome and were then needed for further analysis in stage 1B. The third one combined all the features from the previous 2 lists so that we could evaluate the redundancy between any 2 features from this list.

In stage 1B, we removed any redundant features from the third combined list generated from stage 1A. We used the *MIC* and the *CS* and then calculated these 2 values for all the feature-to-feature combinations in the combined feature list output from stage 1A. We subtracted the *MIC* and the *CS* values from 1 and then used the *1–MIC* and *1–CS* values to determine if any feature was redundant by other features. The threshold

we set was 0.05, based upon our preliminary experimental analysis, so that any combination of 2 features that resulted in both scores being <0.05 was determined to be redundant. Once it was determined that 2 features were redundant, we looked at the number of metrics that selected the features. If one of the features was selected by fewer metrics, that feature was removed from the third combined list. If both features were selected by the same number of metrics and they were redundant, we then looked at the average rank of each feature across the 4 ranked lists by MIC, GI, IG, and CS. The feature with the lower rank was removed from the third combined list. The pseudocode algorithm is detailed for the multimetric, majority-voting filter in Multimedia Appendix 1.

For illustration, we used our dataset as an example to explain our multimetric, majority-voting filter.

Stage 1A: Select the Top N Features Per Metric

Overview

Suppose we want to select the best features for predicting the behavioral outcome, thought problems. This is our *B_Outcome*. F is the set of all input candidate features F_i in the preprocessed clinical records. We then calculate the MIC, 1–GI, IG, and CS scores for all the candidate features in the preprocessed clinical records and our *B_Outcome*, thought problems. We store these results in 4 sets, *MIC, 1–GI, IG,* and *CS*. In this example, our domain experts expected 15 nonredundant input candidate features to be selected; thus, N was set to 15.

Step 1

We first sorted the input features (ie, F_is) according to their MIC, 1–GI, CS, and IG scores. Since N was 15, we then took the top 15 features with the highest values from the MIC set and placed them into a separate set, that is, F_{MIC}. We repeated this with (1–GI), IG, and CS scores and placed the top 15 features into the corresponding sets, that is, $F_{1\text{-}GI}$, F_{CS}, and F_{IG}. At this point, we had the following features in these sets: $F_{MIC}, F_{1\text{-}GI}$, F_{CS}, and F_{IG}. As there were 15 features in each set, we had 60 features across all the 4 sets (Textbox 2).

Textbox 2. Total input features sorted by maximal information coefficient (MIC), 1- Gini index (GI), correlation score (CS), and information gain (IG) scores in descending order.

F_{MIC}

Physical fatigue>overall fatigue>cognitive fatigue>family communication>family concern>IV high-dose methotrexate (MTX)>sleep fatigue>physical activity>family conflict>parental control>family mutuality>age at cancer diagnosis>intrathecal MTX dose>noncranial radiation>cranial radiation therapy

$F_{1–GI}$

Years of education>intrathecal chemotherapy>leukemia risk group>intrathecal MTX dose>living space>physical activity>cognitive fatigue>family communication>physical fatigue>family mutuality>IV high-dose MTX>sleep fatigue>family conflict>age at cancer diagnosis>age at evaluation

F_{CS}

Physical fatigue>overall fatigue>cognitive fatigue>family communication>IV high-dose MTX>family concern>sleep fatigue>family conflict>parental control>physical activity>cranial radiation therapy>noncranial radiation>intrathecal MTX dose>years of education>family mutuality

F_{IG}

Impulsivity (on continuous performance test [CPT; Conner continuous performance test to measure a person's performance in attention, particularly in areas of inattentiveness, impulsivity, variation in response speed, sustained attention, and information processing efficiency] attention test)>inattentiveness (on CPT Attention test)>information processing efficiency (on CPT attention test)>hematopoietic stem cell transplant>response speed variability (on CPT Attention Test)>surgery>sustained attention (on CPT attention test)>physical fatigue>overall fatigue>neurological complications>leukemia risk group >living space>inattentiveness (on CPT attention test)>inflammatory interleukin-7

Step 2

We then created a new set F_{UNION}, the union of sets F_{MIC}, $F_{1–GI}$, F_{CS}, and F_{IG} in step 1, allowing duplicate values. This set F_{UNION} represents all the features that have the top 15 MIC, 1–GI, IG, and CS scores. At this point, the set F_{UNION} contained 60 total features.

Step 3

From the set F_{UNION}, we created the subset *3Metrics+* from the features that were stored in at least 3 of these 4 sets, F_{MIC}, $F_{1–GI}$, F_{CS}, and F_{IG}. These features were then selected as 1 of the top 15 by at least 3 out of the 4 metrics, so these are likely to be highly relevant to predict our *B_Outcome*, thought problems, and were included in the final feature list. By applying this concept, the subset *3Metrics+* contained 10 features.

Step 4

From the set F_{UNION}, we also created a subset *2Metrics* from features that were stored in exactly 2 out of these 4 sets, F_{MIC}, $F_{1–GI}$, F_{CS}, and F_{IG}. These features were selected as the top 15 by 2 out of the 4 metrics only. Thus, they may be relevant to predict the *B_Outcome*, thought problems, but needed to be further analyzed in stage 2 to determine if they should be kept in the final feature list. By applying this concept, the subset *2Metrics* contained 8 features only.

Step 5

We created another set *3+2Metrics*, that is, the union of the sets *3Metrics+* and *2Metrics*, without the duplicate values. At this point, the set *3+2Metrics* contained 18 features, including 10 in the *3Metrics+* set and 8 in the *2Metrics* set (Textbox 3).

Textbox 3. Features in the 3Metrics+ and 2Metrics sets.

