EBM Experimental Biology and Medicine

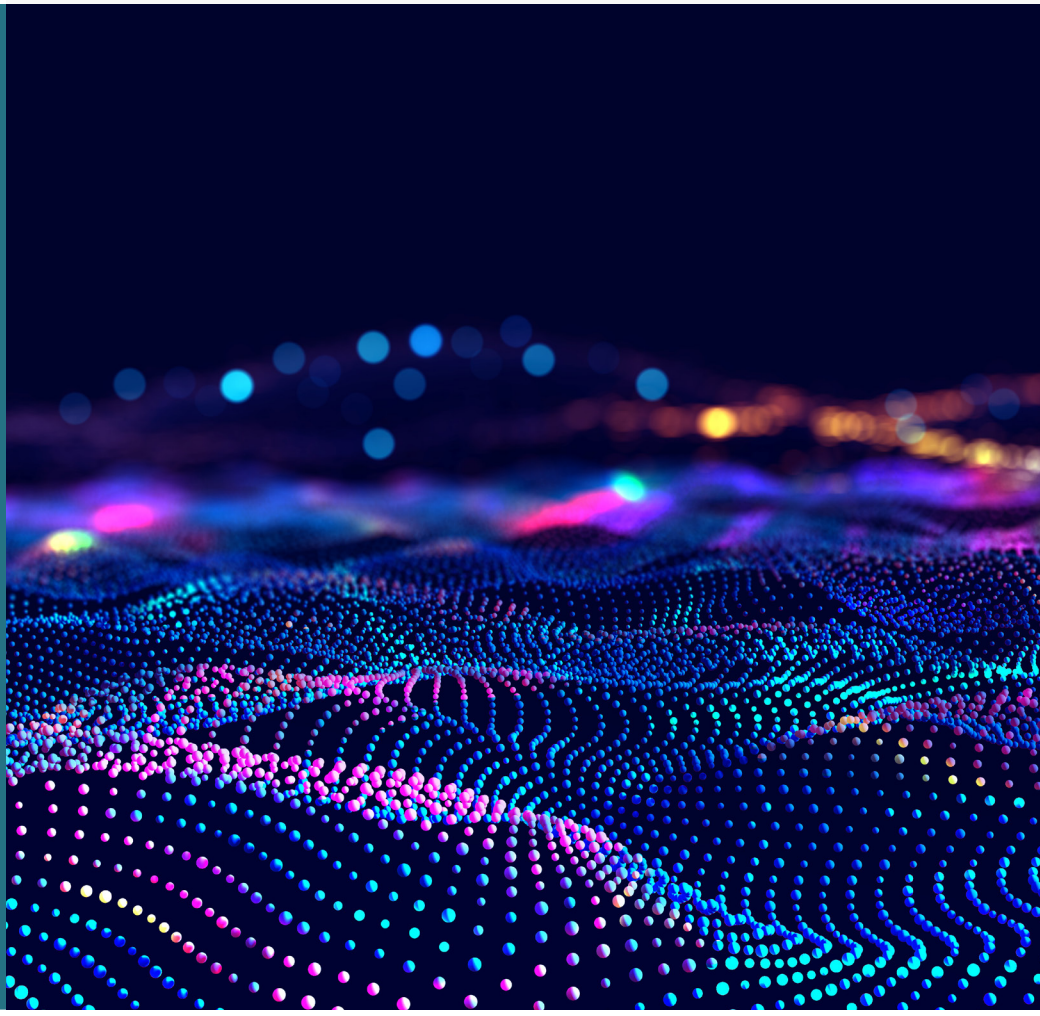
Special Issue

Proceedings of the 10th Annual Conference of the Arkansas Bioinformatics Consortium (AR-BIC) - Real-World Impact of AI

Issue Editors

Huixiao Hong
United States Food and Drug Administration, United States

William Slikker Jr.
Retired, United States



ISBN 978-2-8325-6740-1
DOI 10.3389/978-2-8325-6740-1

frontiers | Publishing Partnerships

EBM Experimental Biology and Medicine

Proceedings of the 10th Annual Conference of the Arkansas Bioinformatics Consortium (AR-BIC) - Real-World Impact of AI

ISSN 1535-3699
ISBN 978-2-8325-6740-1
DOI 10.3389/978-2-8325-6740-1

Generative AI statement

Any alternative text (Alt text) provided alongside figures in the articles in this ebook has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Table of contents

04 Realizing Impact of Artificial Intelligence in Real World Enhances Public Health

Huixiao Hong and William Slikker Jr.

08 Development of a comprehensive open access “molecules with androgenic activity resource (MAAR)” to facilitate risk assessment of chemicals

Fan Dong, Barry Hardy, Jie Liu, Tomaz Mohoric, Wenjing Guo, Thomas Exner, Weida Tong, Joh Dohler, Daniel Bachler and Huixiao Hong

18 Leveraging AI to improve disease screening among American Indians: insights from the Strong Heart Study

Paul Rogers, Thomas McCall, Ying Zhang, Jessica Reese, Dong Wang and Weida Tong

26 Developing predictive models for μ opioid receptor binding using machine learning and deep learning techniques

Jie Liu, Jerry Li, Zoe Li, Fan Dong, Wenjing Guo, Weigong Ge, Tucker A. Patterson and Huixiao Hong

40 A refined set of RxNorm drug names for enhancing unstructured data analysis in drug safety surveillance

Wenjing Guo, Fan Dong, Jie Liu, Aasma Aslam, Tucker A. Patterson and Huixiao Hong

51 AI-powered topic modeling: comparing LDA and BERTopic in analyzing opioid-related cardiovascular risks in women

Li Ma, Ru Chen, Weigong Ge, Paul Rogers, Beverly Lyn-Cook, Huixiao Hong, Weida Tong, Ningning Wu and Wen Zou

62 Enhancing pharmacogenomic data accessibility and drug safety with large language models: a case study with Llama3.1

Dan Li, Leihong Wu, Ying-Chi Lin, Ho-Yin Huang, Ebony Cotton, Qi Liu, Ru Chen, Ruihao Huang, Yifan Zhang and Joshua Xu

71 Optimal transport reveals immune perturbation and fingerprints over time in COVID-19 vaccination

Zexuan Wang, Jiong Chen, Matei Ionita, Qipeng Zhan, Zhuoping Zhou and Li Shen

82 Effect of in utero and lactational exposure to antiretroviral therapy on the gut microbial composition and metabolic function in aged rat offspring

Chandra Mohan Reddy Muthumula, Yaswanthi Yanamadala, Kuppan Gokulan, Kumari Karn, Helen Cunny, Vicki Sutherland, Janine H. Santos and Sangeeta Khare

EBM Experimental Biology and Medicine

TYPE Editorial
PUBLISHED 27 June 2025
DOI 10.3389/ebm.2025.10700

OPEN ACCESS

*CORRESPONDENCE
Huixiao Hong,
huixiao.hong@fda.hhs.gov
William Slikker Jr.,
billslikkerjr@gmail.com

RECEIVED 09 June 2025
ACCEPTED 12 June 2025
PUBLISHED 27 June 2025

CITATION
Hong H, Slikker W Jr. (2025) Realizing Impact of Artificial Intelligence in Real World Enhances Public Health. *Exp. Biol. Med.* 250:10700. doi: 10.3389/ebm.2025.10700

Realizing Impact of Artificial Intelligence in Real World Enhances Public Health

Huixiao Hong[1]* and William Slikker Jr.[2]*

[1]Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, US Food and Drug Administration, Jefferson, AR, United States, [2]Retired, Formerly from the National Center for Toxicological Research, US Food and Drug Administration, Jefferson, AR, United States

KEYWORDS

AI, bioinformatics, machine learning, deep learning, data

Editorial on the Research Topic
Realizing Impact of Artificial Intelligence in Real World Enhances Public Health

This thematic issue is a product of the 10th annual conference of the Arkansas Bioinformatics Consortium (AR-BIC), which was held on February 26–27, 2024, in Little Rock, Arkansas, with a theme "Real World Impact of AI". This conference gathered more than 200 scientists and trainees with diverse scientific interests discussing current research works and future perspectives on realizing the impact of artificial intelligence (AI) in the real world. The conference hosted three pre-conference workshops that provided the attendees with state-of-art knowledge and tools on real world applications of AI, including imaging and genomic data analysis. In the plenary presentations, distinguished scientists gave perspectives on how AI transforms toxicology, drug discovery, and public health, calling attentions to both emerging opportunities and practical applications. Four breakout sessions covered presentations from front-line experts to share their cutting-edge research on application of AI to various vital fields, such as natural language processing (NLP) for pharmacovigilance, ethical frameworks for responsible AI, and novel models in spatial omics and computational toxicology. Of special note is the high-profile dialogue with former US FDA chief scientist Dr. Namandjé Bumpus and the NIEHS/NTP director Dr. Richard Woychik on applications of AI in regulatory sciences. The conference exhibited real world impacts of AI, focusing on transformative roles of AI in bioinformatics and public health. The papers included in this thematic issue are from participants of this conference and demonstrate the scientific efforts of AR-BIC to realize the impact of AI in the real world.

Compounds interacting with the endocrine system can lead to numerous adverse outcomes. The androgen receptor is one important component in the endocrine system and mediates male sex hormones. Chemicals binding to androgen receptor raise concerns on reproductive health. Therefore, a high-quality data source of androgenic activity data is urgently needed to fully utilize artificial intelligence techniques such as machine learning and deep learning to develop reliable models for predicting androgenic activity of compounds. [1] introduced the Molecules with Androgenic Activity Resource

(MAAR) that was developed to facilitate utilization of androgenic activity data for assessing chemical risk. MAAR is an open-access data source designed to provide comprehensive data for developing machine learning and deep learning models and streamlining and improving the evaluation of androgenic activity of compounds. This tool has a user-friendly interface, providing for efficient navigation and download of the androgenic activity data. The open-access nature of MAAR facilitates the use of androgenic activity data in the development of machine learning and deep learning models for assessing chemical risk, supporting regulatory reviewers and scientists in evaluating the endocrine-disrupting potential of compounds.

Pharmacogenomics plays a vital role in precision medicine. However, the available genomic data of drugs are distributed in diverse data sources, making access to the pharmacogenomics data time-consuming and thus hindering the implementation of precision medicine. Therefore, tools are needed to enable rapid and automatic identification of sources that contain high-quality pharmacogenomics data. [2] explored the ability of large language models in this role. They tested the feasibility of Llama3.1–70B in extracting pharmacogenomics data from the FDA Table of Pharmacogenomic Biomarkers in Drug Labeling (https://www.fda.gov/drugs/science-and-research-drugs/table-pharmacogenomic-biomarkers-drug-labeling) as an alternative approach to the most used labor-intensive methods. The results showed a high accuracy in identifying genomic biomarkers of drugs from single labeling texts or mixed texts, demonstrating the effectiveness Llama3.1–70B in analyzing pharmacogenomics data. This study showcases the applicability of large language models to extract pharmacogenomics data from unstructured scientific and regulatory documents, paving the way for promoting precision medicine.

Screening tests for disease is important for improving diagnosis reliability. Performance of disease screening tests are typically measured using metrics such as sensitivity, specificity, and positive predictive value, quantifying the goodness of tests in differentiating between those with and without a disease. It is well known that these performance metrics, especially positive predictive value, are not reliable for traditional screening tests when the prevalence is very low. Machine learning algorithms are gaining popularity in developing predictive models to serve as *in silico* screening tests for disease. However, the screening and diagnostic performance of *in silico* screening tests, particularly for low prevalence cohorts, has not been fully investigated. [3] used The Strong Heart Study (https://strongheartstudy.org/), a study of cardiovascular disease and its risk factors among American Indians, as a case study to evaluate screening test diagnostics of *in silico* models, built with machine learning algorithms logistic regression, artificial neural networks, and random forest, at varying prevalence. Their results revealed that although sensitivity was not greatly affected in these *in silico* screening tests, specificity and positive predictive values dramatically declined when the prevalence decreased. This study demonstrates that machine learning models as disease screening tests have the same limitations as traditional screening tests when the disease prevalence is low in the testing cohort, calling for further studies to explore reliable *in silico* models for disease screening of low prevalence cohorts.

Natural language processing is an artificial intelligence branch and plays an important role in pharmacovigilance studies. Traditional topic modeling, such as Latent Dirichlet Allocation (LDA), has been widely used in text mining. However, LDA has limitations in capturing the semantic relationships in textual data, which is crucial in natural language processing. Bidirectional encoder representations from transformers (BERT) model-based topic modeling, BERTopic, can capture the contextual relationships. [4] integrated artificial intelligence modules to LDA and BERTopic and compared the two methods in evaluating prescription opioid-related cardiovascular risks in women by analyzing PubMed abstracts. Their results showed that that artificial intelligence algorithms can improve the performance of both LDA and BERTopic in identifying adverse events associated with prescription opioid drugs. Their comparison indicated while LDA remains useful for analyzing large-scale text at low computational cost, BERTopic can enhance interpretability and improve semantic coherence for extracting information in textual data.

Opioids are powerful pain-relieving drugs that are widely used in clinical practice. However, opioid addiction is a serious concern and can lead to opioid use disorder. Opioid drugs bind to opioid receptors, including the μ opioid receptor (MOR), attaining analgesic effects. Therefore, to develop pain treatment drugs that binding opioid receptors but are less addictive is one of the approaches to combat the opioid crisis. With the advancement of artificial intelligence and availability of experimental data, machine learning and deep learning have gained interest in new drug development. [5] developed models for predicting MOR binding activity of compounds using various machine learning and deep learning algorithms for assisting the development of less addictive drugs that target MOR. Their models have been assessed using both internal and external validations and have demonstrated robust predictive performance. The results suggest that the developed models could be used to predict MOR binders, potentially assisting in the development of less addictive drugs. This study demonstrates that machine learning and deep learning models can be used to guide the design of less addictive analgesics and ultimately lead to enhanced patient health.

Unstructured data such as textual documents in scientific publications, social media platforms, and clinical reports are often used for drug safety surveillance. One of the tasks in pharmacovigilance studies is to identify adverse events associated with drugs. Usually, different names can be used

for the same drug in textual documents, making it challenging to determining drugs associated with the identified adverse events in drug safety surveillance. Therefore, a comprehensive, non-redundant, and accurate list of drug names is crucial for identification and analysis of adverse events associated with drugs. RxNorm stands out from many sources of drug names as the most popular source used in pharmacovigilance studies. However, the effectiveness of drug names in RxNorm for drug safety surveillance needs to be thoroughly assessed. [6] examined the drug names in RxNorm and developed a refined set of drug names for enhancing unstructured data analysis in drug safety surveillance. They removed duplicates, false drug names, and drug names likely causing inaccurate drug counts in drug safety surveillance from RxNorm, yielding a refined set of drug names. The efficiency and accuracy of the refined drug names were evaluated and compared with the names of original RxNorm using PubMed abstracts. The results demonstrated an increased computational efficiency and decreased false drug names identified for the refined set. Their findings indicate that the refined drug names can improve identification and counting of drugs in unstructured textual data, thereby improving pharmacovigilance.

Mass cytometry is widely used for high-throughput characterization of cellular heterogeneity. Analyzing experimental data from mass cytometry often employ manual gating or clustering technique. [7] proposed quantized optimal transport (QOT), a novel framework derived from optimal transport theory, to analyze mass cytometry data. They used QOT to measure distances between samples based on cellular protein expression profiles by treating the cell-by-protein matrix as a high-dimensional distribution. Their method enables a direct distribution comparison to capture small variations in mass cytometry data and does not need predefined gating strategies. This method was evaluated using two time-series mass cytometry datasets of Coronavirus Disease 2019 (COVID-19) samples. Their leave-one-out analysis identified CD3 and CD45 as immunologically unstable proteins which had the most variation over time during the vaccine response. Their hierarchical clustering based on pairwise Wasserstein distances between samples resulted in the discovery of optimal combinations of immunological markers for grouping samples of different time points from the same patients. This study demonstrates that QOT is a reliable and flexible method for analysis of mass cytometry data of patients to capture immune response heterogeneity, improving the identification of unstable immunological markers and improving patient health.

Antiretroviral therapy (ART) is effective for mitigating human immunodeficiency virus transmission from mother to child. However, there are concerns on potential long-term impacts of ART on offspring health. The gut microbiome contains a huge number of microorganisms, including many types of bacteria. The population of gut bacterial and the produced short-chain fatty acids in the offspring of an ART treated mother can be used to evaluate the health effects of the offspring. [8] investigated the potential long-term effects of ART on offspring health through analyzing gut microbiota populations and short-chain fatty acids concentrations in aged rat offspring with ART exposure *in utero* and during lactation. In this study, pregnant rats received a combination of antiretroviral drugs at two different doses during gestation and lactation, and their offspring's fecal bacterial abundance and short-chain fatty acid concentrations at 12 months of age were analyzed. They found that *Firmicutes* in males were decreased, while *Actinobacteria* in both males and females were increased. However, the metabolic products (short chain fatty acids) and immune factors (IgA) remained stable. This study suggests a need for further understanding of the long-term effects of ART on offspring and points to future pathways for monitoring offspring health.

Collectively, these articles highlight the advances accomplished to demonstrate the impact of AI in the real world, both to revolutionize biomedical research and enhance public health.

This editorial reflects the views of the authors and does not necessarily reflect those of the U.S. Food and Drug Administration.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

1. Dong F, Hardy B, Liu J, Mohoric T, Guo W, Exner T, et al. Development of a comprehensive open access "molecules with androgenic activity resource (MAAR)" to facilitate risk assessment of chemicals. *Exp Biol Med* (2024) **249**:10279. doi:10.3389/ebm.2024.10279

2. Li D, Wu L, Lin Y-C, Huang H-Y, Cotton E, Liu Q, et al. Enhancing pharmacogenomic data accessibility and drug safety with large language models: a case study with Llama3.1. *Exp Biol Med* (2024) **249**:10393. doi:10.3389/ebm.2024.10393

3. Rogers P, McCall T, Zhang Y, Reese J, Wang D, Tong W, et al. AI-powered topic modeling: comparing LDA and BERTopic in analyzing opioid-related cardiovascular risks in women. *Exp Biol Med* (2025) **250**:10341. doi:10.3389/ebm.2024.10341

4. Ma L, Chen R, Ge W, Rogers P, Lyn-Cook B, Hong H, et al. AI-powered topic modeling: comparing LDA and BERTopic in analyzing opioid-related cardiovascular risks in women. *Exp Biol Med* (2025) **250**:10389. doi:10.3389/ebm.2025.10389

5. Liu J, Li J, Li Z, Dong F, Guo W, Ge W, et al. Developing predictive models for μ opioid receptor binding using machine learning and deep learning techniques. *Exp Biol Med* (2025) **250**:10359. doi:10.3389/ebm.2025.10359

6. Guo W, Dong F, Liu J, Aslam A, Patterson TA, Hong H, et al. A refined set of RxNorm drug names for enhancing unstructured data analysis in drug safety surveillance. *Exp Biol Med* (2025) **250**:10374. doi:10.3389/ebm.2025.10374

7. Wang Z, Chen J, Ionita M, Zhan Q, Zhou Z, Shen L, et al. Optimal transport reveals immune perturbation and fingerprints over time in COVID-19 vaccination. *Exp Biol Med* (2025) **250**:10445. doi:10.3389/ebm.2025.10445

8. Muthumula CMR, Yanamadala Y, Gokulan K, Karn K, Cunny H, Sutherland V, et al. Effect of in utero and lactational exposure to antiretroviral therapy on the gut microbial composition and metabolic function in aged rat offspring. *Exp Biol Med* (2024) **250**:10468. doi:10.3389/ebm.2025.10468

EBM Experimental Biology and Medicine

TYPE Original Research
PUBLISHED 19 September 2024
DOI 10.3389/ebm.2024.10279

OPEN ACCESS

*CORRESPONDENCE
Huixiao Hong,
huixiao.hong@fda.hhs.gov
Barry Hardy,
barry.hardy@edelweissconnect.com

RECEIVED 07 June 2024
ACCEPTED 27 August 2024
PUBLISHED 19 September 2024

CITATION
Dong F, Hardy B, Liu J, Mohoric T, Guo W, Exner T, Tong W, Dohler J, Bachler D and Hong H (2024) Development of a comprehensive open access "molecules with androgenic activity resource (MAAR)" to facilitate risk assessment of chemicals. *Exp. Biol. Med.* 249:10279. doi: 10.3389/ebm.2024.10279

Development of a comprehensive open access "molecules with androgenic activity resource (MAAR)" to facilitate risk assessment of chemicals

Fan Dong[1], Barry Hardy[2]*, Jie Liu[1], Tomaz Mohoric[2], Wenjing Guo[1], Thomas Exner[2], Weida Tong[1], Joh Dohler[2], Daniel Bachler[2] and Huixiao Hong[1]*

[1]National Center for Toxicological Research, US Food and Drug Administration, Jefferson, AR, United States, [2]Edelweiss Connect Inc., Durham, NC, United States

Abstract

The increasing prevalence of endocrine-disrupting chemicals (EDCs) and their potential adverse effects on human health underscore the necessity for robust tools to assess and manage associated risks. The androgen receptor (AR) is a critical component of the endocrine system, playing a pivotal role in mediating the biological effects of androgens, which are male sex hormones. Exposure to androgen-disrupting chemicals during critical periods of development, such as fetal development or puberty, may result in adverse effects on reproductive health, including altered sexual differentiation, impaired fertility, and an increased risk of reproductive disorders. Therefore, androgenic activity data is critical for chemical risk assessment. A large amount of androgenic data has been generated using various experimental protocols. Moreover, the data are reported in different formats and in diverse sources. To facilitate utilization of androgenic activity data in chemical risk assessment, the Molecules with Androgenic Activity Resource (MAAR) was developed. MAAR is the first open-access platform designed to streamline and enhance the risk assessment of chemicals with androgenic activity. MAAR's development involved the integration of diverse data sources, including data from public databases and mining literature, to establish a reliable and versatile repository. The platform employs a user-friendly interface, enabling efficient navigation and extraction of pertinent information. MAAR is poised to advance chemical risk assessment by offering unprecedented access to information crucial for evaluating the androgenic potential of a wide array of chemicals.

The open-access nature of MAAR promotes transparency and collaboration, fostering a collective effort to address the challenges posed by androgenic EDCs.

KEYWORDS

androgen receptor, risk assessment, chemicals, database, open access

Impact statement

The prevalence of endocrine-disrupting chemicals (EDCs) and their potential health impacts necessitate robust tools for risk assessment. The androgen receptor is crucial in mediating the effects of male sex hormones, with disruption during critical developmental periods leading to reproductive health issues. To address this, the Molecules with Androgenic Activity Resource (MAAR) was developed. MAAR integrates diverse data sources to create an open-access platform facilitating chemical risk assessment. By offering easy navigation and extraction of androgenic activity data, MAAR enhances transparency and collaboration in addressing the challenges posed by androgenic EDCs.

Introduction

Endocrine-active chemicals are exogenous compounds that affect the endocrine system of humans and other vertebrates. Endocrine activity of chemicals has the potential to cause numerous adverse outcomes, including disrupting physiological function of endogenous hormones and altering homeostasis [1, 2]. Evidence that certain man-made chemicals can disrupt the endocrine system by mimicking endogenous hormones sparked intense international scientific discussion and debate starting some 24 years ago [3]. These discussions culminated in issuance of legislation that reauthorized the Safe Drinking Water Act[1] and authorization of the 1996 Food Quality Protection Act mandating that the US Environmental Protection Agency (EPA) develop a program for screening and testing chemicals with endocrine disrupting potential[2]. In 2015, the US Food and Drug Administration (FDA) published guidance to provide recommendations to sponsors of investigational new drug applications, new drug applications, and biologics license applications regulated by the FDA's Center for Drug Evaluation and Research (CDER) regarding nonclinical studies intended to identify the potential for a drug to cause endocrine-related toxicity[3]. FDA's National Center for Toxicological Research (NCTR) developed a program to meet the need for information systems focused on aggregating knowledge of chemicals with experimental results relevant to endocrine activity. These efforts resulted in the development of the endocrine disruptors knowledge base (EDKB) [4] and estrogenic activity database (EADB) [5], which have been used to help identify endocrine active chemicals, develop predictive toxicology models, and prioritize chemicals for laborious and expensive testing [6–22]. However, as of today, androgenic activity data have not been comprehensively curated into a database.

Androgen receptors (ARs) are ligand-dependent transcription factors that belong to the nuclear receptor superfamily [23]. ARs are widely expressed in various tissues within the body [24]. They are the targets for drugs to treat hormone-related diseases including cancers of prostate, breast, ovary, pancreas, etc. [25]. On the other hand, chemicals can interfere with the endocrine system by interacting with ARs, which result in adverse effects [26]. Therefore, estimation of the androgenic activity of drugs and other chemicals is critical for the evaluation of drug safety and assessment of chemical risk.

Over the past decades, large numbers of chemicals have been assayed for androgenic activity by government agencies, industry, and academic research groups, with the results of these studies reported in the public domain. However, the data are distributed across different and diverse sources, obtained in multiple diverse assays, and stored in different formats, limiting the use of the data in research and regulation. Therefore, a comprehensive and reliable resource to provide open access to the data and enable modeling and prediction of androgenic activity for untested chemicals could facilitate advancement in developing strategies to mitigate the AR-driven toxicity and risk. To enable and optimize the use of the data generated by these studies, we have developed and are maintaining a comprehensive open access resource called Molecules with Androgenic Activity Resource (MAAR) to provide the scientific and regulatory communities with an up-to-date androgenic activity database for evaluating potential endocrine activity of chemicals.

Materials and methods

Data collection

Androgenic activity data were collected from multiple sources which encompass published literature and public

1 http://www.epa.gov/safewater/sdwa/index.html

2 http://www.epa.gov/scipoly/oscpendo/

3 https://www.fda.gov/media/86996/download

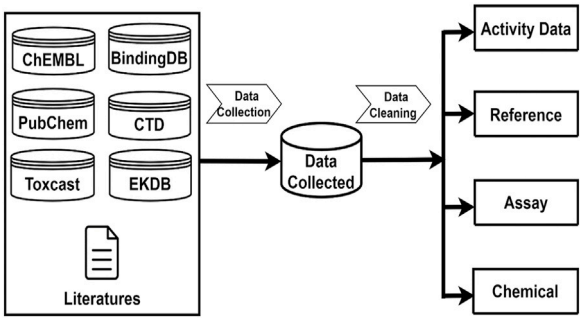


FIGURE 1
Androgenic Activity Data Collection and Curation. Data collection sources including ChemBL, PubChem, BindingDB, EDKB, CTD, Toxcast. After data pre-processing steps, Androgenic Activity Data are curated into four categories: Activity Data information, Reference information, Assay information, Chemical information.

databases including PubChem [27], ChEMBL [28], BindingDB [29], EDKB [4], ToxCast [30], and the Comparative Toxicogenomics Database (CTD) [31]. Java programs were developed to automatically retrieve androgenic activity data points and associated data such as chemical structures, assays, species, and references from these public databases. In addition, androgenic activity data in the literature were manually searched and extracted. The collected androgenic activity data include both quantitative measurements for active compounds and qualitative descriptions for inactive chemicals. Figure 1 gives the sources from which the androgenic activity and related data were collected and the four types of data that were included in this database.

Data curation

After data were collected from individual sources, data were pre-processed and integrated before they were implemented into the database. Given potential duplications in the data collected from different sources, an automated pre-process program was devised to check and remove duplicated data records by comparing chemicals, activity data, assays, and references. This program identified and removed duplicates by comparing CID, ChEMBL ID, PubMed ID, endpoint values, and assay descriptions across data sources. Geometric and optical isomers are considered duplicates only if they have the same CID or ChEMBL ID. This program also ensured data uniformity by transforming all collected activity data into standardized units. For different activity values of a compound from different sources where inconsistencies were found, a manual review was conducted to determine the most reliable value by examining the assay details. Following a cleaning procedure that removed duplicates to keep unique androgenic activity data, the pre-processed data were combined to make the final data that were included in the database. This program was developed in Java. It processes text file containing all activity information, specifies columns used for comparison, and identifies both duplicate and unique activity records to ensure that non-redundant data is included in the final dataset.

TABLE 1 Reference data table.

Data field	Description
Reference ID	Internal ID for reference
PMID	PubMed ID for reference
Journal_Name	Reference journal name
Year	Publication year
Volume	Volume number
Issue	Issue number
First_Page	First page number
Author	Author names
Title	Publication title

TABLE 2 Assay data table.

Data field	Description
Assay ID	Internal ID for assay
Description	Description of assay
Assay_Name	Assay name
Assay_Group	Assay group, e.g., HTS, Reporter gene
Assay_Format	Assay format, e.g., cell-based, protein-based
Assay_Type	Assay type, e.g., Agonist, Antagonist
Species	Species assay based on, e.g., *Homo sapiens*
AID	Bioassay AID in PubChem
ChEMBL Assay ID	Assay ID in ChEMBL.

Data model

The data implemented in the database were organized into four categories: androgen activity data, references, assays, and chemical information. Properties for each of the four categories are summarized in Tables 1–4. The four tables are interconnected through Chemical ID, Assay ID and Reference ID as depicted in the database schema in Figure 2.

TABLE 3 Androgen activity data table.

Data field	Description
Activity ID	Internal ID for activity data
Chemical ID	Internal ID for chemicals
Assay ID	Internal ID for assays
Reference ID	Internal ID for references
Endpoint	Activity endpoint, e.g., IC50, AC50, LogRBA
Relation	Relation to describe activity value, e.g., >, =, <
Value	Activity value from endpoint measurement
Units	Activity data unit, e.g., nM, %
Download	Database where data downloaded, e.g., PubChem, ChEMBL
Curation/Data source	Date source, e.g., literatures, US Patent, Tox21, Abbott Labs

TABLE 4 Chemical data table.

Data field	Description
Chemical ID	Internal chemical ID of chemical
IUPAC_NAME	IUPAC name of chemical
Name	Chemical name used in the system
Synonyms	Chemical synonyms (a string separated by "|")
SMILES	SMILES string of chemical
InChIKey	A fixed-length format directly derived from InChI
InChI	International Chemical Identifier
Molecular_weight	Molecular weight
Molecular_formula	Molecular formula
CAS	Chemical CAS registry number
CID	Compound ID in PubChem
CHEMBL_ID	Compound ID in ChEMBL
BindingDB_ID	Compound ID in BindingDB

Database design

The curated tables were put into a cloud-based database based on EdelweissData that was developed by Edelweiss Connect, GmbH to tackle data management issues in life sciences. Some of the advantages offered by the EdelweissData solution are:

- Each published dataset is assigned a unique URL and is easily accessible through a web browser.
- Published datasets are automatically versioned.

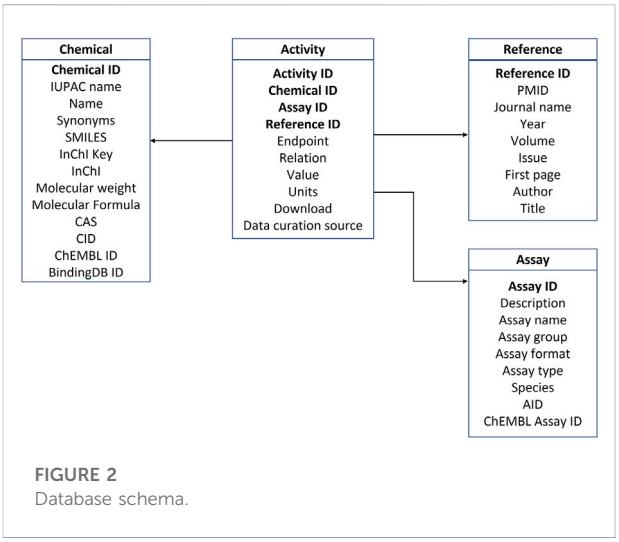


FIGURE 2
Database schema.

- Published datasets are static, i.e., they cannot be changed unless a newer version is published.
- Published datasets are immediately available through a web service and can be consumed by numerous data analysis tools (Python, R, Excel, KNIME, etc.) via REST API (see also section Data model).
- Flexibility - there is no predefined schema for published datasets. Instead, the schema is inferred from the data during publishing. This allows for a quick and easy consumption of datasets with various structures.

Database implementation - EdelweissData

The MAAR Database is built as a simple web application with a back-end supported by EdelweissData and a front-end that lets the user easily explore the database. The most common use cases (such as search by compound or chemical similarity) are well covered by the web application as is, while more advanced queries or analyses could be made through the API (Application Programming Interface) enabled by EdelweissData.

Results

Data collected and curated

In total, 125,519 androgenic activity data points for 13,648 chemicals were collected and curated from multiple sources and included in the MAAR database. These data were obtained from 923 assays. Table 5 lists the statistics of the data collected.

The androgenic activity data are presented in two types. The first type is quantitative value and 48,273 data points are in this type. A quantitative activity value indicates the androgenic

TABLE 5 Statistics of the data collected in the androgenic activity database.

Chemicals			13,648
Activity data	Quantitative data		48,273
	Qualitative data	Active	723
		Inactive	71,630
		Not determined	4,893
	Total		125,519
Assays	Binding		379
	Reporter gene		358
	Cell proliferation		86
	In vivo		60
	HTS		24
	Other		16
	Total		923
Species			6

activity is numerically determined. Another type of data is qualitative androgenic activity that is described using qualitative terms: active or positive indicates a chemical was tested using an assay and activity was observed but could not be numerically determined; inactive or negative means a chemical did not show androgenic activity in an assay; inconclusion or not determined or unspecified implies activity of a chemical in an assay was not able to be determined. There are 77,246 data that are qualitative. The data were generated using 923 assays, including 379 binding assays, 358 reporter gene assays, 86 cell proliferation assays, 60 *in vivo* assays, 24 high-throughput screening assays (HTS), and 16 other assays that could not be clearly put into any type of assays. Species used in the activity testing are also included in the database, including *Bos taurus*, Chimpanzee, Chinese hamster, *Homo sapiens*, *Mus musculus*, and *Rattus norvegicus*. Information on species for some assays could not be determined in the sources, and thus they are missed for the data generated using such assays.

A chemical could be tested in many laboratories using multiple assays. All androgenic activity data for the same chemicals were collected and presented in this database. The distribution of androgenic activity data for the same chemicals is given in Figure 3.

The same chemical is often tested by a variety of assays and has multiple data records. Of the 13,648 chemicals in the database, 2,504 have only one androgenic activity data and the remaining 11,144 have more than one data. Many chemicals have more than 10 androgenic activity data reported and are included in this database. For example, 3,481 chemicals have 11 data records, and 54 chemicals even have more than 40 data records. The androgenic activity data

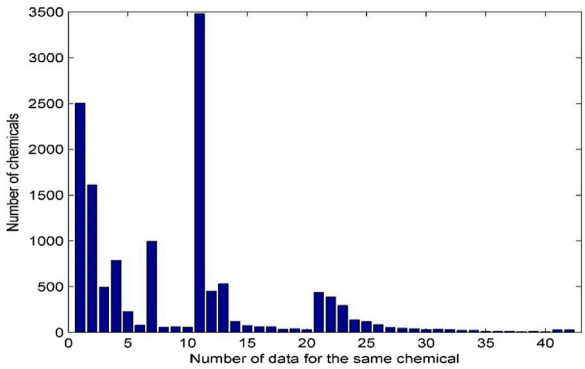


FIGURE 3
Distribution of androgenic activity data for the same chemical. Each bar represents the number of chemicals. X-axis indicates the number of data records for the same chemicals. The chemicals with 41–50 data records were grouped into the bar with x-axis value 41 and the chemicals with more than 50 data were grouped in the last bar with x-axis value 42.

obtained from the same type of assays in different laboratories could be inconsistent. For example, 127 chemicals each have four androgenic activity data generated using binding assays. As shown in Figure 4, 79 chemicals are active for all four data (100% active) and 11 chemicals consistently show inactive (0% active), while the other (37 chemicals) have inconsistent androgenic activity data: one active and three inactive for 15 chemicals, three active and one inactive for six chemicals, and two active and two inactive for 16 chemicals. This database presents all androgenic activity data reported in different sources. Assessing data quality and selecting data for specific applications such as QSAR (quantitative structure-activity relationship) modelling are critically important. Users should make decisions on how to use the data tailored to their applications.

Web resource

The MAAR database is made available through a web portal as an open science resource based on open data provided according to a Creative Commons license. We have established the resource as part of the OpenTox open knowledge infrastructure located at[4]. The main initial functionality supported allows the user to search for compounds or chemically similar compounds in the database (Figure 5). The portal also supports the location of community-generated notebooks providing additional analysis of the data, starting with illustrative examples we have provided (see sections Method and Application Programming Interface).

4 https://opentox.net/MAAR/

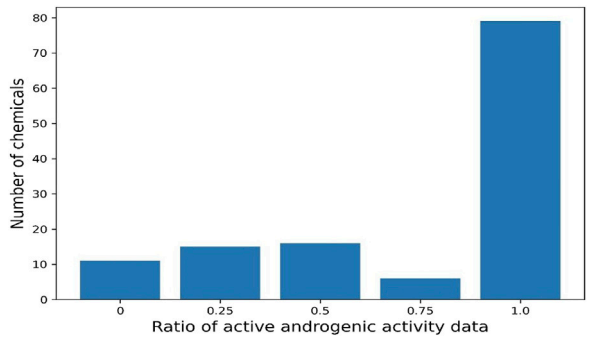


FIGURE 4
Ratio of active data among 4 binding assay data for 127 chemicals. Each bar represents the number of chemicals. X-axis indicates the ratio of active data.

Application programming interface (API)

The MAAR database comes with a versatile API that simplifies the consumption of the data into other applications. Common data analysis tools that support Representational State Transfer (REST) APIs can obtain data in the database through a simple web request. In this way data can be easily transferred into a Python or R script/notebook, KNIME, Microsoft Excel, etc. To make it even easier for users, an example of an API call in Python and curl[5] is provided in the web application and could be copy-pasted to the user's script/notebook. API documentation is available from the EdelweissData main website[6].

For the purpose of demonstrating programmatic data retrieval from the database, we show an example of how a particular dataset could be accessed with a web request. Each assay dataset in the database has its own unique ID and when the URL pointing to that dataset is called the database returns the dataset in the JSON format. For a dataset inside the database the URL for dataset with ID "21b033c5-d048-41f5-b8a1-d5d8492f7048" would be the following: [7]. And the response from the database is shown in Figure 6.

Notebooks

The REST API service mentioned above is very well suited for different interactive notebooks that are nowadays a common tool for data analysis and visualization. There are many different notebooks available today that differ in the language, interactivity, etc. To build an interactive notebook for visualization of the MAAR data we decided to use Observable HQ notebooks[8] as they offer in our opinion a very good user experience even for technically less skilled users. The programming language in Observable notebooks is JavaScript, which is typically not the language of first choice for data analysis, however, it is very well suited for interactive visualizations that work as a web page.

The Observable notebook[9] for the MAAR database is available through a URL and can be easily shared with anyone. The notebook addresses a simple use case where a user wants to search the database for a particular compound. The notebook returns a list of chemically most similar compounds (based on the Tanimoto chemical similarity – see Figure 7, left) together with their activities in the assays. In the next step, users can narrow down the set of activities by filtering the assays based on format, group, type, species, or endpoint. Finally, the subset of compounds (on x axis) and their activities (colors) in various assays (y axis) is displayed as a heatmap (Figure 7, right).

Discussion

Androgens are hormones that play a key role in the development and maintenance of male characteristics. Understanding the androgenic activity of chemicals is important for assessing chemical risk through endocrine disruption. Therefore, androgenic activity data are important for comprehensive chemical risk assessments, providing insight into the potential endocrine-disrupting effects of substances and helping to establish guidelines and regulations to protect human health and the environment.

Vast amounts of androgenic activity data have been generated and reported in the public domain for many chemicals. However, accessing and using androgenic activity data in the public domain may pose several challenges. First, androgenic activity data are contained in different and diverse sources in the public domain. The lack of comprehensive datasets can hinder applications in chemical risk assessment. Second, the importance of data quality and reliability in scientific research cannot be overstated [32, 33]. Sound scientific conclusions rest on the foundation of accurate and trustworthy data. The reliability and accuracy of available androgenic activity data vary. Incomplete or poorly curated datasets can compromise the validity of research findings. Third, androgenic activity data are sourced from various studies, experiments, or databases, leading to heterogeneity in data formats and measurement techniques. Lack of

5 https://en.wikipedia.org/wiki/CURL

6 https://edelweissdata.com/docs/about

7 https://api.aadb.cloud.edelweissconnect.com/datasets/21b033c5-d048-41f5-b8a1-d5d8492f7048/versions/latest

8 https://observablehq.com/

9 https://observablehq.com/@saferworldbydesign/aadb-notebook

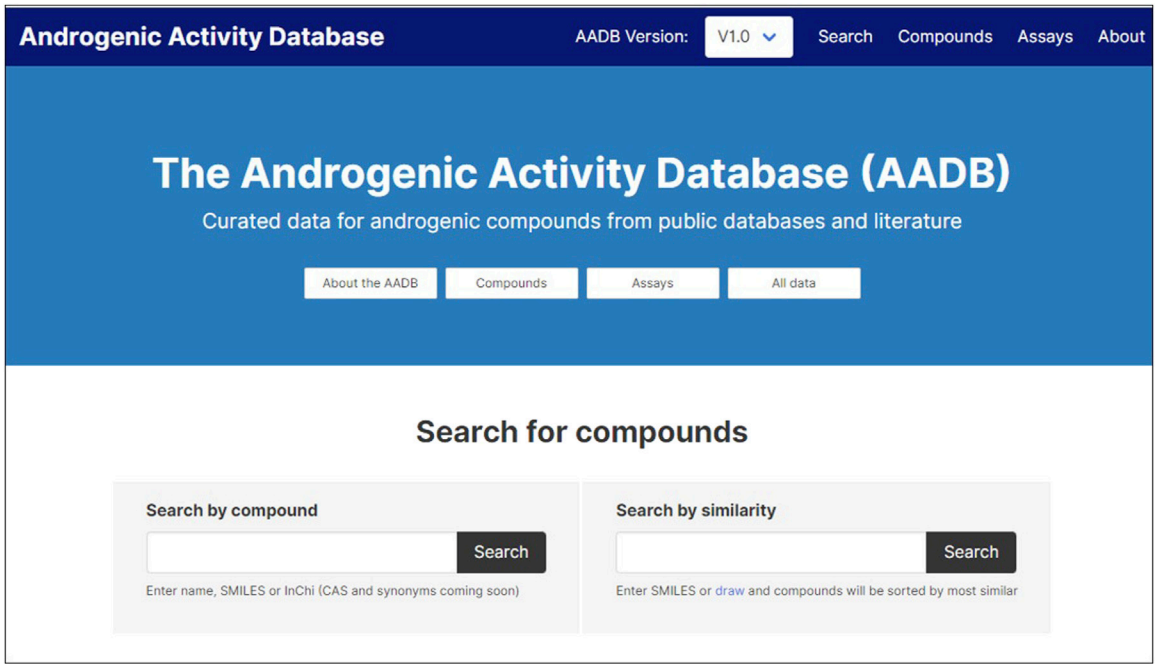


FIGURE 5
User interface of the web resource.

FIGURE 6
JSON response from the database.

standardized protocols for androgenic activity assessment can make it challenging to compare data from different sources. Finally, without sufficient metadata or contextual information, it may be challenging to interpret and utilize androgenic activity data accurately. Inter-laboratory and species-specific variations in androgenic responses can complicate the interpretation of androgenic activity data. Therefore, to facilitate utilization of the available androgenic data in chemical risk assessment, we aim to develop an open resource of androgenic activity data of molecules so that the huge amount of androgenic activity data generated in the scientific community could be used to accelerate and improve chemical risk assessment.

In this article we report the development of an open science data resource for androgenic activity data. We followed

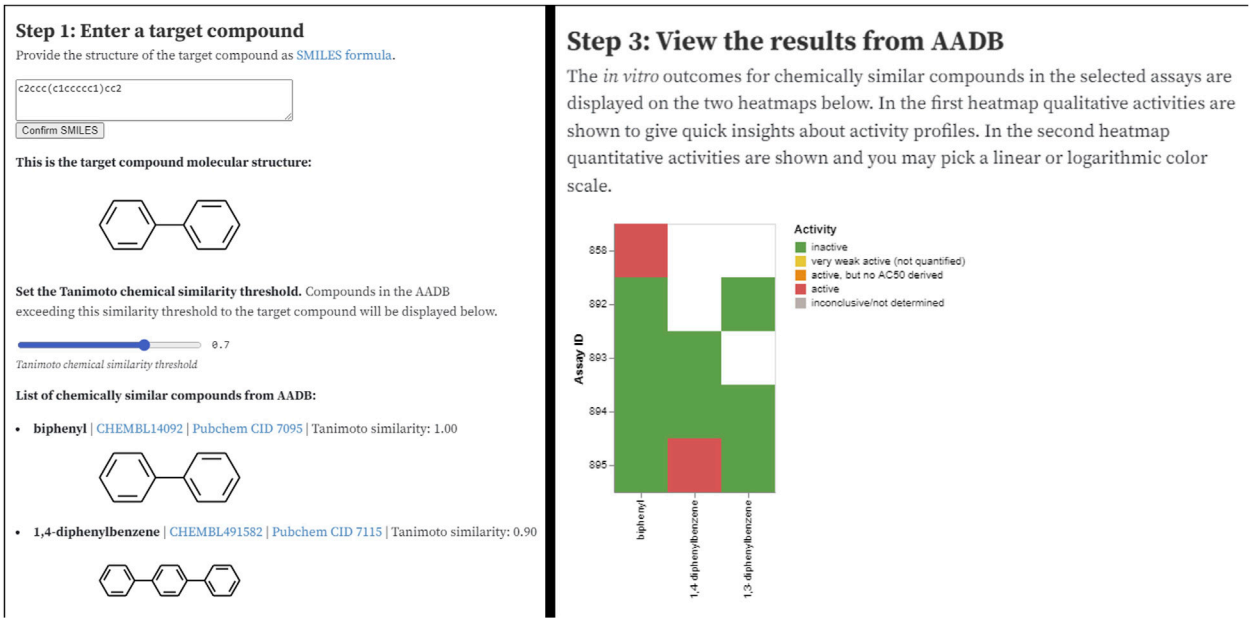


FIGURE 7
Two sections of the AADB Observable notebook: user interface for entering the input parameters (left) and visualization of filtered results (right).

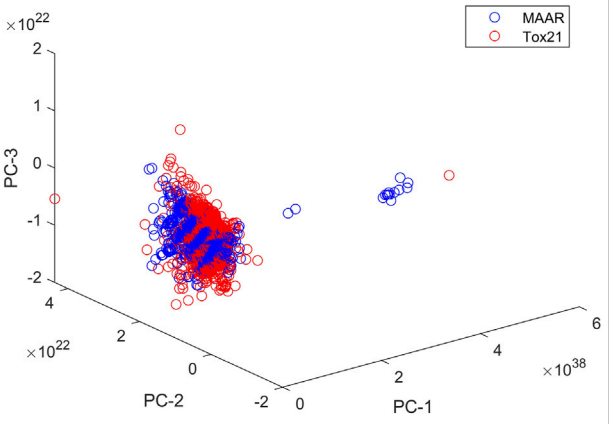


FIGURE 8
Chemical spaces of MAAR and Tox21. Compounds in MAAR and Tox21 are plotted as blue and red circles, respectively. The x-, y-, and z-axes give the first three principal components.

principles established in previous projects including OpenTox[10], OpenRiskNet[11] and EU-ToxRisk[12] [34–41]. The work includes the careful collection and curation of data entering the database and a data model which includes harmonized data to structure the data in a database. Resource functionalities aligned to the FAIR (findability, accessibility, interoperability, and reusability) principles in the preparation and sharing of open science data and supporting further initiatives and use of the project knowledge. We also paid attention to data integrity principles in the construction of the database and the provision of data through harmonized application programming interfaces, supporting the building of web applications making reliable use of the data. This approach should support analysis and modelling goals of the community in making use of the open knowledge resource created by this work.

The MAAR database is an extensive compilation of chemical compounds, systematically curated and annotated for their androgenic properties, providing researchers, regulators, and industry stakeholders with a comprehensive resource for in-depth investigations. To evaluate the structural coverage of chemicals in the MAAR, we computed chemical spaces for both the MAAR and Tox21 [42] datasets using Mold2 descriptors [43, 44]. Following the methodology outlined in our previous studies [19, 45], we performed principal component analysis to represent the chemical space for each dataset. Figure 8 illustrates the first three principal components of the compounds in MAAR and Tox21, demonstrating that the structural coverage of MAAR closely resembles that of Tox21. This comparison confirms that MAAR includes structurally diverse set of compounds, making it suitable for a wide range of applications. Development of the

10 https://opentox.net/
11 https://openrisknet.org/
12 www.eu-toxrisk.eu

MAAR database represents a significant stride towards a more comprehensive and accessible approach to assessing the androgenic activity of chemicals. By providing a centralized platform for data integration and analysis, the MAAR database is poised to enhance our understanding of androgenic endocrine disruption and contribute to the development of effective risk management strategies in the face of evolving chemical landscapes.

Conclusions

We have reported here on a useful curated database for androgenic activity provided as an open science resource to the community, and available to enable searches for relevant information on the presence or absence of evidence on androgenic activity of compounds. We have also provided a model and resource with interfaces supporting additional community members to build additional analysis and modelling applications that work with the database. We hope the resource will prove useful and encourage additional development of the resource including addition of new data and its analysis.

Author contributions

Conceptualization, HH and BH; methodology, FD, TM, JL, TE, WG, TE, BH, and HH; software, TM, TE, JD, and BH; data curation, FD, JL, WG, and HH; writing–original draft preparation, FD, JL, TM, BH, and HH; writing–review and editing, WT, BH, and HH; supervision, BH and HH. All authors have read and agreed to the published version of the manuscript.

Author disclaimer

This article reflects the views of the authors and does not necessarily reflect those of the U.S. Food and Drug Administration. Any mention of commercial products is for clarification only and is not intended as approval, endorsement, or recommendation.