3Metrics+

- Physical fatigue
- Overall fatigue
- Cognitive fatigue
- Family communication
- Sleep fatigue
- Family conflict
- Family mutuality
- Physical activity
- IV high-dose methotrexate (MTX)
- Intrathecal MTX dose

2Metrics

- Leukemia risk group
- Living space
- Family concern
- Cranial radiation therapy
- Years of education
- Family control
- Age at cancer diagnosis
- Noncranial radiation

Step 6

We also created a 1D matrix, *Rank*, which stored the average rank position of each feature in *3+2Metrics* from the sets F_{MIC}, $F_{1–GI}$, F_{CS}, and F_{IG}. For instance, if we consider the feature "physical fatigue," as its position was 1, 9, 1, and 9 in the sets F_{MIC}, F_{1-GI}, F_{CS}, and F_{IG}, respectively, its average position value in *Rank* was equal to 5.

Step 7

Finally, we evaluated whether there were too many or too few features at this stage. We first evaluated the number of features in *3Metrics+*. As *3Metrics+* had 10 features, which was less than N, there was no need to remove any extra features. We then evaluated the number of features in *3+2Metrics*. As there were 18 features in *3+2Metrics*, which was greater than **N**, there was no need to go back to step 1 to find at least 15 features. We now had 3 sets as the outputs: *3Metrics+* with 10 features that were selected by at least 3 metrics; *2Metrics* with 8 features that were selected by exactly 2 metrics; and *3+2Metrics*, with 18 features that included the features from both *3Metrics+* and *2Metrics*.

Stage 1B: Remove Redundant Input Features

At this step, we wanted to remove any redundant features from the features that we selected in stage 1A.

Step 1

We computed 1–MIC(f$_i$, f$_j$) values and 1–CS(f$_i$, f$_j$) values by the developed *compute_MIC* and *compute_CS* functions between any pair of 2 features f$_1$ and f$_2$ in 3+2Metrics. We stored the 1–MIC(f$_i$, f$_j$) values and 1–CS(f$_i$, f$_j$) values in the sets *MIC_Feature_Score* and *CS_Feature_Score*, respectively.

Step 2

We iterated each value in *MIC_Feature_Score* and *CS_Feature_Score* between any pair of 2 features f$_1$ and f$_2$ in *3+2Metrics* and checked if any values were <0.05. We then checked if there was any feature pair that had values <0.05 in both *MIC_Feature_Score* and *CS_Feature_Score*. Suppose we found that the values in *MIC_Feature_Score* and *CS_Feature_Score* that corresponded to the feature pair, "cranial radiation therapy" and "noncranial radiation," were indeed both <0.05, then we select those 2 features as the feature pair that we need to further analyze, as they were categorized as the redundant features at this step. Suppose that "cranial radiation therapy" and "noncranial radiation" were both in the set *2Metrics*, meaning that they were both selected by 2 metrics, then according to the algorithm, they were selected by an equal number of metrics and we must compare their rankings in *Rank* to decide which one must be removed. Suppose that "noncranial radiation" had a lower rank, or a higher score, compared to "cranial radiation therapy," then we remove "noncranial radiation" from the set *3+2Metrics*.

Step 3

After we removed the redundant features from the set *3+2Metrics*, we then split the set *3+2Metrics* into 2 new sets: F_{3M+}, such that its nonredundant features were selected by at least 3 metrics in the set F_{UNION}, and F_{2M}, such that its

nonredundant features were selected by exactly 2 metrics in the set F_{UNION}.

Step 4

We now had 2 sets: F_{3M+} and F_{2M}. The set F_{3M+} had 10 features and the set F_{2M} had 7 features after we removed "noncranial radiation" (Textbox 4).

Textbox 4. Nonredundant features in the set F3M+ and set F2M.

F$_{3M+}$

- Physical fatigue
- Overall fatigue
- Cognitive fatigue
- Family communication
- Sleep fatigue
- Family conflict
- Family mutuality
- Physical activity
- IV high-dose methotrexate (MTX)
- Intrathecal MTX dose

F$_{2M}$

- Leukemia risk group
- Living space
- Family concern
- Cranial radiation therapy
- Years of education
- Parental control
- Age at cancer diagnosis

At this step, we checked if the sum of features from F_{3M+} and F_{2M} was <25. After removing redundant features, we still had 17 features, which was greater than N=15; thus, we do not need to go back to step 1 in stage 1A to find at least 15 features. We can then proceed to stage 2.

Stage 2: A DDN

Overview

In the second stage, the deep neural network had a dropout parameter, where neurons are randomly ignored during construction of the neural network, to avoid model overfitting, which is a problem that the existing wrapper methods have. Thus, our methodology is better suited for finding the best features from the high-dimension, low-sample size dataset. More specifically, after the features were processed by our multimetric majority-voting filter, we passed all the nonredundant features to the deep dropout neural (DDN) network that was designed to determine whether adding any of those features selected by the only 2 metrics to the list of the features selected by at least 3 metrics resulted in a higher F_1-score. Note that this step was not conducted if the number

of the nonredundant features, that is, those features that were already selected by at least 3 metrics in stage 1, had met the domain experts' expectation. Our designed DDN network was a 2-hidden– and 1-output–layer architecture. Due to the limited number of patients' medical records with many input features, our DDN network was likely to quickly overfit a training dataset. To address this issue, we used the grid search algorithm with the K-fold cross-validation (CV) to find the best dropout rate for our network. We also dynamically set the network's hidden layer size using the formula , where I is the number of selected input subset features and O is the number of labels per behavioral outcome [27]. For the remaining network's initialization parameters, default values were used [28]. The goal was to perform the hyperparameter tuning using the grid search algorithm with the K-fold CV to obtain the optimal parameters' values, including the dropout rate, all the network's parameters, and the size of each hidden layer [29].