Data availability

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/FANMISUA/AADB.git.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

Authors BH, TM, TE, JD, and DB were employed by Edelweiss Connect Inc.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Adebayo OA, Adesanoye OA, Abolaji OA, Kehinde AO, Adaramoye OA. First-line antituberculosis drugs disrupt endocrine balance and induce ovarian and uterine oxidative stress in rats. *J Basic Clin Physiol Pharmacol* (2018) **29**(2): 131–40. doi:10.1515/jbcpp-2017-0087

2. Danzo BJ. Environmental xenobiotics may disrupt normal endocrine function by interfering with the binding of physiological ligands to steroid receptors and binding proteins. *Environ Health Perspect* (1997) **105**(3):294–301. doi:10.1289/ehp.97105294

3. Kavlock RJ, Daston GP, DeRosa C, Fenner-Crisp P, Gray LE, Kaattari S, et al. Research needs for the risk assessment of health and environmental effects of endocrine disruptors: a report of the U.S. EPA-sponsored workshop. *Environ Health Perspect* (1996) **104**(Suppl. 4):715–40. doi:10.2307/3432708

4. Ding D, Xu L, Fang H, Hong H, Perkins R, Harris S, et al. The EDKB: an established knowledge base for endocrine disrupting chemicals. *BMC Bioinformatics* (2010) **11**(Suppl. 6):S5. doi:10.1186/1471-2105-11-s6-s5

5. Shen J, Xu L, Fang H, Richard AM, Bray JD, Judson RS, et al. EADB: an estrogenic activity database for assessing potential endocrine activity. *Toxicol Sci* (2013) **135**(2):277–91. doi:10.1093/toxsci/kft164

6. Hong H, Tong W, Fang H, Shi LM, Xie Q, Wu J, et al. Prediction of Estrogen Receptor Binding for 58,000 chemicals Using an Integrated system of a tree-based model with structural alerts. *Environ Health Perspect* (2002) **110**(1):29–36. doi:10.1289/ehp.0211029

7. Tong W, Perkins R, Fa Perkinsng H, Hong H, Xie Q, Branham SW, et al. Development of Quantitative Structure-Activity Relationships (QSARs) and their use for priority setting in the testing strategy of endocrine disruptors. *Regul Res Perspect* (2002) **1**(3):1–16.

8. Hong H, Fang H, Xie Q, Perkins R, Sheehan DM, Tong W. Comparative molecular field analysis (CoMFA) model using a large diverse set of natural, synthetic and environmental chemicals for binding to the androgen receptor. *SAR QSAR Environ Res* (2003) **14**(5-6):373–88. doi:10.1080/10629360310001623962

9. Shi LM, Tong W, Fang H, Xie Q, Hong H, Perkins R, et al. An integrated 4-Phase approach for setting endocrine disruption screening priorities - phase I and II predictions of estrogen receptor binding affinity. *SAR QSAR Environ Res* (2002) **13**(1):69–88. doi:10.1080/10629360290002235

10. Sakkiah S, Guo W, Pan B, Kusko R, Tong W, Hong H. Computational prediction models for assessing endocrine disrupting potential of chemicals. *J Environ Sci Health C* (2018) **36**(4):192–218. doi:10.1080/10590501.2018.1537132

11. Ng HW, Zhang W, Shu M, Luo H, Ge W, Perkins R, et al. Competitive molecular docking approach for predicting estrogen receptor subtype α agonists and antagonists. *BMC Bioinformatics* (2014) **15**(Suppl. 11):S4. doi:10.1186/1471-2105-15-s11-s4

12. Ng HW, Doughty SW, Luo H, Ye H, Ge W, Tong W, et al. Development and validation of decision forest model for estrogen receptor binding prediction of chemicals using large data sets. *Chem Res Toxicol* (2015) **28**(12):2343–51. doi:10.1021/acs.chemrestox.5b00358

13. Ng HW, Shu M, Luo H, Ye H, Ge W, Perkins R, et al. Estrogenic activity data extraction and *in silico* prediction show the endocrine disruption potential of bisphenol A replacement compounds. *Chem Res Toxicol* (2015) **28**(9):1784–95. doi:10.1021/acs.chemrestox.5b00243

14. Hong H, Rua D, Sakkiah S, Selvaraj C, Ge W, Tong W. Consensus modeling for prediction of estrogenic activity of ingredients commonly used in sunscreen products. *Int J Environ Res Public Health* (2016) **13**(10):958. doi:10.3390/ijerph13100958

15. Hong H, Harvey BG, Palmese GR, Stanzione JF, Ng HW, Sakkiah S, et al. Experimental data extraction and *in silico* prediction of the estrogenic activity of renewable replacements for bisphenol A. *Int J Environ Res Public Health* (2016) **13**(7):705. doi:10.3390/ijerph13070705

16. Ye H, Luo H, Ng HW, Meehan J, Ge W, Tong W, et al. Applying network analysis and Nebula (neighbor-edges based and unbiased leverage algorithm) to ToxCast data. *Environ Int* (2016) **89-90**:81–92. doi:10.1016/j.envint.2016.01.010

17. Sakkiah S, Kusko R, Tong W, Hong H. Applications of molecular dynamics simulations in computational toxicology. In: Hong H, editor. *Advances in computational toxicology: methodologies and applications in regulatory science*. Cham: Springer International Publishing (2019). p. 181–212.

18. Sakkiah S, Selvaraj C, Guo W, Liu J, Ge W, Patterson TA, et al. Elucidation of agonist and antagonist dynamic binding patterns in ER-α by integration of molecular docking, molecular dynamics simulations and quantum mechanical calculations. *Int J Mol Sci* (2021) **22**(17):9371. doi:10.3390/ijms22179371

19. Tan H, Wang X, Hong H, Benfenati E, Giesy JP, Gini GC, et al. Structures of endocrine-disrupting chemicals determine binding to and activation of the estrogen receptor α and androgen receptor. *Environ Sci Technol* (2020) **54**(18):11424–33. doi:10.1021/acs.est.0c02639

20. Banerjee A, De P, Kumar V, Kar S, Roy K. Quick and efficient quantitative predictions of androgen receptor binding affinity for screening Endocrine Disruptor Chemicals using 2D-QSAR and Chemical Read-Across. *Chemosphere* (2022) **309**(Pt 1):136579. doi:10.1016/j.chemosphere.2022.136579

21. Wilkes JG, Stoyanova-Slavova IB, Buzatu DA. Alignment-independent technique for 3D QSAR analysis. *J Comput Aided Mol Des* (2016) **30**(4):331–45. doi:10.1007/s10822-016-9909-0

22. Zhang L, Sedykh A, Tripathi A, Zhu H, Afantitis A, Mouchlis VD, et al. Identification of putative estrogen receptor-mediated endocrine disrupting chemicals using QSAR- and structure-based virtual screening approaches. *Toxicol Appl Pharmacol* (2013) **272**(1):67–76. doi:10.1016/j.taap.2013.04.032

23. Lu NZ, Wardell SE, Burnstein KL, Defranco D, Fuller PJ, Giguere V, et al. International Union of Pharmacology. LXV. The pharmacology and classification of the nuclear receptor superfamily: glucocorticoid, mineralocorticoid, progesterone, and androgen receptors. *Pharmacol Rev* (2006) **58**(4):782–97. doi:10.1124/pr.58.4.9

24. Sakkiah S, Ng HW, Tong W, Hong H. Structures of androgen receptor bound with ligands: advancing understanding of biological functions and drug discovery. *Expert Opin Ther Targets* (2016) **20**(10):1267–82. doi:10.1080/14728222.2016.1192131

25. Mooradian AD, Morley JE, Korenman SG. Biological actions of androgens. *Endocr Rev* (1987) **8**(1):1–28. doi:10.1210/edrv-8-1-1

26. Matsumoto T, Sakari M, Okada M, Yokoyama A, Takahashi S, Kouzmenko A, et al. The androgen receptor in health and disease. *Annu Rev Physiol* (2013) **75**:201–24. doi:10.1146/annurev-physiol-030212-183656

27. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res* (2019) **47**(D1):D1102–D1109. doi:10.1093/nar/gky1033

28. Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res* (2019) **47**(D1):D930–D940. doi:10.1093/nar/gky1075

29. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res* (2007) **35**:D198–201. doi:10.1093/nar/gkl999

30. Dix DJ, Houck KA, Martin MT, Richard AM, Setzer RW, Kavlock RJ. The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicol Sci* (2007) **95**(1):5–12. doi:10.1093/toxsci/kfl103

31. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, Wiegers J, Wiegers TC, et al. Comparative toxicogenomics database (CTD): update 2021. *Nucleic Acids Res* (2021) **49**(D1):D1138–43. doi:10.1093/nar/gkaa891

32. Hong H, Xu L, Liu J, Jones WD, Su Z, Ning B, et al. Technical reproducibility of genotyping SNP arrays used in genome-wide association studies. *PLoS One* (2012) **7**(9):e44483. doi:10.1371/journal.pone.0044483

33. Pan B, Kusko R, Xiao W, Zheng Y, Liu Z, Xiao C, et al. Correction to: similarities and differences between variants called with human reference genome HG19 or HG38. *BMC Bioinformatics* (2019) **20**(1):252. doi:10.1186/s12859-019-2776-7

34. Hardy B, Douglas N, Helma C, Rautenberg M, Jeliazkova N, Jeliazkov V, et al. Collaborative development of predictive toxicology applications. *J Cheminformatics* (2010) **2**(7):7. doi:10.1186/1758-2946-2-7

35. Hardy B, Apic G, Carthew P, Clark D, Cook D, Dix I, et al. A toxicology ontology roadmap. *ALTEX* (2012) **29**:129–37. doi:10.14573/altex.2012.2.129

36. Hardy B, Apic G, Carthew P, Clark D, Cook D, Dix I, et al. Toxicology ontology perspectives. *ALTEX* (2012) **29**:139–56. doi:10.14573/altex.2012.2.139

37. Kohonen P, Benfenati E, Bower D, Ceder R, Crump M, Cross K, et al. The ToxBank data warehouse: supporting the replacement of *in vivo* repeated dose systemic toxicity testing. *Mol Inform* (2013) **32**(Issue 1):47–63. doi:10.1002/minf.201200114

38. Fourches D, Muratov E, Tropsha A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J Chem Inf Model* (2010) **50**:1189–204. doi:10.1021/ci100176x

39. Fourches D, Muratov E, Tropsha A. Curation of chemogenomics data. *Nat Chem Biol* (2015) **11**:535. doi:10.1038/nchembio.1881

40. Fourches D, Muratov E, Tropsha A. Trust, but Verify II: A Practical Guide to Chemogenomics Data Curation. *J Chem Inf Model*. (2016) **56**(7):1243–1252. doi:10.1021/acs.jcim.6b00129

41. Tropsha A. Best practices for QSAR model development, validation, and exploitation. *Mol Inform* (2010) **29**(6-7):476–88. doi:10.1002/minf.201000061

42. Idakwo G, Thangapandian S, Luttrell J, Li Y, Wang N, Zhou Z, et al. Structure-activity relationship-based chemical classification of highly imbalanced Tox21 datasets. *J Cheminform* (2020) **12**(1):66. doi:10.1186/s13321-020-00468-x

43. Hong H, Xie Q, Ge W, Qian F, Fang H, Shi L, et al. Mold(2), molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. *J Chem Inf Model* (2008) **48**(7):1337–44. doi:10.1021/ci800038f

44. Hong H, Liu J, Ge W, Sakkiah S, Guo W, Yavas G, et al. Mold2 descriptors facilitate development of machine learning and deep learning models for predicting toxicity of chemicals. In: Hong H, editor. *Machine learning and deep learning in computational toxicology*. Cham: Springer International Publishing (2023). p. 297–321.

45. Liu J, Xu L, Guo W, Li Z, Khan MKH, Ge W, et al. Developing a SARS-CoV-2 main protease binding prediction random forest model for drug repurposing for COVID-19 treatment. *Exp Biol Med (Maywood)* (2023) **248**(21):1927–36. doi:10.1177/15353702231209413

EBM Experimental Biology and Medicine

TYPE Original Research
PUBLISHED 08 January 2025
DOI 10.3389/ebm.2024.10341

OPEN ACCESS

*CORRESPONDENCE
Paul Rogers,
paul.rogers@fda.hhs.gov

RECEIVED 09 August 2024
ACCEPTED 16 December 2024
PUBLISHED 08 January 2025

CITATION
Rogers P, McCall T, Zhang Y, Reese J, Wang D and Tong W (2025) Leveraging AI to improve disease screening among American Indians: insights from the Strong Heart Study.
Exp. Biol. Med. 249:10341.
doi: 10.3389/ebm.2024.10341

Leveraging AI to improve disease screening among American Indians: insights from the Strong Heart Study

Paul Rogers[1]*, Thomas McCall[2], Ying Zhang[3], Jessica Reese[3], Dong Wang[1] and Weida Tong[1]

[1]National Center for Toxicological Research, Division of Bioinformatics and Biostatistics, U.S. Food and Drug Administration, Jefferson, AR, United States, [2]Department of Data Science and Data Analytics, Arkansas State University, Jonesboro, AR, United States, [3]University of Oklahoma Health Sciences Center, Department of Biostatistics and Epidemiology, Oklahoma City, OK, United States

Abstract

Screening tests for disease have their performance measured through sensitivity and specificity, which inform how well the test can discriminate between those with and without the condition. Typically, high values for sensitivity and specificity are desired. These two measures of performance are unaffected by the outcome prevalence of the disease in the population. Research projects into the health of the American Indian frequently develop Machine learning algorithms as predictors of conditions in this population. In essence, these models serve as *in silico* screening tests for disease. A screening test's sensitivity and specificity values, typically determined during the development of the test, inform on the performance at the population level and are not affected by the prevalence of disease. A screening test's positive predictive value (PPV) is susceptible to the prevalence of the outcome. As the number of artificial intelligence and machine learning models flourish to predict disease outcomes, it is crucial to understand if the PPV values for these *in silico* methods suffer as traditional screening tests in a low prevalence outcome environment. The Strong Heart Study (SHS) is an epidemiological study of the American Indian and has been utilized in predictive models for health outcomes. We used data from the SHS focusing on the samples taken during Phases V and VI. Logistic Regression, Artificial Neural Network, and Random Forest were utilized as *in silico* screening tests within the SHS group. Their sensitivity, specificity, and PPV performance were assessed with health outcomes of varying prevalence within the SHS subjects. Although sensitivity and specificity remained high in these *in silico* screening tests, the PPVs' values declined as the outcome's prevalence became rare. Machine learning models used as *in silico* screening tests are subject to the same drawbacks as traditional screening tests when the outcome to be predicted is of low prevalence.

KEYWORDS

artificial intelligence, machine learning, screening test, American Indian, low prevalence

Impact statement

Artificial Intelligence (AI) and Machine Learning (ML) techniques are increasingly integrated into screening and diagnostic models to pinpoint individuals at risk of specific diseases or medical conditions. However, with the rise in popularity of AI and ML, the literature (and internet) is flooded with reports on computer-based prediction and screening tests, often focusing more on showcasing the technique rather than discussing their screening and diagnostic performance. In particular, there is a proliferation of algorithms created for minority groups, including the American Indian. A motivating factor in creating an *in silico* screening exam for American Indians is that this population, as a whole, experiences a greater burden of comorbidities, including diabetes mellitus, obesity, cancer, cardiovascular disease, and other chronic health conditions, than the rest of the U.S. population. This report evaluates these AI algorithms for the American Indian like a screening test in terms of performance in low prevalence situations.

Introduction

Artificial Intelligence (AI) and Machine Learning (ML) techniques are increasingly integrated into screening and diagnostic models to pinpoint individuals at risk for specific diseases or medical conditions [1]. However, with AI's and ML's rise in popularity, the literature (and the internet) is flooded with reports on computer-based prediction and screening tests, often focused more on showcasing techniques than discussing their screening and diagnostic performance. Advances in computer processing speed, increasing numbers of data scientists, low- to no-cost programming libraries, and availability of larger healthcare data sets have driven the proliferation of AI algorithms [2]. Kumar et al. have listed a sampling of prediction algorithms and data sets, including those for outcomes in Alzheimer's disease, cancer, diabetes, chronic heart disease, tuberculosis, stroke, hypertension, skin disease, and liver disease, among others [3]. Notwithstanding the proliferation of algorithms, AI is positioned to considerably enhance the accuracy and efficiency of screening tests. Specifically, ML algorithms can be trained on extensive data sets to discern patterns and make predictive analyses based on those patterns. Rapid expansion of AI technology, coupled with enhanced computing power in health screening, underscores the necessity for evaluating the algorithm's performance and quality of these algorithms [4].

The U.S. Government Accountability Office conducted a technology assessment noting that AI and ML offer advantages in analyzing underserved populations [5]. However, one challenge of utilizing AI in epidemiology pertains to the underrepresentation or absence of minority groups within these algorithms' training data sets [6]. Also, screening test performance may vary in minority populations due to their differences in disease prevalence from non-minority populations.

Many AI and ML methods for predicting disease in non-minority populations are recalibrated for minority groups. For example, an ML algorithm for mortality prediction based on chronic disease was recalibrated for the population of South Korea; this adjusted index showed a greater mortality prediction than the original algorithm [7]. Another effort adjusted this mortality prediction algorithm using hospital discharge abstracts from six countries [8].

In terms of minority status, American Indians are sometimes referred to as the "minority of the minority" or the "invisible minority," given their small population, cultural identity, languages, and histories that set them apart from other groups. Focusing on AI and ML can offer advantages in analyzing these underserved populations, who, like American Indians, bear a greater burden of certain health conditions [9–11]. This study focused on *in silico* AI and ML screening tests explicitly designed for the American Indian population.

The number of *in silico* diagnostic and screening tests has grown exponentially over the last decade, with many of these utilizing data sets based on American Indians. Our study serves as a reminder that *in silico* screening tests, even when classified as AI or ML algorithms, are still subject to the same limitations related to disease prevalence as those of their laboratory-based counterparts.

Popularization of Pima Indian data

Several research articles in the public domain report on AI and ML algorithms for diabetes classification in the Pima Indian population. A contributing factor is the availability of numerous Pima Indian data sets provided to the AI community through platforms like Kaggle, a popular resource for AI and ML algorithm developers [12]. However, these studies often overlooked the differences in disease prevalence among different populations and the potential consequences of applying algorithms trained specifically on one population to another.

Examples of ML algorithms for diabetes classification in Pima Indians sampled from the literature include Support Vector Machines, Radial Basis Function, Kernel Support Vector Machines, K-Nearest Neighbor, Artificial Neural Networks, Fuzzy Support Vector Machine, Naïve Bayes Classifier, J48 Decision Tree, and a Random Forest Classifier [13–15]. Some of the articles in this sample failed to recognize the high prevalence of diabetes among the Pima Indians and the impact of disease prevalence on screening test performance, and tended to focus solely on the methods used to perform the classifications [14].

The Strong Heart Study models

The Strong Heart Study (SHS) has played a significant role in identifying risk factors and patterns related to cardiovascular disease (CVD) in American Indian communities. It included 12 tribes located in Oklahoma, Arizona and the Dakotas. Statistical models developed using SHS data have informed interventions and public health policies targeting CVD. SHS data were also used in developing ML models and risk-based calculators addressing hypertension, diabetes, and coronary heart disease (CHD) [16–18].

AI and ML risks with American Indian data sets

However, potential risks are also associated with using ML in American Indian contexts. One notable concern involves the risk of ML algorithms perpetuating biases and stereotypes about American Indian communities. Specifically, algorithms trained on data sets that reinforce biases and stereotypes about American Indians could inadvertently foster further inequities against a population group frequently underrepresented in AI/ML training data sets for applications such as virtual screening tests. A lack of training data can also result in inaccuracies; a recent example involved an image recognition application identifying an American Indian in native dress as a bird [19]. To mitigate such risks, ML researchers and developers need to collaborate closely with American Indian communities to ensure their technologies are developed ethically and respectfully. Collaborations could entail establishing research partnerships with American Indian communities, involving community members in designing and developing ML models, and ensuring that models are built using unbiased and culturally sensitive data sets, like those of the SHS.

Traditional screening test performance

Traditional screening test performance is typically based on a gold standard in which an individual's true disease status is known to establish the test's sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). The test's sensitivity and specificity inform its effectiveness in identifying the proportion of people in the population with and without the condition of interest [20]. Sensitivity is the ability of the test to correctly identify those with the condition, while specificity is the ability of the test to correctly identify those without it. Sensitivity can be calculated from the column of those truly positive for the condition, while specificity is derived from the column of those truly negative for the condition in Figure 1.

Among these metrics, the PPV holds clinical significance for both healthcare providers and patients. The PPV is a conditional probability that the tested individual has the disease, given that they tested positive. A high PPV indicates effective identification of individuals with the tested condition, guiding further testing, diagnosis, and treatment decisions. The PPV is calculated from the row in Figure 1 that represents those subjects who tested positive for disease.

As disease prevalence decreases, screening test performance decreases, particularly concerning the PPV. This decline can lead to situations where accuracy and sensitivity remain high, giving a false impression of a well-performing test due to the increased number of false positives (FP). For example, in a population of 1,000 people with a disease prevalence of 40%, a test with a sensitivity of 90% and specificity of 80% will produce a PPV of 75%. If the disease prevalence is lowered to 10% in this same population, the PPV drops to 33.33%; hence, the prevalence dominates in screening for rare diseases [21]. Therefore, healthcare providers should consider these factors when interpreting the PPV for further testing and treatment decisions.

While sensitivity and specificity provide information about test performance across populations, PPV is often more relevant in clinical practice. It helps physicians assess the likelihood of disease presence after a positive test result, especially in populations with low disease prevalence. If the disease outcome becomes increasingly rare, the algorithm will likely always predict the absence of disease, leading to high accuracy but poor PPV [22].

This study aimed to develop and evaluate three popular and commonly used AI and ML techniques as *in silico* screening tools for predicting three chronic conditions with differing prevalences in the SHS population: peripheral artery disease (PAD), hypertension, and type 2 diabetes. Specifically, we predicted the disease outcome using epidemiological data with methods including artificial neural networks (ANNs), random forest (RF), and logistic regression (LR). Unlike their traditional laboratory-based counterparts, these *in silico* tests do not have pre-determined sensitivity or specificity; rigorous testing has not been performed using a gold standard to establish these values. Our simulations provided a glimpse of the sensitivity, specificity, and PPV of these *in silico* screening tests, as these values changed in response to differing disease prevalences. We hypothesized that these *in silico* screening tools tailored to the American Indian population would show reduced performance as disease prevalence declines, regardless of the AI or ML method.

This research serves as a reminder that the limitations of screening tests regarding disease prevalence still apply, whether those tests are *in silico* AI or ML algorithms or traditional screening tools.

Materials and methods

LR, ANNs, and RFs are well-known methods for creating *in silico* screening tests. While RFs operate as a nonlinear model, LR

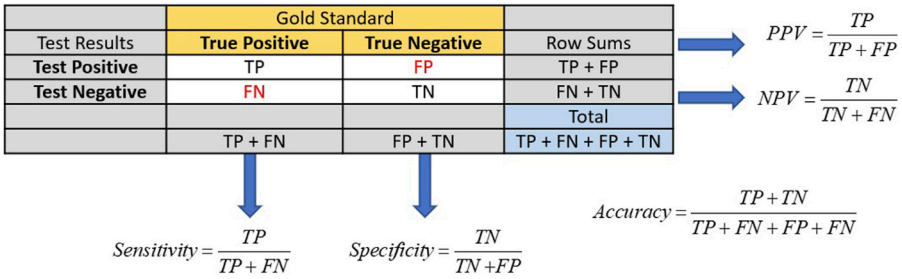


FIGURE 1
Calculations of sensitivity, specificity, PPV, and NPV for screening tests usually have their performance metrics determined via a gold standard. The numbers of true positives and negatives are represented by TP and TN, respectively. Likewise, the numbers of false positives and negatives are represented by FP and FN.

requires a linear relationship with the regression coefficients. ANNs present a more intricate approach, often featuring multiple layers commonly known as deep learning. AI and ML can potentially enhance screening for various medical conditions by illuminating linear and nonlinear data relationships. Nonetheless, it is crucial to acknowledge that the application of AI in medical research and screening exams is still nascent, and concerns over AI algorithms' accuracy and reliability linger.

Longitudinal epidemiological SHS data set

The SHS began in 1988 as a multi-center, population-based longitudinal study of cardiovascular disease (CVD) and its risk factors among American Indians. The study had three phases: a clinical examination, a personal interview, and an ongoing mortality and morbidity survey [23]. Participants from 12 different American Indian tribes were recruited from Arizona, Oklahoma, and the Dakotas, aided by volunteers from each community who promoted participation [24].

Phase II of the SHS, examining changes in risk factors for CVD in the original cohort, occurred between 1993 and 1995. The Strong Heart Family Study (SHFS), launched in Phase III (1998–1999), investigated genetic determinants of cardiovascular disease and extended recruitment to the original cohort's family members aged 18 years and older [25]. Phase IV (2001–2003) involved surveillance of the original cohort plus 90 families to continue the study of genetic markers for CVD [26]. The Phase V exam (2006–2009) continued the SHFS, which began in Phase III; all participants from Phase III and IV were invited to participate in examinations conducted at local Indian Health Service hospitals, clinics, or tribal community facilities [27]. In Phase VI (2014–2018), all surviving participants were invited to complete a medical questionnaire, and continued the morbidity and mortality surveillance continued.

Physicians on the SHS Morbidity and Mortality (M & M) review committee examined the types of health-related events requiring hospital treatment and subsequent causes of mortality, when it occurred. Two of these physicians independently reviewed fatal events for cause, with the results reconciled by a third physician. In addition, one physician reviewed the medical records regarding study participant's non-fatal events to verify specific diagnoses (i.e., stroke). This surveillance occurred yearly for both the original cohort and family cohort participants.

TABLE 1 Age, gender, and medical condition of SHS Phase V participants.

Medical condition	All	Male	Female
N (%)	2,468	977 (39.59)	1,491 (60.41)
Age (years)			
Mean (SD)	45.55 (16.41)	43.74 (16.00)	46.73 (16.58)
Median	44.40	42.70	45.70
Hypertension (%)	948 (38.41)	402 (42.41)	546 (57.59)
Diabetes (%)	631 (25.56)	240 (38.03)	391 (61.97)
PAD (%)	94 (3.81)	32 (34.04)	62 (65.96)

The available 2,468 Phase V SHS participants were divided into development and training cohorts (80%), while the remaining sample (20%) was assigned to a testing cohort. The training cohort generated the model weights, while the testing cohort assessed the algorithm's quality.

A 1-year time-to-event data set for this study was constructed from the examination date in Phase V. The M & M results and all Phases of SHS data will cumulatively provide information on the subject's medical conditions and mortality outcomes. Basic descriptive demographic statistics by gender, age, and comorbidity, including the numbers and percentages for binary variables, are listed in Table 1.

Data labels for hypertension and diabetes already existed within the SHS Phase V data set but did not include a specific label for PAD. The data included the participants' right and left

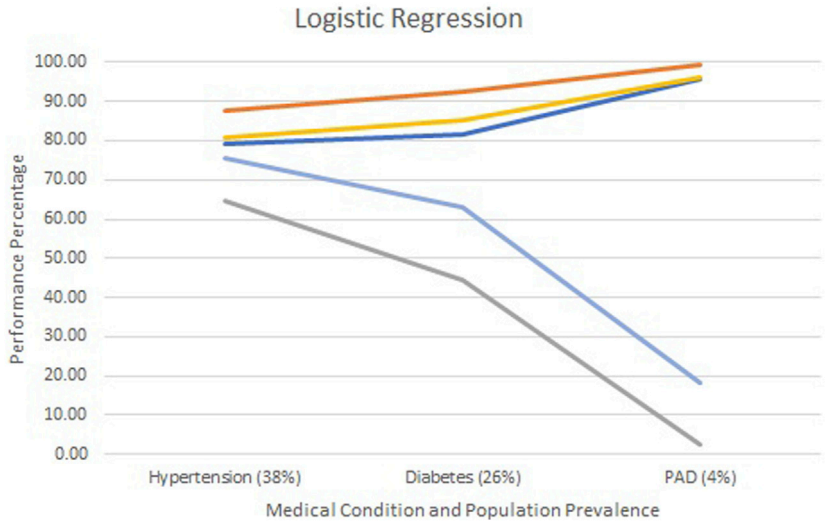


FIGURE 2
Screening test diagnostics for logistic regression. Accuracy Specificity Sensitivity PPV NPV.

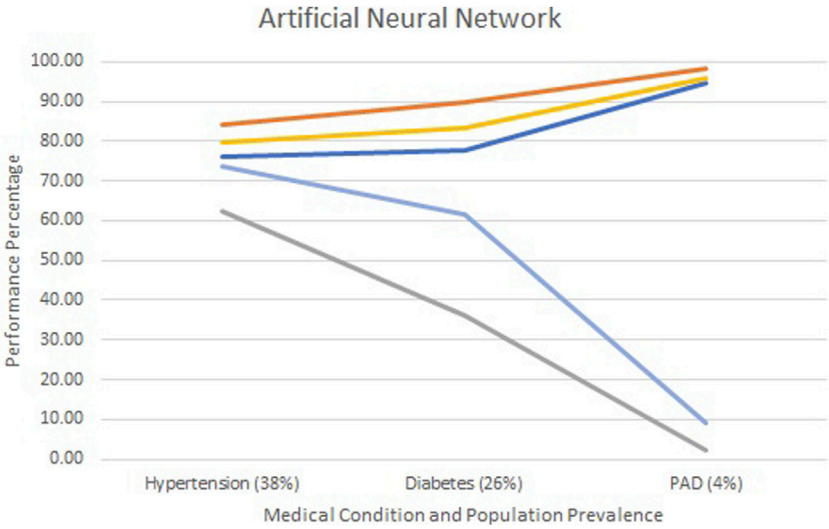


FIGURE 3
Screening test diagnostics for artificial neural networks. Accuracy Specificity Sensitivity PPV NPV.

ankle-brachial indexes (ABIs), which were used to define the presence or absence of PAD. This study used a resting ABI of less than 0.90 on either the right, left, or both sides, similar to that in Virane et al., to indicate a PAD diagnosis. Participants were coded as either 1 or 0 for the presence or absence of PAD, respectively [28, 29].

LR, ANN, and RF were then used to model the PAD, hypertension, and diabetes target features. These models ran 100 unique iterations of splitting and training the data, and producing metrics from the test set. Metrics tracked for the models were accuracy, specificity, sensitivity, PPV, and NPV, which were averaged over 100 iterations for each model type.

SAS version 9.4 was used to assemble the Phases of the SHS into a single data set, while Python version 3.9.7 was used to script the LR, RF, and ANN models.

TABLE 2 Summary of screening test diagnostics by method: logistic regression, artificial neural network, and random forest are represented by LR, ANN, and RF.

Metric	Model and chronic condition								
	Hypertension			Diabetes			PAD		
	LR	ANN	RF	LR	ANN	RF	LR	ANN	RF
Accuracy	78.99	76.00	79.35	81.49	77.66	82.18	95.57	94.56	95.99
Specificity	87.52	84.21	89.24	92.28	89.92	95.31	99.47	98.44	100.00
Sensitivity	64.79	62.41	62.88	44.59	36.06	37.26	2.40	2.07	0.00
PPV	75.61	73.62	77.75	62.88	61.77	70.13	18.06	9.04	0.00
NPV	80.63	79.87	80.12	85.08	83.52	83.88	96.06	96.00	95.99

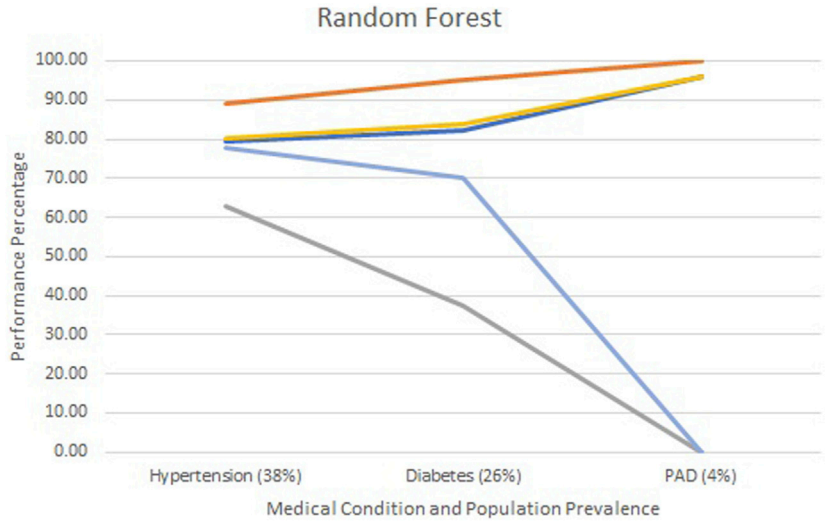


FIGURE 4
Screening test diagnostics for random forest. Accuracy Specificity Sensitivity PPV NPV.

Results

Numbers of SHS participants reporting hypertension, diabetes, and positivity for PAD are reported in Table 1.

More females than males participated in Phase V, comprising over 60% of the study participants. In addition, women reported higher percentages of hypertension, diabetes, and PAD than did men.

Figures 2–4 show each model's accuracy, sensitivity, specificity, NPV, and PPV and reflect similar performance patterns among all models. The PPV and sensitivity measures seem to suffer the most as the outcome prevalence declines, which is what is typically observed for a traditional laboratory-based screening test. PPV and sensitivity decline for all models but remain parallel for LR and appear to converge within the ANN and RF models. As PPV and sensitivity decline for the RF model, they converge to zero at an outcome prevalence of 4% (PAD). Specific numerical values for each model metric are recorded in Table 2 for all three chronic conditions.

All three models reported accuracy and specificity values that increased as the condition's prevalence declined. These two measures are roughly 95% or higher for PAD, regardless of the model selected. Conversely, sensitivity and PPV decreased as the prevalence declined, largely due to the increased number of false positives. Although poor, the LR model reported the greatest PPV of 18% for PAD, as compared to the ANN and RF, which were at 9% and 0%, respectively.

The formulas for sensitivity and PPV in Figure 1 give insight to the effect of false and true positives on these two metrics. Traditional laboratory screening tests' performance metrics are usually determined via a gold standard. As the prevalence of the

condition declined, so did the number of true positives, while that of false positives increased, driving down both the sensitivity and PPV. Accuracy remained high as the true negatives grew, inflating these metrics.

Discussion

LR, ANNs, and RFs are popular methods in the burgeoning world of AI and ML. Although these methods are quite different from one another, we can see that their performance metric trends are similar in screening for disease outcomes with varying prevalences. These performance metrics give the developers of these methods an idea of how a specific *in silico* screening method will perform in the population it was designed to serve based on the prevalence of the outcome.

Although these algorithms may have high predictive power, as measured in terms of predictive accuracy, some are criticized for lacking any causal reasoning [30]. For example, ANNs may give reliable predictions for the end users; however, these end users do not know how the algorithm came to a particular conclusion. Thus, they are "black boxes" contributing little to understanding a condition's cause.

Regardless of the method used, the PPV declined in parallel with the overall prevalence of the condition. The type of *in silico* modeling approach is still subject to the same limitations as those of traditional lab-based screening tests, an important factor to remember as online screening tests become more widespread. This study reminds us that regardless of the approach used, *in silico* AI and ML screening tests are not "magic bullets." Their performance is still limited by the prevalence of the disease in the populations they are intended to serve.

Author contributions

TM performed the modeling and Python coding, while PR completed the statistical analysis and initial manuscript writing. DW and WT contributed to the structure and content of the manuscript. YZ and JR described the Strong Heart Study design and contributed to the creation of the manuscript. All authors contributed to the article and approved the submitted version.

Author disclaimer

This manuscript reflects the views of its authors and does not necessarily reflect those of the U.S. Food and Drug Administration. Any mention of commercial products is for clarification only and is not intended as approval, endorsement, or recommendation.

Data availability

Publicly available datasets were analyzed in this study. This data can be found here: https://strongheartstudy.org/.

Ethics statement

This project was approved by the University of Oklahoma Health Sciences Center Institutional Review Board along with the Strong Heart Study Publications and Presentations Committee (SHS700). In addition, the National Center for Toxicological Research Institutional Review Board approved the project and its publication.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The Strong Heart Study has been funded in whole or in part with federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, Department of Health and Human Services, under contract numbers (75N92019D00027, 75N92019D00028, 75N92019D00029, and 75N92019D00030). The study was previously supported by research grants: R01HL109315, R01HL109301, R01HL109284, R01HL109282, and R01HL109319 and by cooperative agreements: U01HL41642, U01HL41652, U01HL41654, U01HL65520, and U01HL65521.

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

1. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Future Healthc J* (2019) **6**:94–8. doi:10.7861/futurehosp.6-2-94

2. Bohr A, Memarzadeh K. The rise of artificial intelligence in healthcare applications. In: Bohr A, Memarzadeh K, editors *Artificial intelligence in healthcare*. Academic Press (2020). p. i–iii.

3. Kumar Y, Koul A, Singla R, Ijaz MF. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. *J Ambient Intelligence Humanized Comput* (2022) **14**:8459–86. doi:10.1007/s12652-021-03612-z

4. Chalkidou A, Shokraneh F, Kijauskaite G, Taylor-Phillips S, Halligan S, Wilkinson L, et al. Recommendations for the development and use of imaging

test sets to investigate the test performance of artificial intelligence in health screening. *Lancet Digital Health* (2022) **4**:e899–e905. doi:10.1016/S2589-7500(22)00186-8

5. U. S. Government Accountability Office. Artificial intelligence in health care: benefits and challenges of machine learning technologies for medical diagnostics (2024). Available from: https://www.gao.gov/products/gao-22-104629 (Accessed July 1, 2024).

6. Sung J, Hopper JL. Co-evolution of epidemiology and artificial intelligence: challenges and opportunities. *Int J Epidemiol* (2023) **52**:969–73. doi:10.1093/ije/dyad089

7. Choi JS, Kim MH, Kim YC, Lim YH, Bae HJ, Kim DK, et al. Recalibration and validation of the Charlson comorbidity index in an Asian population: the national health insurance service-national sample cohort study. *Sci Rep* (2020) **10**:13715. doi:10.1038/s41598-020-70624-8

8. Quan H, Li B, Couris CM, Fushimi K, Graham P, Hider P, et al. Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. *Am J Epidemiol* (2011) **173**:676–82. doi:10.1093/aje/kwq433

9. Benjamin EJ, Muntner P, Alonso A, Bittencourt MS, Callaway CW, Carson AP, et al. Heart disease and stroke statistics-2019 update: a report from the American heart association. *Circulation* (2019) **139**:e56–e528. doi:10.1161/CIR.0000000000000659

10. Centers for Disease Control and Prevention. CDC and Indian country working together (2024). Available from: https://stacks.cdc.gov/view/cdc/44668 (Accessed July 1, 2024).

11. Indian health Service. Disparities (2024). Available from: https://www.ihs.gov/newsroom/factsheets/disparities/ (Accessed July 1, 2024).

12. Kaggle. Kaggle: your home for data science (2023). Available from: https://www.kaggle.com/ (Accessed April 28, 2023).

13. Chang V, Bailey J, Xu QA, Sun Z. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Comput Appl* (2022) **35**:16157–73. doi:10.1007/s00521-022-07049-z

14. Kaur H, Kumari V. Predictive modelling and analytics for diabetes using a machine learning approach. *Appl Comput Inform* (2022) **18**:90–100. doi:10.1016/jaci2018.12.004

15. Lukmanto R, Irwansyah E. The early detection of diabetes mellitus (DM) using Fuzzy hierarchical model. *Proced Comput Sci* (2015) **59**:312–9. doi:10.1016/j.procs.2015.07.571

16. Lee ET, Howard BV, Wang W, Welty TK, Galloway JM, Best LG, et al. Prediction of coronary heart disease in a population with high prevalence of diabetes and albuminuria: the Strong Heart Study. *Circulation* (2006) **113**:2897–905. doi:10.1161/CIRCULATIONAHA.105.593178

17. Wang W, Lee ET, Fabsitz RR, Devereux R, Best L, Welty TK, et al. A longitudinal study of hypertension risk factors and their relation to cardiovascular disease: the Strong Heart Study. *Hypertension* (2006) **47**:403–9. doi:10.1161/01.HYP.0000200710.29498.80

18. Wang W, Lee ET, Howard BV, Fabsitz RR, Devereux RB, Welty TK. Fasting plasma glucose and hemoglobin A1c in identifying and predicting diabetes: the strong heart study. *Diabetes Care* (2011) **34**:363–8. doi:10.2337/dc10-1680

19. Cipolle AV. How native Americans are trying to debug A.I.'s Biases (2024). Available from: https://www.nytimes.com/2022/03/22/technology/ai-data-indigenous-ivow.html (Accessed July 1, 2024).

20. Gordis L. *Epidemiology*. Philadelphia: Elsevier Saunders (2004). p. 335.

21. van Belle G. *Statistical rules of thumb*. 2nd ed. Hoboken: Basic Books, Inc (2008). p. 272.

22. Yu Z, Wang K, Wan Z, Xie S, Lv Z. Popular deep learning algorithms for disease prediction: a review. *Cluster Comput* (2023) **26**:1231–51. doi:10.1007/s10586-022-03707-y

23. Lee ET, Welty TK, Fabsitz R, Cowan LD, Le NA, Oopik AJ, et al. The Strong Heart Study. A study of cardiovascular disease in American Indians: design and methods. *Am J Epidemiol* (1990) **132**:1141–55. doi:10.1093/oxfordjournals.aje.a115757

24. Stoddart ML, Jarvis B, Blake B, Fabsitz RR, Howard BV, Lee ET, et al. Recruitment of American Indians in epidemiologic research: the strong heart study. *Am Indian Alsk Native Ment Health Res* (2000) **9**:20–37. doi:10.5820/aian.0903.2000.20

25. North KE, Howard BV, Welty TK, Best LG, Lee ET, Yeh JL, et al. Genetic and environmental contributions to cardiovascular disease risk in American Indians: the strong heart family study. *Am J Epidemiol* (2003) **157**:303–14. doi:10.1093/aje/kwf208

26. Strong Heart Study. Strong heart study phase IV operations manual (2024). Available from: https://strongheartstudy.org/portals/1288/Assets/documents/manuals/Phase%20IV%20Operations%20Manual.pdf?ver=2017-11-15-134610-080 (Accessed July 1, 2024).

27. Strong Heart Study. Strong heart study phase IV operations manual (2024). Available from: https://strongheartstudy.org/portals/1288/Assets/documents/manuals/Phase%20V%20Operations%20Manual.pdf?ver=2017-11-15-134617-657 (Accessed July 1, 2024).

28. American Heart Association. How is PAD diagnosed (2024). Available from: https://www.heart.org/en/health-topics/peripheral-artery-disease/diagnosing-pad (Accessed July 1, 2024).

29. Virani SS, Alonso A, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP, et al. Heart disease and stroke statistics-2020 update: a report from the American heart association. *Circulation* (2020) **141**:e139–e596. doi:10.1161/CIR.0000000000000757

30. Pearl J, Mackenzie D. *The book of why: the new science of cause and effect*. Basic Books, Inc. (2018). p. 432.

EBM Experimental Biology and Medicine

TYPE Original Research
PUBLISHED 19 March 2025
DOI 10.3389/ebm.2025.10359

OPEN ACCESS

*CORRESPONDENCE
Huixiao Hong,
huixiao.hong@fda.hhs.gov

RECEIVED 30 August 2024
ACCEPTED 25 February 2025
PUBLISHED 19 March 2025

CITATION
Liu J, Li J, Li Z, Dong F, Guo W, Ge W, Patterson TA and Hong H (2025) Developing predictive models for μ opioid receptor binding using machine learning and deep learning techniques. *Exp. Biol. Med.* 250:10359. doi: 10.3389/ebm.2025.10359

Developing predictive models for μ opioid receptor binding using machine learning and deep learning techniques

Jie Liu[1], Jerry Li[2], Zoe Li[1], Fan Dong[1], Wenjing Guo[1], Weigong Ge[1], Tucker A. Patterson[1] and Huixiao Hong[1]*

[1]U.S. Food and Drug Administration, National Center for Toxicological Research, Jefferson, AR, United States, [2]Department of Computer Science, Rice University, Houston, TX, United States

Abstract

Opioids exert their analgesic effect by binding to the μ opioid receptor (MOR), which initiates a downstream signaling pathway, eventually inhibiting pain transmission in the spinal cord. However, current opioids are addictive, often leading to overdose contributing to the opioid crisis in the United States. Therefore, understanding the structure-activity relationship between MOR and its ligands is essential for predicting MOR binding of chemicals, which could assist in the development of non-addictive or less-addictive opioid analgesics. This study aimed to develop machine learning and deep learning models for predicting MOR binding activity of chemicals. Chemicals with MOR binding activity data were first curated from public databases and the literature. Molecular descriptors of the curated chemicals were calculated using software Mold2. The chemicals were then split into training and external validation datasets. Random forest, k-nearest neighbors, support vector machine, multi-layer perceptron, and long short-term memory models were developed and evaluated using 5-fold cross-validations and external validations, resulting in Matthews correlation coefficients of 0.528–0.654 and 0.408, respectively. Furthermore, prediction confidence and applicability domain analyses highlighted their importance to the models' applicability. Our results suggest that the developed models could be useful for identifying MOR binders, potentially aiding in the development of non-addictive or less-addictive drugs targeting MOR.

KEYWORDS

μ opioid receptor, binding activity, machine learning, deep learning, predictive model

Impact statement

This work is crucial in addressing the opioid crisis by focusing on the development of non-addictive or less-addictive opioid analgesics. Current opioids, while effective for pain relief, pose significant risks of addiction and accidental overdose. By elucidating the structure-activity relationship between the μ opioid receptor (MOR) and its ligands, this study advances the field through the development of machine learning and deep learning models to predict MOR binding activity. Evaluated via rigorous cross-validation, the models showed robust predictive capabilities. This research imparts new insights into the prediction of MOR binding, emphasizing the importance of prediction confidence and applicability domain analyses. The developed models have the potential to identify new MOR binders, significantly impacting the field by guiding the design of analgesics that mitigate the risk of addiction and overdose, ultimately improving patient safety and public health outcomes.

Introduction

The opioid epidemic refers to the widespread misuse, addiction, and overdose deaths associated with prescription opioids and illicit drugs like heroin and synthetic opioids such as fentanyl. For many years this crisis has been a significant public health issue in the United States [1–3]. As reported by the CDC, over 105,000 drug overdose deaths were recorded in the US in 2022 [4]. Notably, between 2020 and 2021, the mortality rate from drug overdoses involving synthetic opioids excluding methadone rose by 22%, whereas deaths involving heroin decreased by 32% [5]. Until recently, the predominant cause of synthetic opioid-related deaths was attributed to fentanyl and its analogs [5, 6]. Therefore, opioid use disorder (OUD) poses a significant public health challenge, contributing to illness and mortality through addiction, overdose, and associated medical complications [7, 8]. Besides the public health issue, the opioid crisis has also caused a severe economic burden. For example, Florence et al. [9] projected the economic toll of the opioid crisis at $1.02 trillion in 2017. This encompasses the staggering costs attributed to lives lost from opioid overdose ($480.8 billion) and the diminished quality of life resulting from OUD ($390.0 billion), collectively representing more than 85% of the overall economic impact.

Numerous efforts have been dedicated to addressing the opioid crisis, encompassing enhanced regulation of opioid prescription practices [10–12], broadened accessibility to addiction treatment and harm reduction services [13–15], public awareness campaigns to highlight opioid risks [16, 17], and steps aimed at decreasing the availability of illicit opioids [9, 18]. The profound addictiveness of opioids is closely linked to the overdose fatalities caused by prescription opioids, heroin, and illicit fentanyl. However, given the important role of prescription opioids as powerful analgesics, outright prohibition of these medications is not feasible [19, 20].

Opioid drugs achieve their analgesic effects by binding to opioid receptors, including the μ opioid receptor (MOR) [21]. MOR is a primary target for analgesics. Since the discovery of MOR in the 1970s, significant efforts have been made to elucidate the relationship between the receptor and its ligands in the hopes of guiding the development of new drugs with high analgesic efficacy, fewer side effects, and a lower risk of tolerance, dependence, and addiction [22, 23]. Numerous morphine-based semi-synthetic opioids (such as oxycodone, heroin) and fully synthesized opioids (such as fentanyl) have been developed; nevertheless, none of these opioids have demonstrated both safety and efficacy as analgesics [24]. Moreover, bringing a new drug to market typically requires an investment of nearly $2.6 billion and over a decade of time [25–27]. With the increasing computational power and data sources, computational modeling using machine learning and deep learning has become a promising approach to reduce the time and cost of new drug development [21, 28–36]. Multiple computational models have been constructed for the binding activity prediction of compounds to diverse opioid receptors [37–42]. Floresta et al. [37] established three quantitative structure-activity relationship models (one field-based 3D model and two molecular fingerprint based 2D k-nearest neighbors (kNN) models) based on a dataset of 115 fentanyl-like compounds. Sakamuru et al. [38] generated models to predict both agonistic and antagonistic activity of multiple opioid receptors, including MOR, based on quantitative high-throughput screening (qHTS) assay data. Pan et al. [39] established a 3D-QSAR model to predict δ opioid receptors binding activity. The training set included 46 compounds collected from five publications. Feng et al. [40] developed machine learning and deep learning models for predicting the inhibitory activity of 75 proteins involved in opioid receptor networks, including models for MOR trained on 4,667 compounds collected from the ChEMBL database, to assess the screening and repurposing potential of more than 120,000 drug candidates targeting four opioid receptors. Leveraging transfer learning, Provasi and Filizola [41] constructed deep learning models for predicting the bioactivity of opioid receptors using ligand-based and structure-based molecular descriptors. Their MOR binding activity predictive model was trained on 87 active compounds from the IUPHAR/BPS Guide to Pharmacology database and 1,058 inactive chemicals from ChEMBL database, with inactivity determined by a -log10 of Ki, IC_{50}, or EC_{50} less than 5. However, this approach raises concerns about model reliability, as many compounds defined as inactive exhibited some agonistic or antagonistic activity, increasing the potential for false negatives. Instead of predicting MOR binding activity, Oh et al. [42] developed machine learning and deep learning

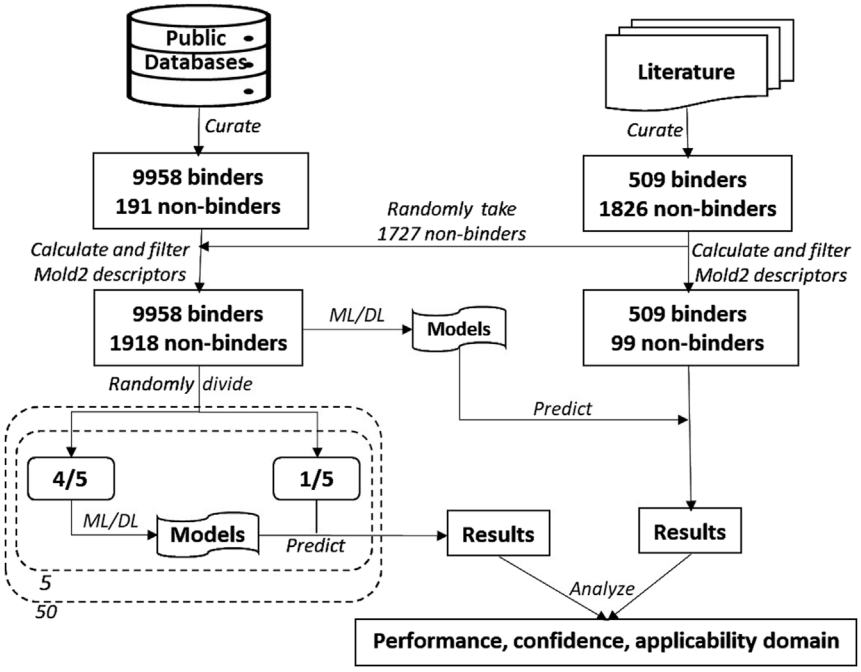


FIGURE 1
Study overview. The data on chemicals and their MOR binding activity were curated from public databases and the literature. The dataset from these databases was augmented with 1,727 non-binding chemicals sourced from the literature, forming the training dataset. The remaining chemicals from the literature constituted the external validation dataset. Molecular descriptors were calculated using Mold2 and subsequently filtered. Five algorithms—random forest, k-nearest neighbors, support vector machine, multi-layer perceptron, and long short-term memory—were used to build predictive models. The training dataset underwent 50 iterations of 5-fold cross-validation. Models constructed using the entire training dataset were then used to predict MOR binding on the external validation dataset. The performance of the models was evaluated based on their cross-validation and external validation predictions, with an additional focus on analyzing prediction confidence and applicability domain.

models for differentiating MOR agonists from antagonists. These models were trained on a small dataset (755 agonists and 228 antagonists) and evaluated with an even smaller dataset (15 agonists and 11 antagonists). The small size of the datasets and the narrow chemical space of the compounds in training these models limit the applicability of the developed models.