Specifically, the subset of features selected by ≥3 metrics in stage 1 was used in building the initial network architecture to produce the baseline F_1-score. This baseline F_1-score tells us how well the network predicts that a cancer survivor will

develop the behavioral outcome of interest, using only the features selected by at least 3 metrics. Afterward, we wanted to see whether adding any subset of features selected by 2 metrics would improve the baseline F_1-score. To achieve this, we tried different combinations among the features selected by 2 metrics; added them on top of the features selected by at least 3 metrics; used all those features to build, train, and optimize our network using the grid search algorithm with the K-fold CV to obtain the optimal parameters' values; and then recorded each new F_1-score. This allowed us to compare F_1-scores between the baseline and the baseline plus additional subsets of features. If any of the new F_1-scores were higher than the baseline, then our final feature list was the one that produced the highest F_1-score. If none of the new F_1-scores were higher than the baseline, then our final feature list was simply the baseline features, that is, the features selected by at least 3 metrics. A step-by-step pseudocode algorithm for our DDN network is detailed in Multimedia Appendix 2).

Let us use our dataset as an example to explain our DDN network. At this stage, we wanted to determine whether any features selected by 2 metrics should be kept in the final feature list on top of the features selected by at least 3 metrics. Our input included the following:

1. F_{3M+} and F_{2M}, which were our outputs from stage 1B.
2. *Drop_Out_Rate*, a set of fine-tuning dropout rates for building a DDN network.
3. *D_Train*, which was the training dataset that only included features in F_{3M+}
4. Z, the set that included all possible subsets from F_{2M}, excluding the null set, where the size of subsets was less than or equal to N minus the size of F_{3M+} so that the total number of features does not exceed N. In our example, the set Z only included all the possible subsets of size ≤5 because we already had 10 features in non_redundant_three_more and N minus 10 was 5. Given that there were 7 features in non_redundant_two, there were 128 possible subsets. However, because we only needed the subsets with size ≤5 and we also excluded the null set, we ended up with a total of 119 different subsets in the set Z.
5. M, a set of lists that add all the possible subsets in the set Z to the set F_{3M+}; thus, there were 119 different lists.
6. *E_Train*, which is the set of training datasets that includes features in each list in M.
7. K, the number of training partitions on *D_Train* and *E_Train* for performing CV.

Step 1

We wanted to find the best dropout rate for the neural network, using the *grid-search* technique, F_1-score, and K-fold CV, on *D_Train*, F_{3M+}, *B_Outcome*, and *Drop_Out_Rate* of a DDN network. K was set to 5. We thus first constructed a neural network using the *create_DDN* function to perform the *grid-search* technique. The neural network was initialized to have a learning rate of 0.001, 500 epochs, used the "Adam" optimizer, used the "Binary Cross Entropy" loss function, had

2 hidden layers with number of neurons and the "Relu" activation function, and 1 output layer with 1 neuron and the activation function "Sigmoid." Suppose using the *grid-search* technique with the *D_Train* training dataset, the F_{3M+} feature set, the *B_Outcome* thought problems, the set of fine-tuning dropout rates *Drop_Out_Rate*, and using 5-fold CV, we found that the best dropout rate was 0.1 (*bestDropOutRate* was set to 0.1).

Step 2

We constructed a deep neural network with the initialized attributes in the *create_DDN* function, *bestDropOutRate*, *D_Train*, F_{3M+}, and *B_Outcome*, and then performed 5-fold CV to obtain the baseline F_1-score, $F1_{Baseline}$.

Step 3

We then iterated through each feature set (ie, $F_{3M+}+Z_r$) in *M* and constructed a deep neural network with the same initialized attributes in the *create_DDN* function, *bestDropOutRate*, *E_Train*, $F_{3M+}+Z_r$, and *B_Outcome*, and then performed 5-fold CV to obtain the F_1-score, F1, for each training dataset in *E_Train*. The hidden layer size of each neural network was calculated using the number of features in M+1, divided by 2. If any F_1-score was greater than $F1_{Baseline}$, the final feature list (ie, *Final_Features*) was set to the feature set (ie, $F_{3M+}+Z_r$) in M in which the F_1-score was obtained.

Step 4

We had the feature list with the best F_1-score (ie, *Final_Features*), which was passed into 3 ML classifiers: logistic regression, naive Bayes, and k-nearest neighbors.

Pilot Experimental Study

In our experimental study, we used a 2018 to 2020 dataset that contained 102 ALL survivors' clinical records collected from a public hospital in Hong Kong. The survivors were aged between 15 and 39 years, had completed treatment, and were >5 years postcancer diagnosis at the time of recruitment. In each patient record, there were >50 features, including demographic factors (eg, age, gender, and education level), cancer treatments received (eg, radiation, chemoradiotherapy, and surgery), inflammatory biomarkers (eg, interleukin-7, monocyte chemoattractant protein-1, and tumor necrosis factor alpha-α), physical health conditions (eg, BMI, sleep fatigue, and cognitive fatigue), family life and socioeconomic descriptors (eg, family conflict, family communication and living space), attention-related outcomes (eg, measures of inattentiveness, impulsivity, and sustained attention), and lifestyle habits (eg, drinking, smoking, and physical activity). The features were obtained from a behavioral assessment that included the traditional Chinese version of the Achenbach System of Empirically Based Assessment youth self-report checklist. It consisted of syndrome scales measuring attention problems, thought problems, internalizing problems (eg, somatic complaints, anxiety and depressive symptoms, and withdrawn behavior), externalizing problems (eg, aggressive behavior, intrusive behavior, and rule-breaking behavior), and sluggish cognitive tempo. The Achenbach System of Empirically Based

Assessment measures were previously validated and used in the local young adult cancer population [9,30]. The inclusion of these features specifically in patient records was based on existing evidence in the literature and data from the local study cohort. The features predicting behavioral outcomes included clinical factors (eg, leukemia risk group, age at cancer diagnosis, and neurological complications), treatment factors (eg, cranial radiation therapy, intrathecal methotrexate dose, intravenous high-dose methotrexate, and hematopoietic stem cell transplant), socioenvironmental factors (eg, living space and family functioning), and lifestyle factors (eg, physical activity and sleep fatigue) [9,30-34].