To enhance performance, robustness, and generalization capability of MOR binding activity prediction models, large sizes of diverse chemicals are needed in training the models. Therefore, this study collected a large size of diverse chemicals to construct machine learning and deep learning models for MOR binding activity prediction. Moreover, multiple machine learning and deep learning algorithms were adopted. We first curated MOR binding activity data of chemicals from public databases and publications. Machine learning and deep learning models were then built using multiple algorithms and validated by cross-validation and external validation. Moreover, prediction confidence and applicability domain (AD) derived from our models offer additional metrics for more appropriate applications of our models. Validation results demonstrate that the developed models could help in identifying compounds that bind to MOR, potentially facilitating the development of opioid drugs with reduced addictive properties.

Materials and methods

Study design

Study design is illustrated in Figure 1. First, chemicals with MOR binding activity data were curated from public databases as the training dataset. Chemicals with qHTS assay data reported in the literature were also curated. After removing chemicals that are contained in the training dataset, some of the inactive chemicals in qHTS assays were added to the training dataset and the rest, including active and inactive chemicals, were used as an external validation set. Molecular descriptors for the chemicals in both training and external validation datasets were then calculated using Mold2 [43, 44]. Five machine learning and deep learning algorithms, including random forest (RF) [45], kNN [46], support vector machine (SVM) [47], multi-layer perceptron (MLP) [48], and long short-term memory (LSTM) network [49], were applied in construct models.

In addition, a consensus model was generated by combining the models built with each of these algorithms. Fifty iterations of 5-fold cross-validations were conducted on the training dataset for estimating the performance of the developed models. Models were constructed on the entire training dataset using these algorithms, and their generalizing capability in predicting MOR binding activity of unseen chemicals was evaluated using the external dataset. Multiple metrics were calculated for measuring model performance. At last, prediction confidence and AD were analyzed based on the predictions from both cross-validations and external validations.

Data sources

Compounds with experimental MOR binding activity data were curated from PubChem[1], BindingDB[2], and ChEMBL[3] databases. Compounds having quantitative MOR binding activity data such as IC_{50}, Ki, and Kd values were designated as binders. For compounds without quantitative binding activity data, the qualitative binding activity description field was used to determine if a compound is MOR binder or non-binder. Compounds marked as "not determined" or "inconclusive" were excluded. Compounds marked as "active" or "positive" were assigned as binders, while chemicals marked as "inactive" or "negative" were treated as non-binders. Compounds with qHTS assay data on MOR used in Sakamuru et al. [38] were downloaded from[4]. The data from columns "OPRM agonist outcome" and "OPRM antagonist outcome" in both the training and validation datasets were used. A compound is inactive in both agonist and antagonist assays was termed as a non-binder, while a compound is active in one of both assays was designated as a binder. The simplified molecular input line entry system (SMILES) strings of compounds in both the public databases and the datasets from the publication were collected for representing their chemical structures.

Data processing

The SMILES strings of the compounds obtained from the public databases (PubChem, BindingDB, and ChEMBL) and the publication (Sakamuru et al. [38]) were first converted to unique SMILES strings using the Online SMILES Translator and Structure File Generator [50]. For compounds with the same unique SMILES strings and the same activity class (binder or non-binder), only one compound was kept. Compounds with the same unique SMILES strings and different activity classes were excluded. Consequently, 10,149 compounds (9,958 binders and 191 non-binders) from the public databases and 2,527 compounds (509 binders and 2,018 non-binders) from Sakamuru et al. [38] remained. Of the 2,527 compounds, 192 non-binders are contained in the 10,149 compounds from the public databases (Supplementary Table S1) and were further removed. Finally, 509 binders and 1,826 non-binders from Sakamuru et al. were used. The binder/non-binder ratios of these two datasets (52.14 and 0.28) are dramatically different. Therefore, we randomly took 1,727 non-binders from the qHTS dataset and added them to the dataset from the public databases, resulting in a dataset of 9,958 binders and 1,918 non-binders as the training dataset (Supplementary Table S2). The remaining 509 binders and 99 non-binders from the qHTS assays were used as the external validation dataset (Supplementary Table S3). The used training and external validation datasets have similar binder/non-binder ratios. The SMILES strings of both training and external validation datasets were used to generate two-dimensional (2D) structures of the compounds using the Online SMILES Translator and Structure File Generator [50]. The SDF files obtained were used for subsequent molecular descriptors calculation.

Descriptors calculation and filtering

Converting chemical structures into machine-readable formats is essential for developing machine learning and deep learning models [51]. In this study we utilized software tool Mold2 for calculating molecular descriptors for the compounds in the training and external validation datasets. Mold2 only accepts SDF (structure data file) representation of chemical structures [43, 44]. Therefore, the unique SMILES strings of compounds were first converted to SDF files for the training and external validation datasets using the Online SMILES Translator and Structure File Generator [50]. The generated SDF files were then input into Mold2 software for calculating molecular descriptors. Mold2 calculated 777 molecular descriptors for each compound.

Molecular descriptors with no or very low information for a dataset can significantly influence the performance of models developed using the dataset. To identify and remove such low informative descriptors, we first excluded 263 descriptors with a constant value for more than 90% of the compounds in the training dataset. Subsequently, we performed Shannon entropy analysis [43, 52–54] on the remaining 514 descriptors of the training dataset. In brief, for each molecular descriptor, the range of descriptor values of the compounds in the training dataset were first divided into 20 groups with equal value intervals. The compounds in the training dataset were then put into these

1 https://pubchem.ncbi.nlm.nih.gov/
2 https://www.bindingdb.org/rwd/bind/index.jsp
3 https://www.ebi.ac.uk/chembl/
4 https://tripod.nih.gov/tox/oprpred/codes.zip

20 groups based their descriptor values. The distribution in the 20 groups, probabilities of compounds in the 20 groups, were calculated by dividing compound counts by the total compounds of the training dataset. At last, Shannon entropy values were computed for each descriptor using Equation 1.

$$H_n(p_1, p_2, \cdots, p_n) = -\sum_{i=1}^{n} p_i \log_2 p_i \quad (1)$$

Where p_i is the probability of group i. 226 descriptors with Shannon entropy less than 2.0 were considered as low informative and removed. The remaining 288 molecular descriptors have Shannon entropy values greater than or equal to 2.0 and were used in subsequent model development. The 288 molecular descriptors are listed in Supplementary Table S4. For the external validation dataset, the same 288 molecular descriptors were kept, and other descriptors were removed.

Scaling descriptor values

The values of different molecular descriptors usually are in quite different scales in a dataset. Using unscaled molecular descriptor values to construct machine learning and deep learning models often result in low performance for most algorithms, depending on their mathematical principles. Therefore, scaling is generally needed before model development. We scaled the values of each molecular descriptor in training and external validation datasets using Equation 2.

$$V = \frac{V_o - Min_{train}}{Max_{train} - Min_{train}} \quad (2)$$

Where V is scaled value, V_o is original value, Min_{train} is the minimum value of the descriptor in the training set, and Max_{train} is the maximum value in the training set.

Model development

MOR binding activity prediction models were built using three machine learning algorithms (RF, kNN, and SVM) and two deep learning algorithms (MLP and LSTM). Numerous machine learning and deep learning algorithms have been developed, each grounded in distinct mathematical principles. kNN is a widely used, simple, and interpretable algorithm. In contrast, RF and SVM are more complex but have demonstrated good performance in various applications. However, the complexity of RF and SVM makes them challenging to interpret. We chose these three to explore the performance difference between simple and complicate machine learning models. Furthermore, MLP and LSTM represent two fundamentally different deep learning architectures: MLP is a feedforward neural network, whereas LSTM is a recurrent neural network. These two were selected to evaluate the performance of deep learning models constructed using algorithms with distinct structural designs.

When building a model using an algorithm, related algorithmic parameters were tuned through inner 5-fold cross validations. Briefly, to optimize algorithmic parameters the training set was randomly split into five folds. Four folds were used to build a model to predict the remaining fold. This process was repeated five times so that each of the five folds was used once and only once as a testing set. The prediction results on all five folds were then used to calculate a Matthews correlation coefficient (MCC) value. This inner 5-fold cross-validation was repeated five times with different random divisions of the training set into five folds. At last, the five MCC values from five iterations of inner 5-fold cross-validations were averaged to estimate performance of models built with a set of parameters. The set of hyperparameters resulting in the highest average MCC value was determined as the optimized parameters for the algorithm and were used to construct a model on the training set.

The hyperparameters tuned in our study are given below. For RF, n_estimators (100 and 200 trees), min_samples_leaf (10 and 20), and max_chemical (1,000 and 2,000) were tuned. For kNN, the parameters n_neighbors (k = 3, 5, and 7) and weights ("uniform," "distance") were optimized. For SVM, a linear kernel was employed and the regularization parameter C (0.1, 1, and 10) was optimized. For MLP, alpha (0.0001, 0.1) and hidden_layer_sizes (100, 300) were optimized. For LSTM, the number of epochs was tuned to 500 with running 5,000 epochs based on the training loss value and accuracy. Other parameters used for LSTM include recurrent layers = 4, features = 200, batch size = 32, and learning rate = 0.0001. For these five algorithms, except the parameters aforementioned, default values were adopted for other algorithmic parameters.

The RF, kNN, SVM, and MLP models were constructed using the packages in Scikit-learn (0.23.2) [55] in Python (3.8.5) [56], while the LSTM models were developed using a PyTorch package (2.0.1) [57] in Python (3.8.5).

In addition to models developed using the five algorithms, a consensus model was constructed for a training set using the five individual models. Each of these models was built using distinct algorithms, potentially utilizing different features from the same training dataset. Consensus modeling capitalizes on the strengths of each model, aggregating their predictions to deliver more reliable, robust, and accurate results. This approach enhances the overall performance by minimizing the weaknesses inherent in any single model. The consensus model combines outcomes from its five individual models using a majority voting strategy: if three or more individual models predict a compound as MOR binder, the compound is determined as a binder, otherwise, it is predicted as non-binder.

Model evaluation

Model performance was evaluated using two strategies: 5-fold cross-validation and external validation. In a 5-fold cross-validation, the entire training set was first randomly divided into five equal or close to equal folds. Four of the five folds were then used to tune algorithmic parameters using the inner 5-fold cross-validations for each of the machine learning and deep learning algorithms. The tuned parameters were then used to train models on the four folds, and the trained models were used to predict the remaining fold. This process was repeated five times so that each of the five folds was used as a testing set only once. At last, performance metrics values were calculated using prediction results from all five testing sets to estimate model performance. The 5-fold cross-validation was repeated 50 times to reach statistically robust estimations on model performance.

The training dataset was randomly split into five subsets, four subsets for training and one subset for testing. This random splitting was repeated five times to ensure all compounds were used for both training and testing.

External validation was employed to evaluate the generalization of the constructed models using the entire training set. The same parameter tuning process was applied to the whole training set. The optimized parameters were then used to develop models using the entire training set. Finally, the developed models were used to predict MOR binding activity for compounds in the testing set.

Performance metrics

Five metrics were used to measure model performance, including accuracy, sensitivity, specificity, balanced accuracy, and MCC. These metrics were derived by comparing model predictions with actual binding activity data. They were calculated using Equations 3–7.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{5}$$

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \tag{6}$$

$$\text{MCC} = \frac{\text{TP*TN} - \text{FP*FN}}{\sqrt{(TP + FP)^{*}(TP + FN)^{*}(TN + FP)^{*}(TN + FN)}} \tag{7}$$

Where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

Prediction confidence analysis

Predictions produced by our machine learning or deep learning models provide not only class assignments but also probabilities that quantify the likelihoods of these class assignments. The prediction probability of a prediction not only classifies the compounds as MOR binder or non-binder, but also measures the confidence of the prediction. Prediction confidence analysis is conducted to evaluate if prediction confidence can be used as an additional valuable parameter to inform better utilization of a model in applications, such as decision-making and safety assessment. The prediction confidence of a prediction is derived from the prediction probability using Equation 8 [43, 53, 54, 58].

$$\text{Prediction confidence} = \frac{|\text{prob} - 0.5|}{0.5} \tag{8}$$

where *prob* is the probability of a compound predicted as a MOR binder from a machine learning and deep learning model. Prediction confidence values are between 0 and 1. The larger the value the more confidence in the prediction.

To examine the relationship between prediction confidence and prediction performance for predictions of a model in 5-fold cross-validations or external validation, the prediction confidence value range (between 0 and 1) was divided into 10 even bins with the interval of 0.1. Next, the predictions were allocated to the 10 bins according to their prediction confidence values. Lastly, performance metrics were separately calculated for predictions in each the 10 bins.

Applicability domain (AD) analysis

AD of a model represents the structural space of chemicals utilized to train the model. Chemicals falling within the AD of a model exhibit structural similarities to the training chemicals, thus yielding more accurate predictions. Therefore, AD analysis plays a crucial role in evaluating the predictions made by computational models [59–61]. In this study, the AD of a model was defined by the boundaries of all descriptors ranging from the minimum to the maximum values of chemicals used in training the model. More specifically, we first computed the AD of a model using the training chemicals. Next, the distance of a chemical to the AD was calculated using Equation 9.

$$Distance = \sqrt{d_1^{\,2} + d_2^{\,2} + \ldots + d_n^{\,2}} \tag{9}$$

Where d_i (i = 1, 2, . . ., n) is the distance of the chemical to the AD for molecular descriptor *i*. If the value of molecular descriptor *i* falling in the value range of the same molecular descriptors of the training chemicals, d_i was set to zero. Therefore, when all molecular descriptor values of a chemical

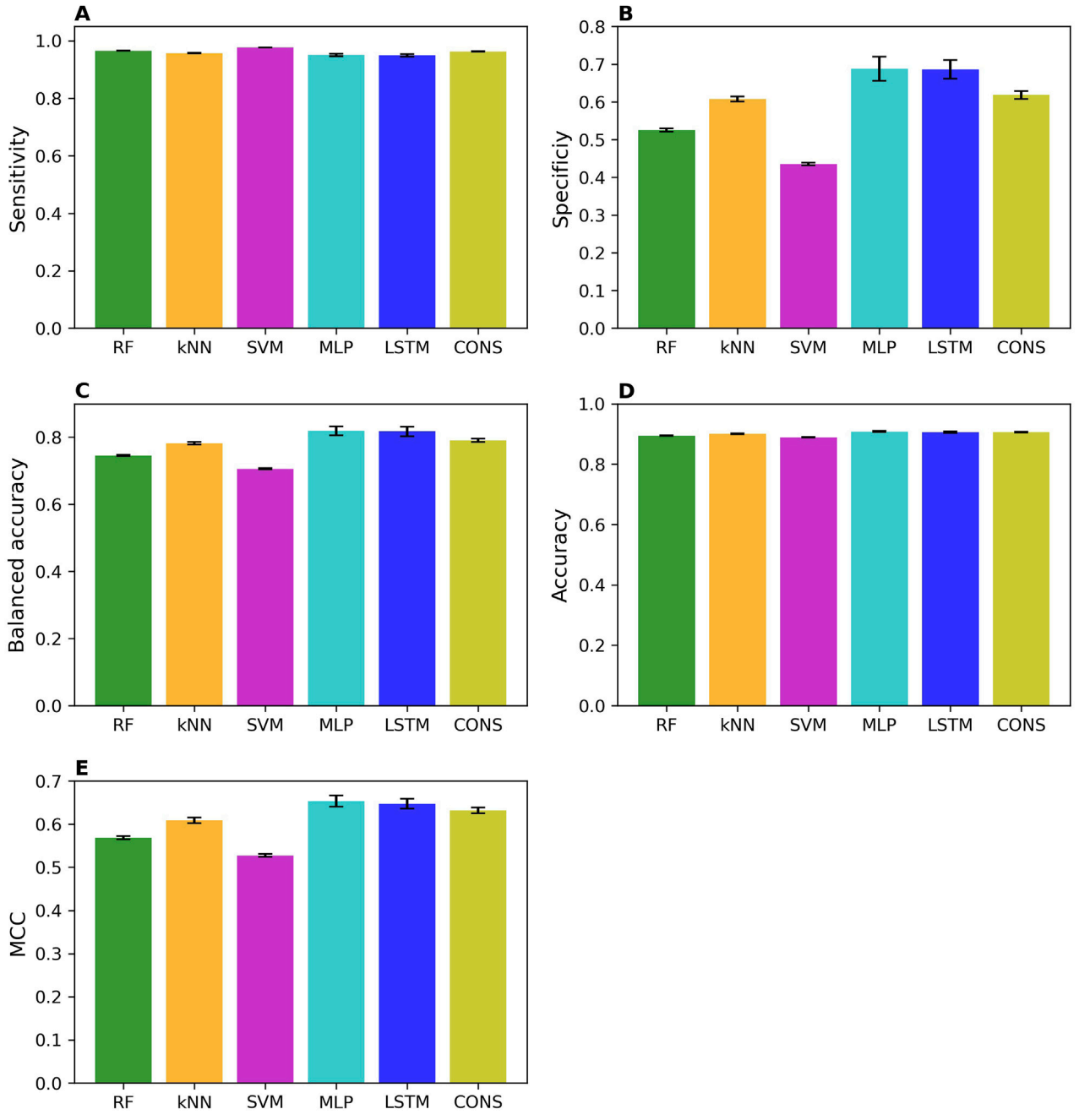


FIGURE 2
Performance of cross-validations. Performance of 50 iterations of 5-fold cross-validations was measured using sensitivity **(A)**, specificity **(B)**, balanced accuracy **(C)**, accuracy **(D)**, and MCC **(E)**. The average values of these metrics across the 50 iterations are represented by color bars, corresponding to different algorithms indicated by the x-axis labels (RF, random forest; kNN, k-nearest neighbors; SVM, support vector machine; MLP, multi-layer perceptron; LSTM, long short-term memory; and CONS, consensus model). The standard deviations are displayed as error bars on top of the color bars.

fall within the molecular descriptor value ranges of training chemicals, the *Distance* value is calculated to be zero according to Equation 9 and the chemical is considered inside the model's AD. If the value of any molecular descriptor is outside the descriptor value boundary of the training set, the *Distance* value is greater than zero and the chemical is considered outside the model's AD. At last, performance of predictions inside and outside AD was compared for the models constructed in the 5-fold cross-validations and the model built in the external validation.

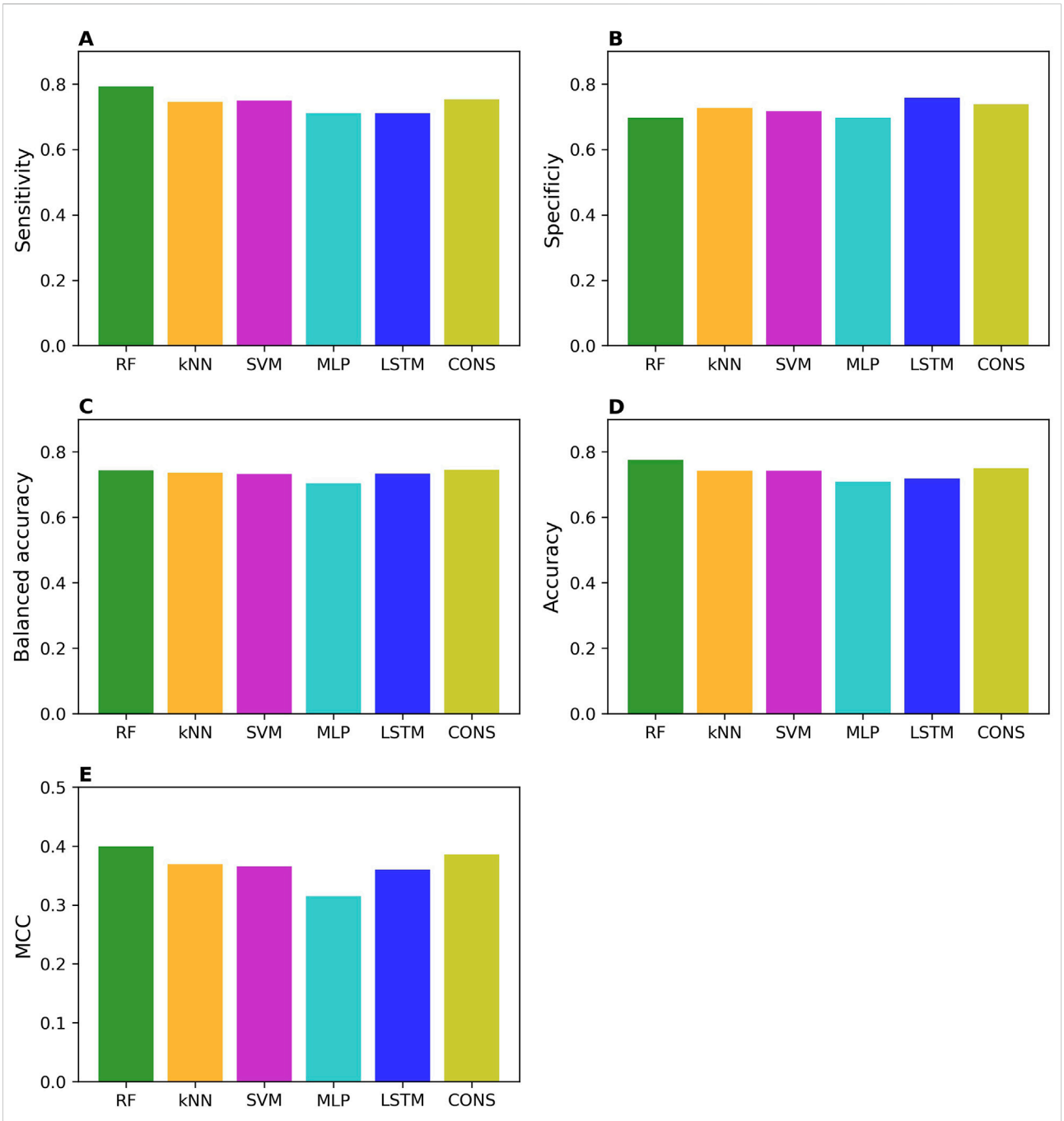


FIGURE 3
Performance of external validations. The performance was assessed using sensitivity **(A)**, specificity **(B)**, balanced accuracy **(C)**, accuracy **(D)**, and MCC **(E)**. The values of these metrics are represented by color bars for models developed using different algorithms, as indicated by the x-axis labels (RF, random forest; kNN, k-nearest neighbors; SVM, support vector machine; MLP, multi-layer perceptron; LSTM, long short-term memory, and CONS, consensus model).

Results

Model performance

The prediction performances of the machine learning (RF, kNN, and SVM), deep learning (MLP and LSTM), and consensus models from the 50 iterations of 5-fold cross-validations were summarized in Figure 2 in sensitivity (Figure 2A), specificity (Figure 2B), balanced accuracy (Figure 2C), accuracy (Figure 2D), and MCC (Figure 2E). Overall, all models performed well, as indicated by the averaged performance metrics values (the bars in Figure 2).

More specifically, performance metrics accuracy (0.89 – 0.91), balanced accuracy (0.71 – 0.82), and MCC (0.53 – 0.65) are high for all models, indicating good overall performance. Not surprisingly, all models performed much better on MOR binders than non-binders, with much higher averaged sensitivity (0.95 – 0.98) than specificity (0.44 – 0.69) because the training dataset has a greater number of MOR binders (9,958) than non-binders (1,918). Notably, all performance metrics exhibit small standard deviations among the 50 iterations of cross-validations (the sticks atop the bars in Figure 2), suggesting that the machine learning, deep learning, and consensus models were relatively unaffected by the random partitioning of the whole training dataset into five folds. Interestingly, the two deep learning models (the cyan and blue bars in Figure 2) outperformed the three machine learning models (the green, light brown, and magenta bars in Figure 2), especially in specificity (Figure 2B). These results demonstrate that a large dataset, such as the one training dataset in this study with 11,876 chemicals, is needed for deep learning algorithms to show superiority over conventional machine learning algorithms, especially for the minority class of an imbalanced dataset. Surprisingly, the consensus models did not surpass all member models. They performed better than the three machine learning models but worse than the two deep learning models. Our results indicate that though consensus modeling remains as an effective approach to combine models constructed using conventional machine learning algorithms, it deserves further investigation on if and how a consensus approach can be applied to models built with deep learning algorithms as we only used one consensus strategy, majority voting.

The external validation performances on the models trained with the whole training set are summarized in Figure 3. The external validation results indicate that the models performed well, with good performance metrics values; sensitivity of 0.71–0.79, specificity of 0.70–0.76, balanced accuracy of 0.70–0.74, accuracy of 0.71–0.78, and MCC of 0.32–0.40. As expected, they slightly underperformed the models in the cross-validations. Strikingly, specificity and sensitivity are very close in the external validations, in contrast to the cross-validation results where sensitivity is expectedly higher than specificity. This difference may attribute to the nature of data in the training and external validation datasets. The MOR binders in the training dataset are determined by conventional low-put assays, resulting in models that perform well in predicting chemicals tested with the same assays. In contrast, most of the MOR non-binders in the training dataset are results from qHTS assays, and thus the trained models performed better on the external MOR non-binders determined by the same qHTS assays. Our results suggest that caution should be exercised in the validation and application of machine learning and deep learning models.

Surprisingly, not like in the cross-validations, the two deep learning models did not consistently outperform all three machine learning models in the external validations.

Prediction confidence analysis

The prediction confidence analysis was performed on the results of cross-validations and external validation. The accuracy values of the predictions at 10 confidence levels from the cross-validations are shown in Figure 4A for all models. The accuracy of predictions is improved when their prediction confidence level is increased, for all models. Similar trends were observed for sensitivity, specificity, balanced accuracy, and MCC as depicted in Supplementary Figures S1–S4, respectively. Moreover, more predictions fall in higher confidence levels for all models except SVM as shown in Figure 4B.

The prediction confidence analysis was also conducted on the results of external validation. The accuracy values of the predictions at 10 confidence levels from the external validations are shown in Figure 4C for all models. The trends are similar to those observed in the cross-validations: higher prediction confidence levels correspond to greater prediction accuracy. However, the trend lines are less smooth than those in the cross-validations, due to the significantly fewer predictions at each confidence level. The sensitivity, specificity, balanced accuracy, and MCC values of the predictions at 10 confidence levels from the external validations exhibit similar trends as shown in Supplementary Figures S5–S8, respectively. Notably, the SVM model had very few predictions at high confidence levels, 3, 1, and 3 at confidence levels 0.7–0.8, 0.8–0.9, and 0.9–1.0, respectively. Therefore, performance metrics at these confidence levels were not calculated because they would not be statistically meaningful. The number of predictions is plotted against prediction confidence level in Figure 4D. In general, the number of predictions does not differ much in confidence levels except the SVM model which had fewer predictions at higher confidence levels.

Applicability domain analysis

The distances to the AD of the compounds predicted in the cross-validations and external validations were computed. Prediction accuracy values inside and outside the AD were calculated separately and are illustrated in Figure 5A for the cross-validations and Figure 5B for the external validation. It was clear that the compounds inside the AD were predicted more accurately than those outside the AD by all models, in both cross-validations and external validations. The sensitivity, specificity, balanced accuracy, and MCC values of predictions inside and outside AD for all models are presented in Supplementary

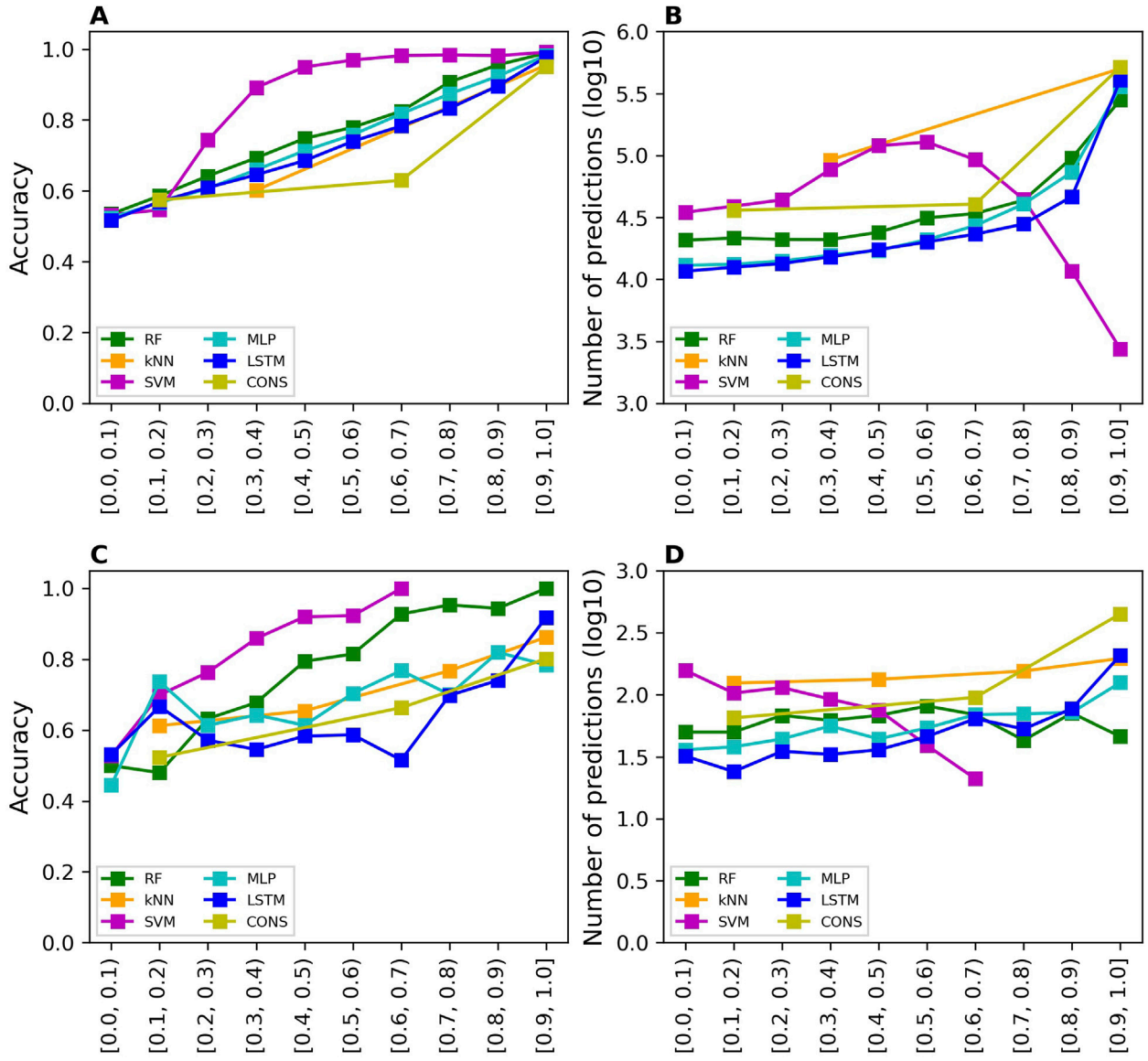


FIGURE 4
Prediction confidence analysis results. The analysis of prediction confidence is depicted by plotting prediction accuracy values and the number of predictions at various confidence levels. **(A, B)** show the results for cross-validations, while **(C, D)** display the results for external validations. The x-axis tick labels represent the different confidence levels. The models developed using different algorithms are distinguished by various colors, as indicated in the color legend (RF, random forest; kNN, k-nearest neighbors; SVM, support vector machine; MLP, multi-layer perceptron; LSTM, long short-term memory, and CONS, consensus model).

Figures S9–S12, respectively, for the cross-validations, and in Supplementary Figures S13–S16 for the external validations.

In terms of overall performance metrics (accuracy, balanced accuracy, and MCC), all models performed better inside the AD than outside it. However, examining performance on binders and non-binders, it was found that the compounds outside the AD had higher specificity than those inside the AD for all models in both cross-validations and external validations. This indicates that all models performed better on MOR binders than on non-binders. Moreover, the deep learning models achieved higher MCC and balanced accuracy than the machine learning models, both inside and outside the AD, in both cross-validations and external validation.

The results demonstrated that AD analysis is beneficial for evaluating the reliability of predictions from both machine learning and deep learning models. It is worth noting that the MCC values for predictions outside the AD in the external validations are zeros for SVM, MLP, and LSTM models

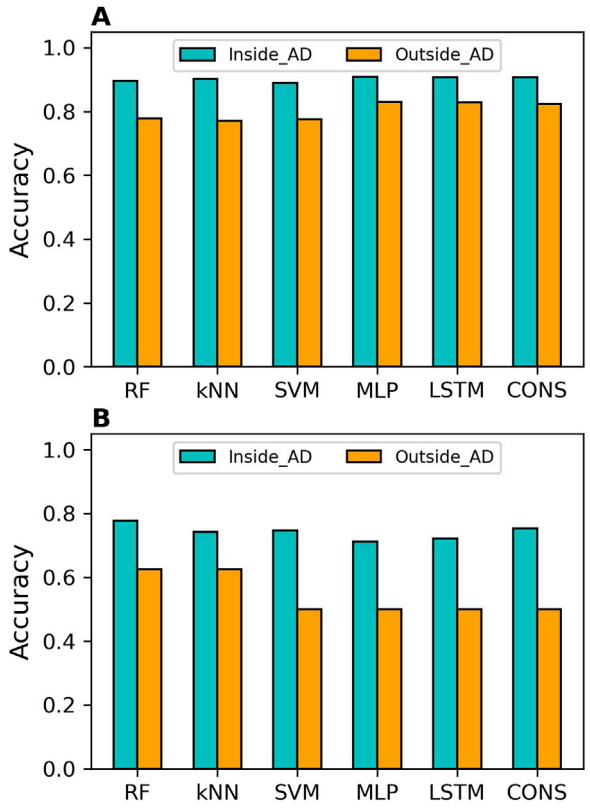


FIGURE 5
Applicability domain (AD) analysis results. The AD analysis is presented for cross-validations **(A)** and external validations **(B)**. Accuracy for predictions within the AD is shown in cyan bars, while accuracy for predictions outside the AD is displayed in orange bars. The models developed using different algorithms are represented by the x-axis labels (RF, random forest; kNN, k-nearest neighbors; SVM, support vector machine; MLP, multi-layer perceptron; LSTM, long short-term memory, and CONS, consensus model).

(Supplementary Figure S16). This might not be statistically robust due to the small number (8) of compounds.

Discussion

The opioid epidemic is a severe public health crisis in the United States, leading to an increasing number of deaths and imposing a substantial economic burden. Opioids are potent analgesics, but many are addictive and prone to cause an overdose. Hence, the non- or less-addictive drugs that target MOR are needed. Developing new drugs is a lengthy and costly process, often taking about a decade and billions of dollars. Therefore, computational approaches provide a promising and efficient way to aid drug development. In this study, we collected chemicals with MOR binding activity data from multiple databases and the literature. We then constructed and evaluated machine learning and deep learning models using the curated data for MOR binding activity prediction.

The curated data are imbalanced, which is common in the real world. The data collected from databases have a greater number of MOR binders than non-binders. Conversely, the qHTS data acquired from the literature have more MOR non-binders than binders. Hence, to maintain a consistent ratio of binders to non-binders in both the training and testing datasets, 1,727 non-binders were randomly taken from the qHTS data and added to the data collected from databases to form the training dataset. The remaining qHTS data were used as the external validation dataset. The same prevalence of MOR binders in both the training and external validation datasets can reduce the impact of difference in prevalence on external validation results, enhancing the reliability of extrapolation assessment for the developed models.

Both the training and external validation datasets are biased toward MOR binders. In two-class classification models, accuracy tends to favor the majority class, which, in this study, is the MOR binders. This bias affects the evaluation of prediction performance, especially in imbalanced datasets. Therefore, we also employed balance accuracy and MCC to measure overall performance.

Interestingly, all models had higher sensitivity than specificity, especially in the cross-validations (Figures 2A, B). This discrepancy arises because the training dataset contains a greater number of MOR binders than non-binders, enabling the models to learn the structures of MOR binders better than those of non-binders. Consequently, this results in more accurate predictions for binders (higher sensitivity) compared to non-binders (lower specificity). These findings suggest that incorporating more non-binders into the training dataset would likely improve the prediction performance of machine learning and deep learning models. To address this issue, future efforts should focus on incorporating a more balanced representation of non-binders in the training dataset. This would help reduce the imbalance and, in turn, enhance the robustness and reliability of the models. Additionally, we recommend that the scientific community place equal value on the publication of inactive results, alongside active findings. Acknowledging and appreciating inactive data will not only contribute to a more balanced dataset but also foster greater transparency and scientific rigor in the field.

It is worth noting that the prediction performance of all models in the external validations is worse than in the cross-validations. This discrepancy is not surprising because the MOR binding activity data in the training and external validation datasets are obtained from different types of assays. The training dataset, except for 1,727 non-binders, consists of results from traditional assays, whereas the external validation dataset is generated using the qHTS assay. Hence, the poorer performance of the machine learning and deep learning models in the external validations compared to the cross-validations

cannot be fully attribute to the models extrapolating to different chemicals. The difference in experimental methods that produced the MOR binding activity data in the two datasets likely contribute to this performance discrepancy.

To confirm this hypothesis, we examined the concordance between the two types of data. There are 192 compounds in the training dataset with qHTS assay data that were excluded from the external validation dataset. Of these 192 compounds, 186 are binders and only 6 are non-binders according to traditional assays, while the qHTS assay results show 42 binders and 150 non-binders. Comparison revealed that 41 binders and five non-binders are common in the two methods. Hence, most binders from the qHTS assay (41 out of 42) can be predicted using traditional assay results, whereas only a few non-binders (5 out of 150) can be predicted. This low concordance between the qHTS assay data and traditional assay data confirms our hypothesis.

Various consensus strategies can be employed to combine multiple individual models into a unified consensus model. In this study, consensus models were generated using a majority voting strategy based on the predictions of five individual models. To further investigate the potential impact of different consensus strategies, we also applied an average prediction probability approach. In this method, the mean of the MOR binder prediction probabilities from the five individual models was calculated and used as the binder probability for the consensus model. If the consensus probability of a compound was greater than or equal to 0.5, it was predicted as a binder; otherwise, it was classified as a non-binder. The consensus models derived from both the majority voting and average prediction probability strategies exhibited similar performance, as shown in Supplementary Figure S17.

A machine learning and deep learning model not only predicts the classes of a sample but also quantifies the likelihood of the sample belonging to the predicted class. In this study, the models output a probability indicating the likelihood of a compound being a binder. This probability is used not only to predict the compound as a MOR binder or non-binder, but also to measure the confidence of the prediction. To evaluate the usefulness of this probability, prediction confidence analysis was conducted on predictions in both cross-validations and external validations. The results (Figure 4) suggest that prediction confidence derived from the developed models offers an additional valuable metric for their applications.

The interpretability of deep learning models remains a key challenge, particularly in complex domains such as predicting binding activity for MOR. While the two deep learning models achieved higher predictive performance than the three machine learning models, the black-box nature makes deep learning models difficult to directly understand how individual molecular descriptors influence MOR binding. To enhance interpretability, techniques such as feature importance analysis, SHAP (Shapley Additive Explanations), and LIME (Local Interpretable Model-agnostic Explanations) can be used to identify the most influential molecular descriptors. However, achieving a fully transparent understanding of deep learning model behavior remains an ongoing research challenge. In the context of this study, we focus on predictive accuracy but acknowledge the need for further work on improving model interpretability for better insight into the underlying mechanisms of MOR binding.

AD serves as a critical metric for evaluating the uncertainty of predictions from machine learning or deep learning models. Compounds within the chemical space of the training compounds, or within the AD of a model, are expected to be predicted more accurately than those outside the AD [60, 61]. Our AD analysis of the predictions in the 5-fold cross-validations (Figure 5A) and external validations (Figure 5B) revealed that predictions within the AD are more accurate than those outside the AD for all models. Therefore, developing a model based on a training dataset with a broader chemical space can improve its applicability to a wider range of chemicals.

The scope of application of a predictive model is determined by its training dataset. The models developed in this study are based on a large dataset derived from traditional low-throughput experiments. In these models, any compound that exhibits binding activity—regardless of how weak—is classified as a MOR binder. As a result, the cross-validation results reflect the accuracy of predictions for chemicals in these binding assays, while the external validation results assess the models' ability to generalize and predict the binding activity of compounds in qHTS assays. To improve the reliability of predictions for MOR binding activity in either traditional or qHTS assays, stratified sampling should be employed to generate training and testing datasets. To demonstrate this, we applied this strategy. The details of the process are provided in the Supplementary Material, and the results are summarized in Supplementary Figures S18–S19.

In conclusion, machine learning (RF, kNN, and SVM) and deep learning (MLP and LSTM) models were constructed for MOR binding activity prediction. These models were evaluated using 5-fold cross-validations and external validations. The models achieved good performance in both evaluation methods. Results from prediction confidence analysis and AD analysis demonstrated the importance of prediction confidence and AD in evaluating the reliability of the models' predictions. Our findings suggest that the developed models have the potential to identify MOR binders, which could assist in the development of non-addictive or less-addictive drugs targeting MOR.

Author contributions

JiL, JeL, WGu, and WGe developed and evaluated the model. ZL and FD curated and processed data. JL wrote the first draft of

the manuscript. HH and TP revised the manuscript and generated the final version. All authors contributed to the article and approved the submitted version.

Author disclaimer

This article reflects the views of the authors and does not necessarily reflect those of the U.S. Food and Drug Administration.

Data availability

Publicly available datasets were analyzed in this study. This data can be found in the article/Supplementary Material.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This research was supported in part by an appointment to the Research Participation Program at the National Center for Toxicological Research (JeL and ZL), administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the U.S. Food and Drug Administration. This research was also supported in part by the Translational Science Interagency Fellowship program that is jointly sponsored by the National Center for Advancing Translational Sciences, National Institutes of Health and the U.S. Food and Drug Administration (ZL).

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.ebm-journal.org/articles/10.3389/ebm.2025.10359/full#supplementary-material

References

1. Compton P. The United States opioid crisis: big pharma alone is not to blame. *Prev Med* (2023) **177**:107777. doi:10.1016/j.ypmed.2023.107777

2. Skolnick P. Treatment of overdose in the synthetic opioid era. *Pharmacol and Ther* (2022) **233**:108019. doi:10.1016/j.pharmthera.2021.108019

3. Hebert AH, Hill AL. Impact of opioid overdoses on US life expectancy and years of life lost, by demographic group and stimulant co-involvement: a mortality data analysis from 2019 to 2022. *The Lancet Reg Health - Americas* (2024) **36**: 100813. doi:10.1016/j.lana.2024.100813

4. Ahmad FB, Cisewski JA, Rossen LM, Sutton P. *Provisional drug overdose death counts*. National Center for Health Statistics (2024). Available online at: https://www.cdc.gov/nchs/nvss/vsrr/drug-overdose-data.htm (Accessed January 10, 2024).

5. Spencer M, Minino A, Warner M. *Drug overdose deaths in the United States, 2001–2021*. NCHS Data Brief No. 457 (2022). Available online at: https://www.cdc.gov/nchs/products/databriefs/db457.htm.

6. Ciccarone D. The rise of illicit fentanyls, stimulants and the fourth wave of the opioid overdose crisis. *Curr Opin Psychiatry* (2021) **34**:344–50. doi:10.1097/yco.0000000000000717

7. Lee YK, Gold MS, Blum K, Thanos PK, Hanna C, Fuehrlein BS. Opioid use disorder: current trends and potential treatments. *Front Public Health* (2023) **11**: 1274719. doi:10.3389/fpubh.2023.1274719

8. Taylor JL, Samet JH. Opioid use disorder. *Ann Intern Med* (2022) **175**: ITC1–ITC16. doi:10.7326/aitc202201180

9. Florence C, Luo F, Rice K. The economic burden of opioid use disorder and fatal opioid overdose in the United States, 2017. *Drug and Alcohol Dependence* (2021) **218**:108350. doi:10.1016/j.drugalcdep.2020.108350

10. Dowell D, Ragan KR, Jones CM, Baldwin GT, Chou R. Prescribing opioids for pain - the new CDC clinical practice guideline. *N Engl J Med* (2022) **387**:2011–3. doi:10.1056/nejmp2211040

11. Kampman K, Jarvis M. American society of addiction medicine (ASAM) national practice guideline for the use of medications in the treatment of addiction involving opioid use. *J Addict Med* (2015) **9**:358–67. doi:10.1097/adm.0000000000000166

12. Judd D, King CR, Galke C. The opioid epidemic: a review of the contributing factors, negative consequences, and best practices. *Cureus* (2023) **15**:e41621. doi:10.7759/cureus.41621

13. Hoffman KA, Ponce Terashima J, McCarty D. Opioid use disorder and treatment: challenges and opportunities. *BMC Health Serv Res* (2019) **19**:884. doi:10.1186/s12913-019-4751-4

14. Ward MK, Guille C, Jafry A, Gwanzura T, Pryce K, Lewis P, et al. Digital health interventions to support women with opioid use disorder: a scoping review. *Drug and Alcohol Dependence* (2024) **261**:111352. doi:10.1016/j.drugalcdep.2024.111352

15. Yeo Y, Johnson R, Heng C. The public health approach to the worsening opioid crisis in the United States calls for harm reduction strategies to mitigate the harm from opioid addiction and overdose deaths. *Mil Med* (2022) **187**:244–7. doi:10.1093/milmed/usab485

16. *National opioids crisis: help and resources*. US Department of Health and Human Services (2025). Available online at: https://www.hhs.gov/opioids/index.html (Accessed January 12, 2024).

17. *Rx awareness campaign*. US Centers for Disease Control and Prevention (2025). Available online at: https://www.cdc.gov/rx-awareness/index.html (Accessed January 12, 2024).

18. *Fact sheet: addressing addiction and the overdose epidemic* (2024). Available online at: https://www.whitehouse.gov/briefing-room/statements-releases/2022/03/01/fact-sheet-addressing-addiction-and-the-overdose-epidemic/ (Accessed January 12, 2024).