After preprocessing the data and using our 2-stage feature selection algorithm, we selected 15 input features, expected by our medical investigators, to train and test our 3 ML classifiers, that is, logistic regression, naive Bayes, and k-nearest neighbors, to predict 6 behavioral outcomes (ie, anxiety and depression, thought problems, attention problems, internalizing problems, externalizing problems, and sluggish cognitive tempo) that our medical investigators would like to focus on. Due to their

clinical importance recommended by our medical investigators, we also added 3 more clinically relevant features (ie, gender, current age, and age at diagnosis) to the final feature list if those features had not been already selected by our 2-stage feature selection approach.

Ethical Considerations

Approval of this study was obtained from the Joint Chinese University of Hong Kong – New Territories East Cluster Clinical Research Ethics Committee (2017.701). Written informed consent was obtained from all participants.

Results

Overview

The experimental results included the F_1-score, precision, and recall on the testing data (Table 1). Note that for each feature selection method category, those scores are the average values of prediction performance among all the 3 ML classifiers for every behavioral outcome.

Table 1. Average F1-scores.

Behavioral outcome	Filter	Wrapper	Embedded	Hybrid	Our method	Percentage change (our method vs highest baseline)
Anxiety and depression						
F_1-score	0.624	0.437	0.585	0.449	*0.738* [a]	+18.27
Precision score	0.562	0.407	0.563	0.424	*0.708*	+25.75
Recall score	0.813	0.519	0.630	0.580	0.778	–4.31
Thought problems						
F_1-score	0.490	0.438	0.477	0.394	*0.511*	+4.29
Precision score	0.522	0.385	0.590	0.496	0.448	–24.07
Recall score	0.537	0.556	0.463	0.383	*0.611*	+9.89
Attention problems						
F_1-score	0.348	0.417	0.440	0.350	*0.568*	+29.10
Precision score	0.290	0.360	0.350	0.329	*0.515*	+43.10
Recall score	0.463	0.519	0.630	0.424	*0.667*	+5.87
Internalizing problems						
F_1-score	0.533	0.706	0.619	0.637	0.700	–0.85
Precision score	0.583	0.665	0.665	0.668	0.618	–7.49
Recall score	0.587	0.762	0.651	0.651	*0.857*	+12.47
Externalizing problems						
F_1-score	0.219	0.459	0.267	0.265	0.278	–39.43
Precision score	0.230	0.417	0.278	0.297	*0.444*	+6.47
Recall score	0.222	0.556	0.259	0.259	0.222	–60.07
Sluggish cognitive tempo						
F_1-score	0.560	0.463	0.582	0.489	*0.639*	+9.79
Precision score	0.542	0.409	0.577	0.494	0.570	–1.21
Recall score	0.654	0.568	0.617	0.568	*0.741*	+13.30

[a]Italicized values indicate that our score was higher than the other 4 methods.

XSL•FO

RenderX

Our 2-stage feature selection approach outperformed or leveled the existing feature selection methods to support the prediction of 5 out of 6 behavioral outcomes (ie, anxiety and depression, thought problems, attention problems, internalizing problems, and sluggish cognitive tempo) in terms of the average F_1-scores (Table 1). Although the wrapper method outperformed our feature selection approach to support the prediction of externalizing problems, our approach's performance was more stable, as the F_1-score variance was smaller. Thus, our feature selection approach still outperforms the other 3 existing feature selection methods.

In addition, our feature selection approach outperformed or leveled the existing feature selection methods to support the prediction of 5 out of 6 behavioral outcomes (ie, anxiety and depression, attention problems, internalizing problems, externalizing problems, and sluggish cognitive tempo) in terms of precision scores (Table 1). Although the embedded method outperformed our feature selection approach to support the prediction of thought problems, our approach's performance variance was much smaller, which implies our approach was more stable.

Finally, our feature selection approach outperformed the existing feature selection methods to support the prediction of 4 out of 6 behavioral outcomes (ie, thought problems, attention problems, internalizing problems, and sluggish cognitive tempo) in terms of recall scores (Table 1). Although the filter and wrapper method outperformed our feature selection approach to support the prediction of anxiety and depression and externalizing problems, our approach's performance variance was much smaller as well.

As the F_1-scores were calculated from both precision and recall scores, we can infer that our feature selection approach improves the F_1-scores largely because it increases the recall scores as opposed to the precision scores (Table 1). Overall, the experimental results show promising evidence that our method improves the ML classifiers' prediction performance to support better early detection of long-term behavioral outcomes in survivors of cancer.

Radial Feature Charts

Radial feature charts were generated for each of the 6 behavioral outcomes analyzed, including anxiety and depression, thought problems, attention problems, internalizing problems, externalizing problems, and sluggish cognitive tempo (Figure 3). Each chart includes the top 15-plus features selected by our proposed methodology. The size and the color of each red slice is measured by the unified metric value of each feature, which is calculated by averaging the scores of the metrics that select each feature during stage 1A of our proposed method.