19. Alogaili F, Abdul Ghani N, Ahmad Kharman Shah N. Prescription drug monitoring programs in the US: a systematic literature review on its strength and weakness. *J Infect Public Health* (2020) **13**:1456–61. doi:10.1016/j.jiph.2020.06.035

20. McLawhorn JM, Stephany MP, Bruhn WE, Crow LD, Coldiron BM, Hruza GJ, et al. An expert panel consensus on opioid-prescribing guidelines for dermatologic procedures. *J Am Acad Dermatol* (2020) **82**:700–8. doi:10.1016/j.jaad.2019.09.080

21. Jia X, Ciallella HL, Russo DP, Zhao L, James MH, Zhu H. Construction of a virtual opioid bioprofile: a data-driven qsar modeling study to identify new analgesic opioids. *ACS Sustainable Chem and Eng* (2021) **9**:3909–19. doi:10.1021/acssuschemeng.0c09139

22. Li Z, Liu J, Dong F, Chang N, Huang R, Xia M, et al. Three-dimensional structural insights have revealed the distinct binding interactions of agonists, partial agonists, and antagonists with the μ opioid receptor. *Int J Mol Sci* (2023) **24**:7042. doi:10.3390/ijms24087042

23. Kozell LB, Eshleman AJ, Wolfrum KM, Swanson TL, Bloom SH, Benware S, et al. Pharmacologic characterization of substituted nitazenes at μ, κ, and Δ opioid receptors suggests high potential for toxicity. *The J Pharmacol Exp Ther* (2024) **389**: 219–28. doi:10.1124/jpet.123.002052

24. Pasternak GW, Pan YX. Mu opioids and their receptors: evolution of a concept. *Pharmacol Rev* (2013) **65**:1257–317. doi:10.1124/pr.112.007138

25. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: new estimates of R&D costs. *J Health Econ* (2016) **47**:20–33. doi:10.1016/j.jhealeco.2016.01.012

26. Simoens S, Huys I. R&D costs of new medicines: a landscape analysis. *Front Med (Lausanne)* (2021) **8**:760762. doi:10.3389/fmed.2021.760762

27. Schlander M, Hernandez-Villafuerte K, Cheng CY, Mestre-Ferrandiz J, Baumann M. How much does it cost to research and develop a new drug? A systematic review and assessment. *Pharmacoeconomics* (2021) **39**:1243–69. doi:10.1007/s40273-021-01065-y

28. Qi X, Zhao Y, Qi Z, Hou S, Chen J. Machine learning empowering drug discovery: applications, opportunities and challenges. *Molecules* (2024) **29**:903. doi:10.3390/molecules29040903

29. Hong H, Tong W, Xie Q, Fang H, Perkins R. An *in silico* ensemble method for lead discovery: decision forest. *SAR QSAR Environ Res* (2005) **16**:339–47. doi:10.1080/10659360500203022

30. Xie Q, Ratnasinghe LD, Hong H, Perkins R, Tang ZZ, Hu N, et al. Decision forest analysis of 61 single nucleotide polymorphisms in a case-control study of esophageal cancer; a novel method. *BMC Bioinformatics* (2005) **6**(Suppl. 2):S4. doi:10.1186/1471-2105-6-s2-s4

31. Tan H, Wang X, Hong H, Benfenati E, Giesy JP, Gini GC, et al. Structures of endocrine-disrupting chemicals determine binding to and activation of the estrogen receptor α and androgen receptor. *Environ Sci Technol* (2020) **54**:11424–33. doi:10.1021/acs.est.0c02639

32. Idakwo G, Thangapandian S, Luttrell J, Li Y, Wang N, Zhou Z, et al. Structure-activity relationship-based chemical classification of highly imbalanced Tox21 datasets. *J Cheminform* (2020) **12**:66. doi:10.1186/s13321-020-00468-x

33. Guo W, Liu J, Dong F, Hong H. Unlocking the potential of AI: machine learning and deep learning models for predicting carcinogenicity of chemicals. *J Environ Sci Health C* (2024) **43**:23–50. doi:10.1080/26896583.2024.2396731

34. Dong F, Guo W, Liu J, Patterson TA, Hong H. BERT-based language model for accurate drug adverse event extraction from social media: implementation, evaluation, and contributions to pharmacovigilance practices. *Front Public Health* (2024) **12**:1392180. doi:10.3389/fpubh.2024.1392180

35. Liu J, Khan MKH, Guo W, Dong F, Ge W, Zhang C, et al. Machine learning and deep learning approaches for enhanced prediction of hERG blockade: a comprehensive QSAR modeling study. *Expert Opin Drug Metab and Toxicol* (2024) **20**:665–84. doi:10.1080/17425255.2024.2377593

36. Guo W, Liu J, Dong F, Song M, Li Z, Khan MKH, et al. Review of machine learning and deep learning models for toxicity prediction. *Exp Biol Med (Maywood)* (2023) **248**:1952–73. doi:10.1177/15353702231209421

37. Floresta G, Rescifina A, Abbate V. Structure-based approach for the prediction of mu-opioid binding affinity of unclassified designer fentanyl-like molecules. *Int J Mol Sci* (2019) **20**:2311. doi:10.3390/ijms20092311

38. Sakamuru S, Zhao J, Xia M, Hong H, Simeonov A, Vaisman I, et al. Predictive models to identify small molecule activators and inhibitors of opioid receptors. *J Chem Inf Model* (2021) **61**:2675–85. doi:10.1021/acs.jcim.1c00439

39. Pan C, Meng H, Zhang S, Zuo Z, Shen Y, Wang L, et al. Homology modeling and 3D-QSAR study of benzhydrylpiperazine delta opioid receptor agonists. *Comput Biol Chem* (2019) **83**:107109. doi:10.1016/j.compbiolchem.2019.107109

40. Feng H, Elladki R, Jiang J, Wei GW. Machine-learning analysis of opioid use disorder informed by MOR, DOR, KOR, NOR and ZOR-based interactome networks. *Comput Biol Med* (2023) **157**:106745. doi:10.1016/j.compbiomed.2023.106745

41. Provasi D, Filizola M. Enhancing opioid bioactivity predictions through integration of ligand-based and structure-based drug discovery strategies with transfer and deep learning techniques. *J Phys Chem B* (2023) **127**:10691–9. doi:10.1021/acs.jpcb.3c05306

42. Oh M, Shen M, Liu R, Stavitskaya L, Shen J. Machine learned classification of ligand intrinsic activities at human mu-opioid receptor. *ACS Chem Neurosci* (2024) **15**:2842–52. doi:10.1021/acschemneuro.4c00212

43. Hong H, Thakkar S, Chen M, Tong W. Development of decision forest models for prediction of drug-induced liver injury in humans using A large set of FDA-approved drugs. *Sci Rep* (2017) **7**:17311. doi:10.1038/s41598-017-17701-7

44. Hong H, Xie Q, Ge W, Qian F, Fang H, Shi L, et al. Mold(2), molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. *J Chem Inf Model* (2008) **48**:1337–44. doi:10.1021/ci800038f

45. Breiman L. Random forests. *Mach Learn* (2001) **45**:5–32. doi:10.1023/A:1010933404324

46. Cover TM, Hart PE. Nearest neighbor pattern classification. *IEEE Trans Inf Theor* (1967) **13**:21–7. doi:10.1109/tit.1967.1053964

47. Cortes C, Vapnik V. Support-vector networks. *Mach Learn* (1995) **20**:273–97. doi:10.1007/bf00994018

48. Rosenblatt F. The perceptron - a probabilistic model for information-storage and organization in the brain. *Psychol Rev* (1958) **65**:386–408. doi:10.1037/h0042519

49. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* (1997) **9**:1735–80. doi:10.1162/neco.1997.9.8.1735

50. *Online SMILES translator and structure file generator*. NIH/National Cancer Institue (2025). Available online at: https://cactus.nci.nih.gov/translate/ (Accessed August 9, 2023).

51. Wigh DS, Goodman JM, Lapkin AA. A review of molecular representation in the age of machine learning. *Wires Comput Mol Sci* (2022) **12**:e1603. doi:10.1002/wcms.1603

52. Godden JW, Stahura FL, Bajorath J. Variability of molecular descriptors in compound databases revealed by Shannon entropy calculations. *J Chem Inf Comput Sci* (2000) **40**:796–800. doi:10.1021/ci000321u

53. Liu J, Guo W, Dong F, Aungst J, Fitzpatrick S, Patterson TA, et al. Machine learning models for rat multigeneration reproductive toxicity prediction. *Front Pharmacol* (2022) **13**:1018226. doi:10.3389/fphar.2022.1018226

54. Liu J, Xu L, Guo W, Li Z, Khan MKH, Ge W, et al. Developing a SARS-CoV-2 main protease binding prediction random forest model for drug repurposing for COVID-19 treatment. *Exp Biol Med (Maywood)* (2023) **248**:1927–36. doi:10.1177/15353702231209413

55. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res* (2011) **12**:2825–30.

56. *Python* (2025) Available online at: https://www.python.org/downloads/release/python-385/ (Accessed August 16, 2023).

57. *PyTorch* (2025) Available online at: https://pytorch.org/get-started/pytorch-2.0/ (Accessed August 16, 2023).

58. Hong H, Rua D, Sakkiah S, Selvaraj C, Ge W, Tong W. Consensus modeling for prediction of estrogenic activity of ingredients commonly used in sunscreen products. *Int J Environ Res Public Health* (2016) **13**:958. doi:10.3390/ijerph13100958

59. Weaver S, Gleeson MP. The importance of the domain of applicability in QSAR modeling. *J Mol Graphics Model* (2008) **26**:1315–26. doi:10.1016/j.jmgm.2008.01.002

60. Klingspohn W, Mathea M, Ter Laak A, Heinrich N, Baumann K. Efficiency of different measures for defining the applicability domain of classification models. *J Cheminform* (2017) **9**:44. doi:10.1186/s13321-017-0230-2

61. Kar S, Roy K, Leszczynski J. Applicability domain: a step toward confident predictions and decidability for qsar modeling. *Methods Mol Biol* (2018) **1800**:141–69. doi:10.1007/978-1-4939-7899-1_6

EBM Experimental Biology and Medicine

TYPE Original Research
PUBLISHED 02 May 2025
DOI 10.3389/ebm.2025.10374

OPEN ACCESS

*CORRESPONDENCE
Huixiao Hong,
huixiao.hong@fda.hhs.gov

RECEIVED 09 September 2024
ACCEPTED 22 April 2025
PUBLISHED 02 May 2025

CITATION
Guo W, Dong F, Liu J, Aslam A, Patterson TA and Hong H (2025) A refined set of RxNorm drug names for enhancing unstructured data analysis in drug safety surveillance.
Exp. Biol. Med. 250:10374.
doi: 10.3389/ebm.2025.10374

A refined set of RxNorm drug names for enhancing unstructured data analysis in drug safety surveillance

Wenjing Guo, Fan Dong, Jie Liu, Aasma Aslam, Tucker A. Patterson and Huixiao Hong*

National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, AR, United States

Abstract

Adverse drug events are harms associated with drug use, whether the drug is used correctly or incorrectly. Identifying adverse drug events is vital in pharmacovigilance to safeguard public health. Drug safety surveillance can be performed using unstructured data. A comprehensive and accurate list of drug names is essential for effective identification of adverse drug events. While there are numerous sources for drug names, RxNorm is widely recognized as a leading resource. However, its effectiveness for unstructured data analysis in drug safety surveillance has not been thoroughly assessed. To address this, we evaluated the drug names in RxNorm for their suitability in unstructured data analysis and developed a refined set of drug names. Initially, we removed duplicates, the names exceeding 199 characters, and those that only describe administrative details. Drug names with four or fewer characters were analyzed using 18,000 drug-related PubMed abstracts to remove names which rarely appear in unstructured data. The remaining names, which ranged from five to 199 characters, were further refined to exclude those that could lead to inaccurate drug counts in unstructured data analysis. We compared the efficiency and accuracy of the refined set with the original RxNorm set by testing both on the 18,000 drug-related PubMed abstracts. The results showed a decrease in both computational cost and the number of false drug names identified. Further analysis of the removed names revealed that most originated from only one of the 14 sources. Our findings suggest that the refined set can enhance drug identification in unstructured data analysis, thereby improving pharmacovigilance.

KEYWORDS

adverse drug events, pharmacovigilance, natural language processing, database, DrugBank

Impact statement

Adverse drug events are a significant concern for public health, necessitating accurate detection in drug safety surveillance. While unstructured data is a valuable source for identifying adverse drug events, effective analysis depends on a comprehensive and accurate list of drug names. Although RxNorm is recognized for providing standardized drug names, its effectiveness in unstructured data analysis remains unassessed. Our research refined the list of RxNorm drug names to improve its suitability for unstructured data analysis. By removing duplicates, excessively long names, false names, and replaceable names, we created a more accurate and efficient list of drug names. Testing this refined set on drug-related PubMed abstracts revealed improved accuracy and reduced computational costs compared to the original RxNorm list. This refined list of drug names enables more accurate monitoring of adverse drug events, providing a valuable tool for improving drug safety surveillance and protecting public health.

Introduction

Adverse drug events (ADEs) are harmful responses to medications that pose significant risks to patients with millions of deaths and hospitalization annually [1]. Effective monitoring of ADEs through drug safety surveillance is crucial for protecting public health. Drug safety surveillance begins in clinical trials, where new drugs are rigorously tested for safety and efficacy. However, clinical trials are limited by short exposure periods and the size and diversity of the tested population [2]. Therefore, post-market drug safety surveillance is crucial to identify potential ADEs in a large population, particularly for drugs repurposed to treat COVID-19. For example, originally developed for the treatment of hepatitis C, Remdesivir was later evaluated for antiviral activity against other viruses and, in 2020, received FDA approval for the treatment of COVID-19. Traditionally, post-market surveillance relies on spontaneous adverse event reporting systems [3, 4]. In the United States, the Food and Drug Administration's Adverse Event Reporting System (FAERS) [5] collects adverse event reports, medication error reports, and product quality complaints from various sources, including the MedWatch program. FAERS has been widely used to investigate drug safety issues [6–9]. However, FAERS relies on voluntary reporting, which can result in underreporting and delays in identifying ADEs. In recent years, unstructured text data has become valuable sources for investigating ADEs.

To effectively analyze unstructured data for drug safety surveillance, it is important to identify drugs and associated ADEs. One challenge for identifying drugs in unstructured data is different names used for the same drugs. The active ingredient, generic names, trade names, brand names, and even street names can be used to indicate the same drug in unstructured text. Using acetaminophen, a commonly used analgesic, as an example, Tylenol, Paracetamol, Panadol, Anacin, Feverall, Mapap, Ofirmev, Tempra, and APAP (the abbreviation for its chemical name, N-acetyl-para-aminophenol) are names used for the same drug in unstructured documents. The use of various names for the same drugs in unstructured data complicates accurate identification of drugs, making the standardization and normalization of drug names essential.

Various methods have been used in the standardization and normalization of drug names, including dictionary-based methods [10], rule-based systems [11–16], advanced machine learning models [17–20], and hybrid approaches [19]. Dictionary-based methods use comprehensive drug dictionaries built from various sources to identify drug names [10]. In these methods, a comprehensive dictionary like RxNorm is essential to ensure accurate recognition of complex or less common drug names [21].

Rule-based systems, on the other hand, rely on predefined patterns or contextual rules to identify drug names. These rules can be either composition-based, focusing on systematic naming conventions, or context-based, extracting names based on surrounding text features [22, 23]. Despite the rigidity and extensive manual effort required to develop and maintain these rules and dictionaries—especially given the evolving nature of language and the introduction of new terminology—both dictionary and rule-based methods remain crucial for establishing a baseline of accurate drug identification.

To enhance the matching and normalization processes, similarity algorithms such as Levenshtein distance [24], cosine similarity [25], and Jaccard index [25] can be used. These techniques measure the similarity between drug names and help link various names of the same drug to a standard drug name [26, 27], further improving the accuracy of drug name standardization.

With the increasing availability of annotated datasets, machine learning-based models have gained significant popularity in this field [10, 17–20, 28]. Notable techniques such as Conditional Random Forest (CRF) [29], Hidden Markov Models (HMM), Recurrent Neural Networks (RNN) [30], and Bi-directional Long Short-Term Memory CRF (BI-LSTM-CRF) [31–33], and Bidirectional Encoder Representations from Transformers (BERT) [15] have been employed for drug name identification and normalization. These models leverage various features, including domain-specific attributes and word representation features, to improve accuracy.

Hybrid approaches have also emerged, integrating multiple methods to capitalize on the strengths of different models while mitigating their weaknesses [19]. For example, a semi-supervised machine learning technique known as feature coupling generalization was applied to refine a drug name dictionary, which was constructed from sources such as DrugBank and

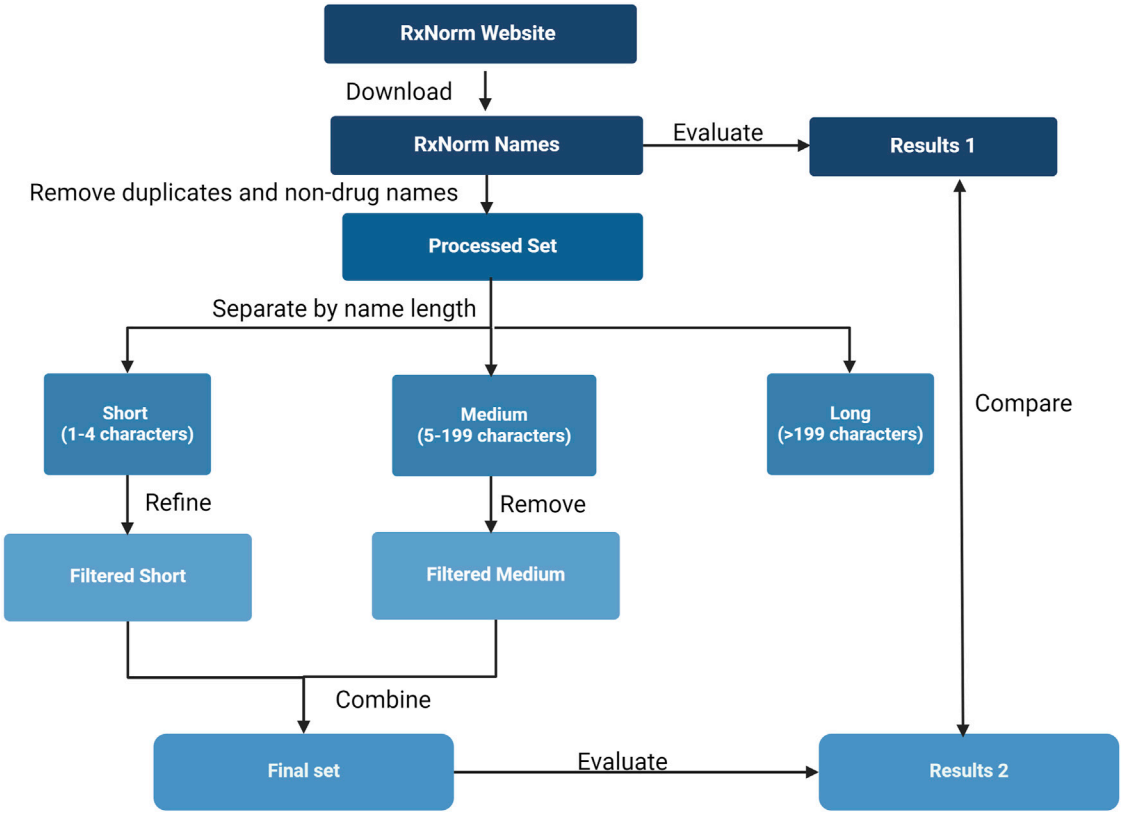


FIGURE 1
Study overview. The flowchart illustrates the procedures used to generate and evaluate a refined set of drug names from RxNorm, including extraction of drug names from the RxNorm website, removal of duplicates, filtering false names, discarding names that likely lead to redundant occurrence counts in unstructured data analysis, and evaluating accuracy and efficiency of the refined set.

PubMed, to enhance drug name recognition in unstructured textual data [19].

To create a drug name dictionary, different names for the same drug are linked to a standardized name. A comprehensive dictionary is essential for accurate drug identification and normalization. RxNorm [34], a standardized vocabulary developed by the National Library of Medicine (NLM), plays a key role in these processes. RxNorm compiles drug names from 13 different sources and further standardizes them under its own unique terminology, RxNorm, bringing the total to 14 distinct sources, enabling consistent linkage of various drug names across different databases. The integration of RxNorm with both rule-based and machine learning approaches enhances the identification and normalization of drug names.

Although RxNorm is widely used in clinical settings, such as electronic health records and clinical decision support systems [35], it faces several limitations when analyzing unstructured data. One significant issue is the extensive variability in the length of drug names within RxNorm, which can range from one to over 2000 characters. These extremely short or long names are seldom found in unstructured text. Moreover, RxNorm includes distinct entries, various drug formats, and dosages, which are typically omitted when discussing experience with drugs in unstructured text. Even when such details are mentioned, they are often inconsistent and incomplete.

Additionally, RxNorm's approach of combining drug names with specific dosages as separate entries can lead to multiple hits for the same drug in a single text. For example, "Acetaminophen" and "Acetaminophen 325 mg" are distinct entries in RxNorm. If both terms are included in a drug name dictionary, a sentence like "Acetaminophen 325 mg caused my mom's liver injury" could lead to two matches—one for "Acetaminophen" and another for "Acetaminophen 325 mg"— resulting in redundant counts of the adverse event. These complexities stress the need for a refined set of drug names to improve the accuracy and efficiency of drug identification in unstructured data.

The purpose of this study is to develop an enhanced set of drug names from RxNorm, specifically tailored for identifying drug names in unstructured data for drug safety surveillance. By refining the existing drug names in RxNorm, this study aims to address current limitations and

improve the accuracy and efficiency of drug identification in unstructured data.

Materials and methods

Study design

The workflow for generating this refined set and assessing its accuracy and efficiency is depicted in Figure 1. Initially, a comprehensive list of drug names was downloaded from the RxNorm database. This was followed by a systematic process of removing duplicates, incorrect names, and names that could potentially cause inaccurate counts in unstructured data analysis. Drug names were classified into three categories and filtered out by those with fewer than 4 characters, those with between 5 and 199 characters, and those with 200 or more characters.

Data sources

RxNorm file released on July 3, 2023 (RxNorm_full_07032023.zip) was downloaded from RxNorm repository [36]. The "RXNCONSO.RRF" file within this package was used to extract drug names. Specifically, drug names were obtained from the "STR" (string) column, while their corresponding types were identified from the "TTY" (type of terms) column, which includes categories such as brand name, synonyms, and others.

To ensure relevance, name types not associated with specific drugs were excluded based on the guidelines provided in the RxNorm technical documentation [37]. For instance, terms like dose form, dose form group, and special category—which describe routes of administration rather than specific drugs—were removed. The source of each drug name is indicated in the "SAB" (source abbreviation) column: ATC (Anatomical Therapeutic Chemical Classification System), CVX (Vaccines Administered), DB (DrugBank), GS (Gold Standard Drug Database), MMSL (Micromedex RED BOOK), MMX (Micromedex), MSH (Medical Subject Headings), MTHCMS (CMS Formulary Reference File), MTHSPL (FDA Structured Product Labeling), NDDF (First Databank), RXNORM (RxNorm itself), SNOMED (SNOMED Clinical Terms), USP (United States Pharmacopeia), and VANDF (Veterans Health Administration National Drug File).

To evaluate the extracted drug names, a dataset of 18,000 drug-related PubMed abstracts was prepared. These abstracts were retrieved by searching PubMed using the keyword "drug" via the Entrez Programming Utilities [38] (E-Utlilities) developed by the National Center for Biotechnology Information (NCBI). To comply with NCBI guidelines, we designated an email address for Entrez queries. On 22 May 2024, we generated a search query using the keyword "drug" without imposing any timeframe restrictions, ensuring the retrieval of all available abstracts up to that date. Entrez was used to retrieve 20,000 PubMed abstract IDs matching this query. Due to the limitation on the number of abstracts that can be fetched in a single request, we retrieved the IDs in two batches, with each batch containing 10,000 IDs. Abstracts were fetched and output for each batch. Although 20,000 IDs were obtained, 18,520 abstracts were successfully retrieved due to some missing entries. Ultimately, we used the first 18,000 abstracts, choosing this round number to simplify subsequent calculations.

Refinement of RxNorm drug names

The first step is to remove duplicates and exclude drug names that are not associated with specific drugs. This includes eliminating terms that describe dose form, dose form group, and special category—such as "oral tablet," "chewable product," and "medical supplies"—since these are not linked to particular drugs and should, therefore, be excluded. Brand and generic drug names were retained to ensure comprehensive drug identification. For example, both Daytrana (patch) and Ritalin (oral tablet) were included as brand names for methylphenidate. This approach ensures that drug identification focuses on the medication itself while preventing redundant counts based on formulation differences. However, we recognize that ADEs can sometimes be associated with the delivery method rather than the active ingredient. For instance, systemic methylphenidate may be linked to behavioral effects like aggression, while transdermal formulations such as Daytrana may cause localized reactions like rash.

For drug names with four or fewer characters such as APAP (Acetaminophen), ASA (Aspirin), and HCTZ (Hydrochlorothiazide), their use frequency in unstructured data were tested in 18,000 drug-related PubMed abstracts to remove those that would rarely appear in drug-related documents. Drug names that were not found in these abstracts were considered rare and removed. We used the "en_core_web_sm" model from the spaCy [39] natural language processing (NLP) library to identify and count occurrences of these drug names within the abstracts. Each abstract was tokenized, and both tokens and drug names were converted to lowercase for consistency. We then compared each token against the list of drug names, recording an occurrence whenever a match was found. Drug names with zero occurrences were excluded from the final list.

For drug names with five to 199 characters, we examined their potential redundant occurrences in unstructured data analysis. If a drug name contains another drug name, leading to redundant counts, it was discarded. To identify distinct drug names that overlap with discarded names but not with other distinct names, we split each drug name into words using the Python's "re.split" function (version 3.11.7 in Anaconda). The names were then sorted by word count. We checked if the words

TABLE 1 Summary of removed words for each drug name type.

Name type	Percentage of removed words
Duplicates	23.61
Non-drug Names	0.09
Drug Names with less than 5 characters	0.06
Drug Names with 5-199 characters	65.78
Drug Names with >200 characters	1.53

in a drug name contained all words of another name. If a drug name that does contain all the words of any other names, it was removed. Drug names with 199 or more characters were removed entirely, as they are unlikely to appear in real-world unstructured texts.

Assessment of the refined set

To evaluate the efficiency and accuracy of the refined set of drug names in unstructured data analysis, we conducted drug identification on the 18,000 drug-related PubMed abstracts. The refined and original drug names were converted to lowercase and tokenized using the "en_core_web_sm" in spaCy. These tokenized drug names were used to create matching patterns, which were added to spaCy's PhraseMatcher. Each abstract was tokenized, and the PhraseMatcher compared each sequence of tokens against the created matching patterns. When a match was found, the drug name was recorded.

Efficiency was measured by comparing the computational time required for both the refined and original RxNorm drug name sets. Accuracy was calculated as the ratio of drug names identified within the abstracts to the total number of drug names, for both the refined and original sets.

Results

Refinement of drug names

Table 1 provides a summary of the percentages of words removed at each stage of the refinement process, offering a clearer overview of the impact of our filtering criteria.

Download and processing of drug names

To refine the drug names in RxNorm, we downloaded the RxNorm file released on July 3, 2023, from the RxNorm website [40]. The "RXNCONSO" file in the downloaded zipped files was used to obtain drug names and other related information, with drug names stored in the "STR" column. A total of 1,143,201 drug names were retrieved from which 269,931 duplicates were identified and removed. Then, we examined the types of the retained drug names to remove those not containing specific drug information. According to the RxNorm technical documentation [41], three term types (DF, DFG, SC) pertain to administrative details rather than specific drugs. We removed 1,009 drug names belonging to these categories.

Drug names with four or fewer characters

We used 18,000 drug-related PubMed abstracts to evaluate the occurrence of drug names with four or fewer characters. Out of 1260 drug names, 687 had zero occurrences and were discarded. The occurrences of the remaining drug names with the abstracts are provided in Supplementary Table S1.