The variables represent the documented risk factors associated with the development of behavioral problems in the literature. They include (1) sociodemographic variables (ie, age at evaluation and gender), (2) clinical variables (ie, age at cancer diagnosis, intrathecal chemotherapy, intrathecal methotrexate dose, IV high-dose methotrexate, and inflammatory interleukin-8 levels), and (3) socioenvironmental and lifestyle variables (ie, sleep, fatigue, physical activity, and family functioning). Physicians can interpret the charts by seeing which features have the darkest color and largest size, indicating higher unified metric values and thus greater associations with the behavioral problem of interest. Those features can then be further used to devise customized prevention plans and advice.

Figure 3. Radial feature charts.
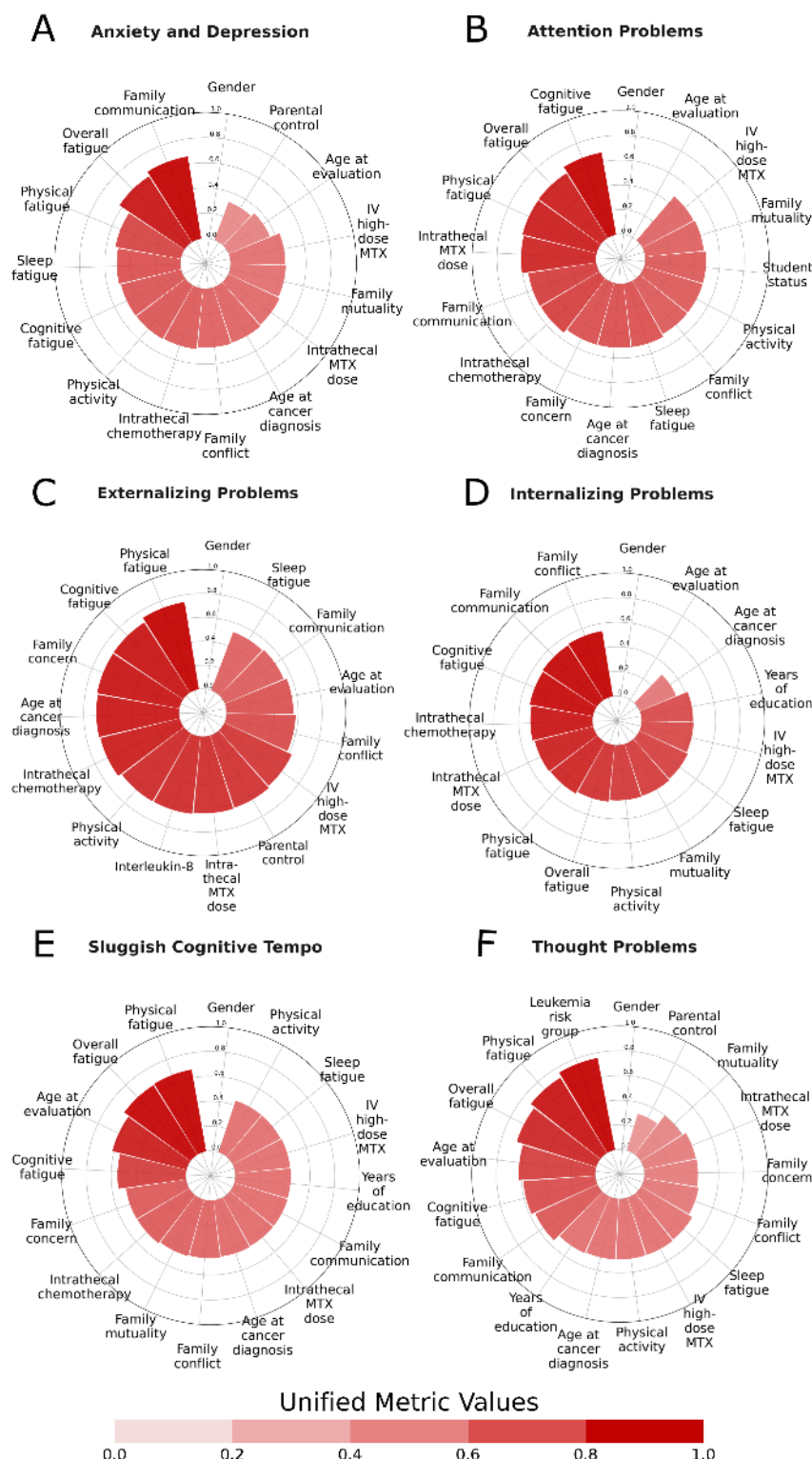
Discussion

Principal Findings

In this work, we sought to develop a prognostic ML framework and feature selection approach to predict the trajectory of functional outcomes in a specific population: survivors of ALL. Our hybrid deep learning–based feature selection approach outperforms or equals the existing feature selection methods

assessed (ie, filter, wrapper, embedded, and hybrid) for 5 out of 6 long-term behavioral outcomes. Even in cases where our feature selection method did not outperform existing methods, our approach's performance variance was much smaller and thus more stable. We observed that the performance of the model was significantly weaker in predicting externalizing problems than internalizing problems. This may be attributed to the complex phenotypic nature of externalizing behaviors, such as antisocial or aggressive behaviors and conduct problems. In

addition, there are other factors that may predict externalizing problems that were not considered in this study. For example, our previous work showed that increased screen time during the COVID-19 pandemic was associated with inattentiveness and impulsivity in pediatric survivors of cancer in China, but screen time was not included in the data [35]. Social support and rehabilitation, which are important interventions addressing behavioral functioning and mental health in young Chinese survivors of cancer, were also not assessed in this study [36]. From the data, we infer that our feature selection approach improves F_1-scores from ML classifiers compared to existing feature selection methods largely because it increases the recall scores as opposed to the precision scores. We also developed radial feature charts that can quickly and effectively help clinicians understand which predictor variables were most important in predicting long-term behavioral outcomes. Overall, the experimental results show promising evidence that our method improves ML classifiers' prediction performance on high-dimension low-sample size data, which can support better early detection of long-term behavioral outcomes in survivors of cancer.

Limitations

Our study was limited to a pilot study with young Chinese survivors of leukemia. As one's neurodevelopment and social skills are often dependent on cultural norms, our findings may not be extrapolated or applicable to other populations. However, the contemporary treatment for childhood ALL is similar in most countries or regions, consisting of high-dose methotrexate, intrathecal chemotherapy, and a standard set of intravenous and oral chemotherapy drugs as the backbone. Therefore, we reasoned that our findings may still be generalizable to the existing population of individuals in the health care system of Hong Kong who have survived leukemia over the past decade. In addition, although clinical domain experts assisted with additional input for the features that were kept in ML classifiers, there remains room for human error, and domain experts' opinions may occasionally differ from what features would optimize ML classifiers' performance. Furthermore, as this is a cross-sectional study, it was not possible to delineate the causal relationship between the risk factors and behavioral outcomes. The model developed through this study should be validated in a larger cohort with prospective collection of outcome data to better reflect the trajectories of functional outcomes in these young survivors as they advance from young to middle adulthood. Finally, additional biases may have influenced the data, such as those related to patients who had access to hospital care and were willing to share their data with our clinical investigators.

Comparison With Prior Work

Our findings reinforce existing evidence that adverse behavioral outcomes in survivors of cancer are a complex and multifactorial phenotype. Most preexisting research is focused on either disease- or treatment-related factors as predictors of cognitive dysfunction. However, socioenvironmental factors play an important role in the neurodevelopment of these young survivors. Our findings showed the interaction and unique contribution of the socioenvironmental factors, such as family

dynamics and lifestyle factors, on anxiety, depression, and sluggish cognitive symptoms in survivors. Studies have found associations of parents' psychological distress on the child's cognitive and behavioral outcomes [8,37]. Environmental events can elicit a biological stress response that results in neurological reactions to that stress. This is especially relevant in the context of Hong Kong and Mainland China, where much emphasis is now placed on ameliorating the adverse health effects of the urban environment in children and adolescents. The findings provide directions for the development of multidisciplinary services and interventions. For example, social workers can pay more attention to the occupational or employment challenges of young survivors who experience fatigue symptoms from treatment and manifest adverse behavioral outcomes. The study findings can help us identify high-risk subgroups from dysfunctional families or households struggling with financial problems and conflicts. Interventions that promote self-confidence and positive peer interaction can be implemented during the early survivorship phase when young survivors transit back to their full-time school or work.

Our results also build upon existing computational methods and feature selection approaches for predicting behavioral outcomes in survivors of cancer. Traditional computational methods in the clinical and social sciences typically use regression analysis to model the relationship between ≥2 variables for prediction. However, modeling human behavioral data is challenging due to its multifactorial nature, heterogeneity, nonlinearity of data, and class imbalance [10,11]. As a result, the model can only account for a small proportion of variance, with limited utility in clinical settings. For example, we have reported that cranial radiation, chronic health conditions, and poor physical activity are associated with worse cognitive and behavioral outcomes in Chinese survivors of childhood leukemia [9]. However, these factors only accounted for 22.9% to 35.8% of the variance in the traditional regression models. Identifying an effective computational method that minimizes algorithmic bias, such as the 2-stage feature selection algorithm within the clinical domain–guided framework outlined in this study, can maximize the use of clinical and behavioral data for predictive purposes. Such prognostic models will aid in informing strategies aimed at changing behavior and designing social and clinical interventions.

Conclusions

Future studies can validate our prediction model in other Chinese populations of survivors of cancer sharing similar cultural norms in mainland China and Taiwan, as well as validate the model in larger samples with a longitudinal prospective cohort study design. In addition, studies can further investigate the real-world feasibility of incorporating such algorithms into health care systems as risk stratification tools to assist clinicians and psychologists in identifying patients at risk of adverse behavioral outcomes. Incorporation of diverse populations, larger sample sizes, and similar prediction models in future studies may provide deeper insights into the interaction among clinical, treatment, socioeconomic, and lifestyle factors and their impact on functional outcomes, ultimately enabling the incorporation of such multifactorial insights to improve strategies for the personalized care of patients with cancer.

Given that we are working with such small cancer survivor datasets, even a slight improvement in prediction performance from ML classifiers can make a substantial difference in helping survivors of cancer. Our data-driven, clinical domain–guided approach can potentially address the problem of "high dimension low sample size." The pilot analysis shows that this approach has allowed us to identify a set of interacting clinical and socioenvironmental characteristics that predicted behavioral outcomes in survivors.

In late 2019, the American Cancer Society had a special call for attention to financial, social, and emotional concerns that uniquely affect young survivors of cancer [38]. Currently, in Hong Kong, there are no centralized cancer programs for adolescent and young adult patients. From a clinical perspective, identifying the unique factors associated with interindividual differences in functional outcomes will help clinicians to identify individualized modifiable risk factors. This will contribute to the development of a personalized, patient-centered cancer care program for local patients with cancer. From a research perspective, this project serves as a pilot study to apply ML-based prognostic technology, guided by clinical knowledge, on a combination of objective data (ie, clinical and demographics variables) and subjective data (ie, behavioral and patient-reported variables). The framework and algorithms developed through this analysis can be applied to address clinically relevant research questions in patients with other chronic diseases. The aim of this application is in line with the recent call by the government of the Hong Kong Special Administrative Region to harness data-driven analytics to formulate health care policies [39].

Acknowledgments

Conflicts of Interest

None declared.

Multimedia Appendix 1
Step-by-step pseudocode algorithm for the multimetric, majority-voting filter.
[DOCX File , 22 KB - bioinform_v6i1e65001_app1.docx]

Multimedia Appendix 2
Step-by-step pseudocode algorithm for the DDN network.
[DOCX File , 18 KB - bioinform_v6i1e65001_app2.docx]

References

1. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. CA Cancer J Clin 2023 Jan;73(1):17-48 [FREE Full text] [doi: 10.3322/caac.21763] [Medline: 36633525]

2. Brinkman TM, Recklitis CJ, Michel G, Grootenhuis MA, Klosky JL. Psychological symptoms, social outcomes, socioeconomic attainment, and health behaviors among survivors of childhood cancer: current state of the literature. J Clin Oncol 2018 Jul 20;36(21):2190-2197 [FREE Full text] [doi: 10.1200/JCO.2017.76.5552] [Medline: 29874134]

3. Yeh JM, Ward ZJ, Chaudhry A, Liu Q, Yasui Y, Armstrong GT, et al. Life expectancy of adult survivors of childhood cancer over 3 decades. JAMA Oncol 2020 Mar 01;6(3):350-357 [FREE Full text] [doi: 10.1001/jamaoncol.2019.5582] [Medline: 31895405]

4. Dixon SB, Liu Q, Chow EJ, Oeffinger KC, Nathan PC, Howell RM, et al. Specific causes of excess late mortality and association with modifiable risk factors among survivors of childhood cancer: a report from the Childhood Cancer Survivor Study cohort. Lancet 2023 Apr 29;401(10386):1447-1457 [FREE Full text] [doi: 10.1016/S0140-6736(22)02471-0] [Medline: 37030315]

5. Lam CS, Lee CP, Chan JW, Cheung YT. Prescription of psychotropic medications after diagnosis of cancer and the associations with risk of mortality in Chinese patients: a population-based cohort study. Asian J Psychiatr 2022 Dec;78:103290. [doi: 10.1016/j.ajp.2022.103290] [Medline: 36209707]

6. Suh E, Stratton KL, Leisenring WM, Nathan PC, Ford JS, Freyer DR, et al. Late mortality and chronic health conditions in long-term survivors of early-adolescent and young adult cancers: a retrospective cohort analysis from the Childhood Cancer Survivor Study. Lancet Oncol 2020 Mar;21(3):421-435 [FREE Full text] [doi: 10.1016/S1470-2045(19)30800-9] [Medline: 32066543]

7. Alias H, Morthy SK, Zakaria SZ, Muda Z, Tamil AM. Behavioral outcome among survivors of childhood brain tumor: a case control study. BMC Pediatr 2020 Feb 05;20(1):53 [FREE Full text] [doi: 10.1186/s12887-020-1951-3] [Medline: 32020861]

8. Patel SK, Wong AL, Cuevas M, Van Horn H. Parenting stress and neurocognitive late effects in childhood cancer survivors. Psychooncology 2013 Aug 25;22(8):1774-1782 [FREE Full text] [doi: 10.1002/pon.3213] [Medline: 23097416]

9. Peng L, Yang LS, Yam P, Lam CS, Chan AS, Li CK, et al. Neurocognitive and behavioral outcomes of Chinese survivors of childhood lymphoblastic leukemia. Front Oncol 2021;11:655669 [FREE Full text] [doi: 10.3389/fonc.2021.655669] [Medline: 33959507]

10. Kliegr T, Bahník Š, Fürnkranz J. Advances in machine learning for the behavioral sciences. Am Behav Sci 2019 Jul 24;64(2):145-175. [doi: 10.1177/0002764219859639]

11. Turgeon S, Lanovaz MJ. Tutorial: applying machine learning in behavioral research. Perspect Behav Sci 2020 Dec 10;43(4):697-723 [FREE Full text] [doi: 10.1007/s40614-020-00270-y] [Medline: 33381685]

12. Thejas GS, Garg R, Iyengar SS, Sunitha NR, Badrinath P, Chennupati S. Metric and accuracy ranked feature inclusion: hybrids of filter and wrapper feature selection approaches. IEEE Access 2021;9:128687-128701. [doi: 10.1109/access.2021.3112169]

13. Cherrington M, Thabtah F, Lu J, Xu Q. Feature selection: filter methods performance challenges. In: Proceedings of the 2019 International Conference on Computer and Information Sciences. 20192019 Presented at: ICCIS '19; April 3-4, 2019; Sakaka, Saudi Arabia p. 1-4 URL: https://ieeexplore.ieee.org/document/8716478 [doi: 10.1109/iccisci.2019.8716478]

14. Lin X, Li C, Ren W, Luo X, Qi Y. A new feature selection method based on symmetrical uncertainty and interaction gain. Comput Biol Chem 2019 Dec;83:107149. [doi: 10.1016/j.compbiolchem.2019.107149] [Medline: 31751882]

15. Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. Applications of support vector machine (SVM) learning in cancer genomics. Cancer Genomics Proteomics 2018 Jan 2;15(1):41-51 [FREE Full text] [doi: 10.21873/cgp.20063] [Medline: 29275361]

16. Jonas R, Cook J. LASSO regression. Br J Surg 2018;105(10):1348. [doi: 10.1002/bjs.10895]

17. Hoerl RW. Ridge regression: a historical context. Technometrics 2020 Oct 23;62(4):420-425. [doi: 10.1080/00401706.2020.1742207]

18. Alhamzawi R, Ali HT. The Bayesian elastic net regression. Commun Stat Simul Comput 2017 Jun 20;47(4):1168-1178. [doi: 10.1080/03610918.2017.1307399]

19. Usman AU, Hassan S, Tukur K. Application of dummy variables in multiple regression analysis. Int J Recent Sci Res 2015;7(11):7440-7442 [FREE Full text] [doi: 10.4324/9781315748788-15]

20. Berger VW, Zhou Y. Kolmogorov–Smirnov test: overview. Wiley StatsRef 2014:63 [FREE Full text] [doi: 10.1002/9781118445112.stat06558]

21. González-Estrada E, Cosmes W. Shapiro–Wilk test for skew normal distributions based on data transformations. J Stat Comput Simu 2019 Aug 27;89(17):3258-3272. [doi: 10.1080/00949655.2019.1658763]

22. Saculinggan M, Balase EA. Empirical power comparison of goodness of fit tests for normality in the presence of outliers. J Phys Conf Ser 2013 Apr 26;435:012041. [doi: 10.1088/1742-6596/435/1/012041]

23. Patro S. Normalization: a preprocessing stage. arXiv Preprint posted online March 19, 2015 [FREE Full text]

24. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell 2002 Jun 01;16:321-357. [doi: 10.1613/jair.953]

25. Akoglu H. User's guide to correlation coefficients. Turk J Emerg Med 2018 Sep;18(3):91-93 [FREE Full text] [doi: 10.1016/j.tjem.2018.08.001] [Medline: 30191186]

26. Kornbrot D. Point biserial correlation. Wiley StatsRef 2014:22. [doi: 10.1002/9781118445112.stat06227]

27. Lawrence S, Giles CL, Tsoi AC. What size neural network gives optimal generalization? Convergence properties of backpropagation. Institute for Advanced Computer Studies, University of Maryland. 1998. URL: https://api.drum.lib.umd.edu/server/api/core/bitstreams/bf781aeb-eb41-4803-a2ac-7d915d0ac791/content [accessed 2024-04-29]

28. Zollanvari A. Deep learning with Keras-TensorFlow. In: Zollanvari A, editor. Machine Learning With Python: Theory and Implementation. Cham, Switzerland: Springer; 2023:351-391.

29. Belete DM, Huchaiah MD. Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. Int J Comput Appl 2021 Sep 12;44(9):875-886. [doi: 10.1080/1206212X.2021.1974663]

30. Cheung YT, Ma CT, Li MC, Zhou KR, Loong HH, Chan AS, et al. Associations between lifestyle factors and neurocognitive impairment among Chinese adolescent and young adult (AYA) survivors of sarcoma. Cancers (Basel) 2023 Jan 28;15(3):799 [FREE Full text] [doi: 10.3390/cancers15030799] [Medline: 36765757]

31. Cheung YT, To KK, Hua R, Lee CP, Chan AS, Li CK. Association of markers of inflammation on attention and neurobehavioral outcomes in survivors of childhood acute lymphoblastic leukemia. Front Oncol 2023 Jun 21;13:1117096 [FREE Full text] [doi: 10.3389/fonc.2023.1117096] [Medline: 37416531]

32. Krull KR, Hardy KK, Kahalley LS, Schuitema I, Kesler SR. Neurocognitive outcomes and interventions in long-term survivors of childhood cancer. J Clin Oncol 2018 Jul 20;36(21):2181-2189 [FREE Full text] [doi: 10.1200/JCO.2017.76.4696] [Medline: 29874137]

33. Mavrea K, Efthymiou V, Katsibardi K, Tsarouhas K, Kanaka-Gantenbein C, Spandidos D, et al. Cognitive function of children and adolescent survivors of acute lymphoblastic leukemia: a meta-analysis. Oncol Lett 2021 Apr 05;21(4):262 [FREE Full text] [doi: 10.3892/ol.2021.12523] [Medline: 33664825]

34. van der Plas E, Modi AJ, Li CK, Krull KR, Cheung YT. Cognitive impairment in survivors of pediatric acute lymphoblastic leukemia treated with chemotherapy only. J Clin Oncol 2021 Jun 01;39(16):1705-1717. [doi: 10.1200/JCO.20.02322] [Medline: 33886368]

35. Cai J, Cheung YT, Au-Doung PL, Hu W, Gao Y, Zhang H, et al. Psychosocial outcomes in Chinese survivors of pediatric cancers or bone marrow failure disorders: a single-center study. PLoS One 2022 Dec 13;17(12):e0279112 [FREE Full text] [doi: 10.1371/journal.pone.0279112] [Medline: 36512620]
36. Zhu W. The impact of social support on the mental health of cancer patients: evidence from China. Psycho Oncol 2024;18(1):69-77. [doi: 10.32604/po.2023.046593]
37. Hile S, Erickson SJ, Agee B, Annett RD. Parental stress predicts functional outcome in pediatric cancer survivors. Psychooncology 2014 Oct 10;23(10):1157-1164. [doi: 10.1002/pon.3543] [Medline: 24817624]
38. Bhatia S, Pappo AS, Acquazzino M, Allen-Rhoades WA, Barnett M, Borinstein S, et al. Adolescent and Young Adult (AYA) oncology, version 2.2024, NCCN clinical practice guidelines in oncology. J Natl Compr Canc Netw 2023 Aug;21(8):851-880. [doi: 10.6004/jnccn.2023.0040] [Medline: 37549914]
39. Leung KY, Lee HY. Implementing the smart city: who has a say? Some insights from Hong Kong. Int J Urban Sci 2021 Nov 08;27(sup1):124-148. [doi: 10.1080/12265934.2021.1997634]

Abbreviations

ALL: acute lymphocytic leukemia
CFS: correlation-based feature selection
CS: correlation score
CV: cross-validation
DDN: deep dropout neural network
GI: Gini index
IG: information gain
Lasso: least absolute shrinkage and selection operator
MIC: maximal information coefficient
ML: machine learning
MRMR: maximum relevance minimum redundancy
SBS: sequential backwards selection
SFS: sequential forward selection
SMOTE-NC: synthetic minority oversampling technique for nominal and continuous
SS: stepwise selection

XSL•FO
RenderX

XSL•FO

RenderX