We further analyzed the sources of the 687 discarded names. Our analysis showed that the majority originated from a single source among the 14 in RxNorm, indicating that drug names from a single source are unlikely to appear in unstructured drug-related texts. This result is not surprising, as these names lack corroboration from other sources. We also examined the source distribution of these 557 names. As shown in Figure 2A, DrugBank had the highest number (289), followed by SNOMEDCT_US (84) and MSH (84). In total, DrugBank, SNOMEDCT_US, and MSH, contained 628, 250, and 233 drug names with four or fewer characters, respectively. This indicates that approximately 46%, 34%, and 36% of such names from DrugBank, SNOMEDCT_US, and MSH were excluded. In contrast, sources like NDDF and MTHSPL had fewer names of this length and a lower removal rate, with only 1 out of 60 from NDDF and 6 out of 62 from MTHSPL being removed.

Drug names with five to 199 characters

For drug names with five to 199 characters, we excluded those that could lead to redundant occurrence counts in unstructured data analysis. For example, using both original drug names "Acetaminophen" and "Acetaminophen 325 MG Oral Tablet" to identify adverse events for drugs in the text "my brother had headache after take acetaminophen 325 MG tablet", might lead to two counts for the adverse event "headache" when only one should be recorded. Therefore, drug names that contain other names were removed, while distinct names without overlaps were retained. Out of 853,472 names with five to 199 characters, 101,491 are distinct names and were retained, whereas 751,981 names, which contain other names, were removed.

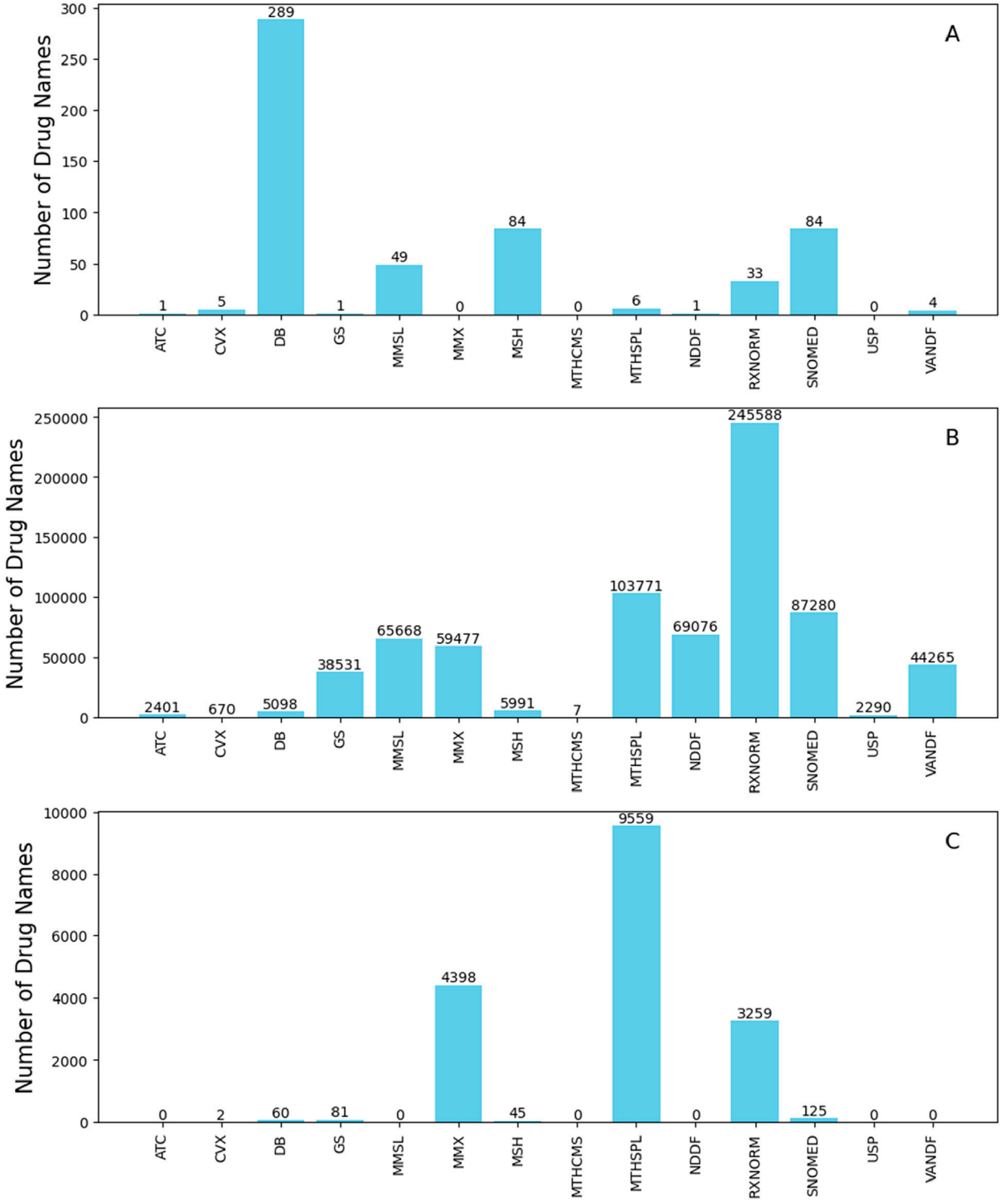


FIGURE 2
Source distribution of the removed drug names that only originate from a single source for names with four or fewer characters **(A)**, names with five to 199 characters **(B)**, and names with 200 or more characters **(C)**. The y-axes give number of names and x-axes depict name sources. Abbreviations: ATC (Anatomical Therapeutic Chemical Classification System), CVX (Vaccines Administered), DB (DrugBank), GS (Gold Standard Drug Database), MMSL (Micromedex RED BOOK), MMX (Micromedex), MSH (Medical Subject Headings), MTHCMS (CMS Formulary Reference File), MTHSPL (FDA Structured Product Labeling), NDDF (First Databank), RXNORM (RxNorm itself), SNOMED (SNOMED Clinical Terms), USP (United States Pharmacopeia), and VANDF (Veterans Health Administration National Drug File).

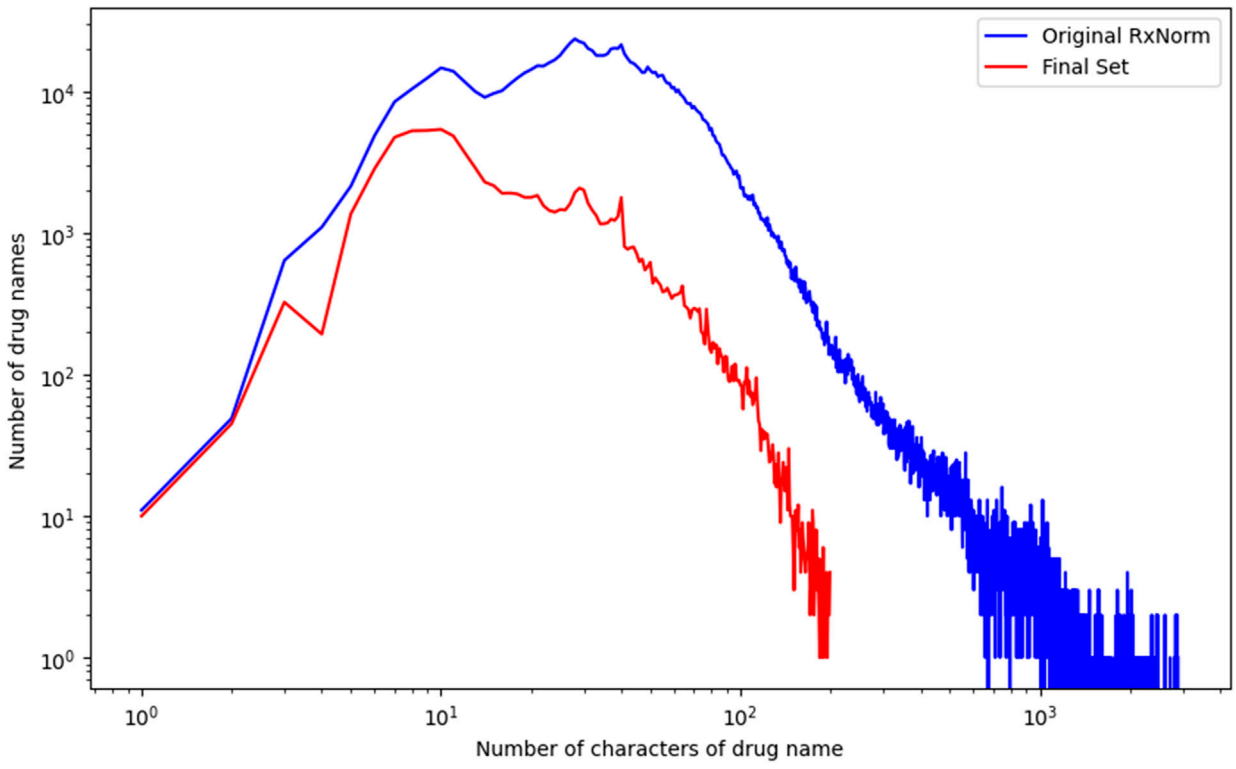


FIGURE 3
Comparison of name length between the refined set and the original RxNorm set. The y-axis shows the number of drug names, and the x-axis indicates name length. Name lengths were color coded in red for the refined sets and in blue for the original RxNorm set.

A significant portion of the removed names (730,113 out of 751,981) originate from only one of the 14 sources in RxNorm. The source distribution of these removed single-sourced names is shown in Figure 2B. Most of these drug names came from RxNorm, followed by MTHSPL, SNOMEDCT_US, NDDF, and MSSL. Specifically, RxNorm, MTHSPL, SNOMEDCT_US, NDDF, and MMSL provided 279,465, 121,035, 108,421, 99,054, and 91,270 drug names with five to 199 characters, respectively. The removal rates for these names are notably high: 87.8% for RxNorm, 85.7% for MTHSPL, 80.4% for SNOMEDCT_US, 71.9% for MMSL, and 69.7% for NDDF. In contrast, only 16.4% (5,098 out of 31,041) of the names with five to 199 characters from DrugBank were removed.

Drug names with 200 or more characters

Drug names with 200 or more characters are rarely used in unstructured data and, therefore, were excluded. A total of 17,529 such drug names were found in RxNorm and excluded. All these names originated from a single source, with the source distribution depicted in Figure 2C.

Evaluation of the refined drug names set

The refined set of drug names include 573 names with four or fewer characters and 101,491 names with five to 199 characters. We analyzed the distribution of drug name lengths between the refined set and the original RxNorm set. As shown in Figure 3, longer drug names were less likely to be retained in the refined set. This suggests that longer drug names are more prone to generating redundant occurrence counts in unstructured data analysis compared to shorter drug names and were thus discarded.

To evaluate the efficiency and accuracy of the refined set of drug names, we used 18,000 drug-related PubMed abstracts. Our results revealed that 3,065 names were identified in the abstracts, with lengths ranging from 1 to 46 characters. When we evaluated the original RxNorm set using the same abstracts, we found 4,471 names with lengths ranging from 1 to 66 characters. The additional 1,046 names that RxNorm identified in the abstracts were either false drug names or names likely leading to redundant occurrence counts in unstructured data analysis. These names were excluded from the refined set, with the majority originating from DrugBank and SNOMEDCT_US, as shown in Figure 4. Our results reveal that the refined set of drug names improved

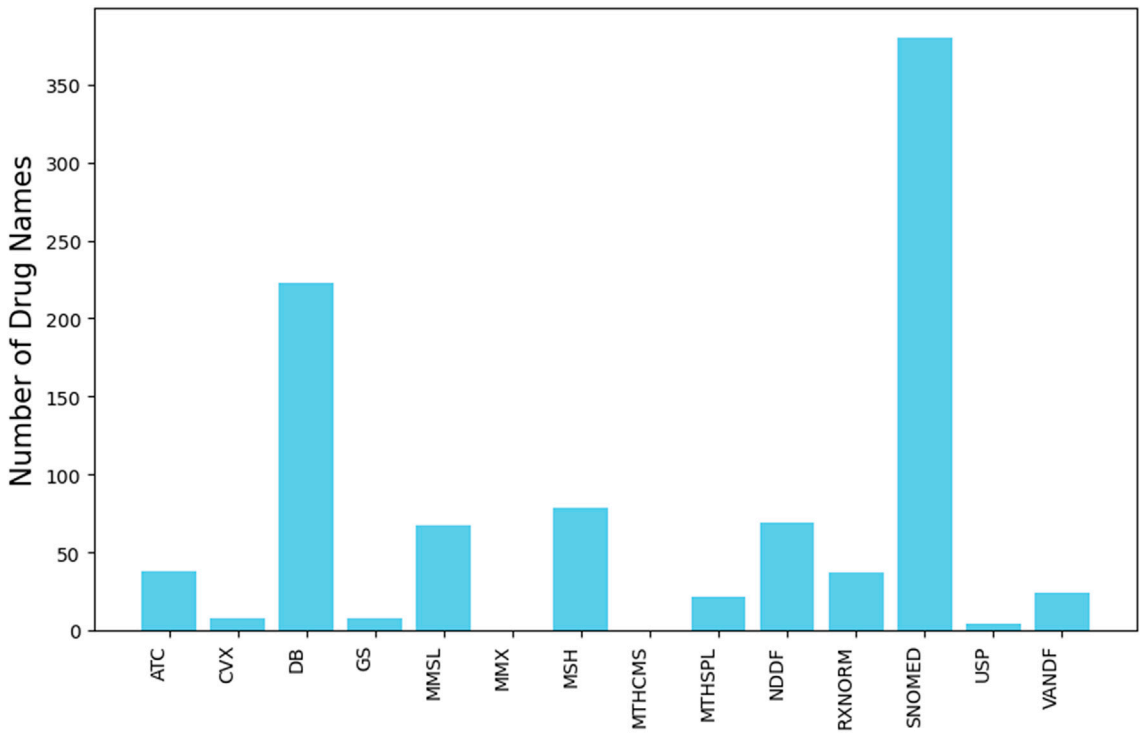


FIGURE 4
Source of original RxNorm drug names that were excluded from the refined set but identified in the PubMed abstracts. The y-axis represents number of drug names and the x-axis depicts sources. Abbreviations: ATC (Anatomical Therapeutic Chemical Classification System), CVX (Vaccines Administered), DB (DrugBank), GS (Gold Standard Drug Database), MMSL (Micromedex RED BOOK), MMX (Micromedex), MSH (Medical Subject Headings), MTHCMS (CMS Formulary Reference File), MTHSPL (FDA Structured Product Labeling), NDDF (First Databank), RXNORM (RxNorm itself), SNOMED (SNOMED Clinical Terms), USP (United States Pharmacopeia), and VANDF (Veterans Health Administration National Drug File).

drug identification accuracy in analyzing unstructured texts compared to the original RxNorm set.

The efficiency of the refined set of drug names was measured using the computational time required to analyze the abstracts. The analysis using the refined set took 1,910 s, while using the original RxNorm set took 6,301 seconds—over three times longer. Our results demonstrate a significant improvement in efficiency when analyzing unstructured data, making the refined set more suitable for real-time drug safety surveillance.

Discussion

Artificial intelligence is increasingly playing a critical role in evaluating drug safety and chemical toxicity. By harnessing machine learning algorithms and computational models, artificial intelligence can predict adverse effects, identify toxic compounds, and improve pharmacovigilance efforts. There are two main types of data involved: structured and unstructured. Due to their distinct formats and organization, machine learning techniques are applied differently to each. Structured data is well-organized and easily interpretable by machines, making it a natural fit for a wide range of safety assessments and toxicity endpoints [40–53]. In contrast, unstructured data lacks a predefined format, which makes it more challenging to process and analyze. To effectively apply machine learning techniques, such as natural language processing and recurrent neural networks, to unstructured data in pharmacovigilance, a reliable and comprehensive set of drug names is essential.

In this study, we generated a refined set of drug names from RxNorm to improve the accuracy and efficiency of drug identification in unstructured data. The original RxNorm set contained duplicates, non-specific drug names, and names that were either too long or too short, which hindered effective drug identification in unstructured data. Our objective was to exclude such names from analysis of unstructured texts. The refined set was evaluated using 18,000 drug-related PubMed abstracts, demonstrating enhanced accuracy and efficiency in drug identification, thereby potentially improving drug safety surveillance through unstructured data analysis.

Single-sourced drug names, originated from only one of the 14 sources in RxNorm, are generally less reliable than names corroborated by multiple sources. These single-sourced names tend to cause incorrect identification or generate redundant

occurrence counts when analyzing unstructured data, affecting both the accuracy and efficiency of drug identification. Our results revealed that the majority of the removed names were single-sourced, highlighting the importance of utilizing drug names validated by multiple sources.

Furthermore, most of the removed single-sourced names originated from FDA Structured Label, RxNorm, and SNOMEDCT_US. These sources serve distinct roles in drug information management. The FDA Structured Product Label provides comprehensive regulatory drug details, including dosage, formulation, and safety information, to ensure clarity and reduce medication errors. RxNorm standardizes drug names by linking ingredients, strengths, and dosage forms, facilitating interoperability across electronic health systems. SNOMED CT, on the other hand, is primarily used for clinical documentation and coding within electronic health records.

RxNorm integrates drug names from multiple external sources; however, not all names from contributing databases are necessarily included. Furthermore, many drug names appear in multiple sources within RxNorm, potentially leading to redundant listings. To mitigate this, our analysis systematically identified and removed duplicate drug names contributed by multiple sources, ensuring that each unique drug name was counted only once. While these structured resources are essential for clinical and regulatory use, their detailed naming conventions can complicate drug identification in unstructured data. Refining these names is crucial to enhance their applicability in text-based analyses.

On the other hand, sources like DrugBank and MSH showed varying levels of reliability across different lengths of drug names. For drug names with four or fewer characters, DrugBank had a relatively high removal rate of 46%, indicating that many of these names are unlikely to appear in unstructured data. However, the removal rate for DrugBank drug names with five to 199 characters significantly reduced to 16.4%, suggesting that these names are more reliable in unstructured data analysis. Similarly, MSH had a high removal rate of 36% for names with four or fewer characters and a lower rate of 24% for names with five to 199 characters. Our results suggest that more caution is needed when using short names from DrugBank and MSH in unstructured data analysis for drug safety surveillance compared to their longer names.

Despite the improvements in accuracy and efficiency demonstrated by the refined set, some limitations should be noted. First, our refined set of drug names is not error-free for unstructured data analysis, and some unsuitable names may persist. For example, short drug names in the refined set might include common words that, depending on the context, do not refer to drugs. Second, as RxNorm is primarily composed of professionally used names, it may not capture the variations found in street names or slang used in non-professional documents. Third, because RxNorm is updated monthly, regular updates are necessary to maintain the accuracy and relevance of the refined set. Finally, our evaluation was limited to 18,000 drug-related PubMed abstracts. Although we focused on abstracts containing the keyword "drug" to increase the likelihood of identifying drug names, these abstracts may not represent other unstructured real-world data. We selected the keyword "drug" to maximize the inclusion of abstracts that explicitly mention specific drug names. Alternative terms such as "medications" or "pharmacologic" were not used, as they are often associated with broader discussions on treatment strategies, pharmacological mechanisms, or drug classes rather than individual drug names. Additionally, a composite search incorporating all relevant MeSH terms was not conducted to ensure consistency with prior studies that employed keyword-based retrieval for drug-related text analysis. This approach maintains methodological alignment while optimizing the extraction of relevant drug name mentions.

Further efforts are needed to enhance the refined set. One such effort involves evaluating the set more comprehensively using diverse unstructured data. Additionally, the refined set could be improved by integrating advanced algorithms and machine learning techniques. Machine learning algorithms, particularly those involving similarity measurements, could be trained to recognize and link synonymous drug names, thereby improving accuracy. Natural language processing techniques like BERT could also be employed to better understand the context in which drug names appear, further enhancing accuracy. Finally, developing automated processes for updating the drug names in the dataset is crucial. As RxNorm updates its dataset monthly, maintaining the refined set through an automated update process will ensure its continued reliability for unstructured data mining in drug safety surveillance.

Conclusion

The development of the refined set of drug names from RxNorm has shown significant improvements in the accuracy and efficiency of drug identification in unstructured data. This refined dataset could be valuable for extracting drug-related information from unstructured data, thereby supporting more effective monitoring and management of drug safety through unstructured data analysis. Our study also highlights the importance of addressing the limitations of existing drug names when used for unstructured data mining, particularly in the context of drug safety surveillance.

Author contributions

WG and HH designed the work. WG, FD, JL, and AA conducted data analysis. WG and HH wrote the first draft. TP revised the manuscript. All authors contributed to the article and approved the submitted version.

Author disclaimer

This article reflects the views of the authors and does not necessarily reflect those of the U.S. Food and Drug Administration.

Data availability

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This study was funded by the US Food and Drug Administration (FDA). This research was supported in part by an appointment to the Research Participation Program at the National Center for Toxicological Research (AA), administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the U.S. Food and Drug Administration.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.ebm-journal.org/articles/10.3389/ebm.2025.10374/full#supplementary-material

References

1. Classen DC, Pestotnik SL, Evans RS, Lloyd JF, Burke JP. Adverse drug events in hospitalized patients. Excess length of stay, extra costs, and attributable mortality. *Jama* (1997) **277**:301–6. doi:10.1001/jama.1997.03540280039031

2. Harpaz R, DuMouchel W, Shah NH, Madigan D, Ryan P, Friedman C. Novel data-mining methodologies for adverse drug event discovery and analysis. *Clin Pharmacol Ther* (2012) **91**:1010–21. doi:10.1038/clpt.2012.50

3. Alomar M, Tawfiq AM, Hassan N, Palaian S. Post marketing surveillance of suspected adverse drug reactions through spontaneous reporting: current status, challenges and the future. *Ther Adv Drug Saf* (2020) **11**:2042098620938595. doi:10.1177/2042098620938595

4. Waller PC. Making the most of spontaneous adverse drug reaction reporting. *Basic and Clin Pharmacol and Toxicol* (2006) **98**:320–3. doi:10.1111/j.1742-7843.2006.pto_286.x

5. U.S. Food and Drug Administration. Questions and answers on FDA's adverse event reporting system (FAERS). Available online at: https://www.fda.gov/drugs/surveillance/questions-and-answers-fdas-adverse-event-reporting-system-faers#:~:text=What%20is%20FAERS%3F,that%20were%20submitted%20to%20FDA (Accessed January 8, 2024).

6. Guo W, Pan B, Sakkiah S, Ji Z, Yavas G, Lu Y, et al. Informing selection of drugs for COVID-19 treatment through adverse events analysis. *Scientific Rep* (2021) **11**:14022. doi:10.1038/s41598-021-93500-5

7. Xu W, Zhu L, Wang J, Shi L, Tang X, Chen Q, et al. Safety assessment of Yasmin: real-world adverse event analysis using the FAERS database. *Eur J Obstet and Gynecol Reprod Biol* (2024) **301**:12–8. doi:10.1016/j.ejogrb.2024.07.048

8. Zhao B, Zhang X, Chen M, Wang Y. A real-world data analysis of acetylsalicylic acid in FDA Adverse Event Reporting System (FAERS) database. *Expert Opin Drug Metab and Toxicol* (2023) **19**:381–7. doi:10.1080/17425255.2023.2235267

9. Le H, Hong H, Ge W, Francis H, Lyn-Cook B, Hwang YT, et al. A systematic analysis and data mining of opioid-related adverse events submitted to the FAERS database. *Exp Biol Med (Maywood)* (2023) **248**:1944–51. doi:10.1177/15353702231211860

10. Hettne KM, Stierum RH, Schuemie MJ, Hendriksen PJ, Schijvenaars BJ, Mulligen EM, et al. A dictionary to identify small molecules and drugs in free text. *Bioinformatics* (2009) **25**:2983–91. doi:10.1093/bioinformatics/btp535

11. Sohn S, Clark C, Halgrim SR, Murphy SP, Chute CG, Liu H. MedXN: an open source medication extraction and normalization tool for clinical text. *J Am Med Inform Assoc* (2014) **21**:858–65. doi:10.1136/amiajnl-2013-002190

12. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. *J Am Med Inform Assoc* (2010) **17**:19–24. doi:10.1197/jamia.m3378

13. Ermshaus A, Piechotta M, Rüter G, Keilholz U, Leser U, Benary M. preon: fast and accurate entity normalization for drug names and cancer types in precision oncology. *Bioinformatics* (2024) **40**:btae085. doi:10.1093/bioinformatics/btae085

14. Fung KW, Bodenreider O, Aronson AR, Hole WT, Srinivasan S. Combining lexical and semantic methods of inter-terminology mapping using the UMLS. *Stud Health Technol Inform* (2007) **129**:605–9.

15. Miftahutdinov Z, Kadurin A, Kudrin R, Tutubalina E. Medical concept normalization in clinical trials with drug and disease representation learning. *Bioinformatics* (2021) **37**:3856–64. doi:10.1093/bioinformatics/btab474

16. Vasilakes J, Fan Y, Rizvi R, Bompelli A, Bodenreider O, Zhang R. Normalizing dietary supplement product names using the RxNorm model. *Stud Health Technol Inform* (2019) **264**:408–12. doi:10.3233/SHTI190253

17. Patrick J, Li M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *J Am Med Inform Assoc* (2010) **17**:524–7. doi:10.1136/jamia.2010.003939

18. Chapman AB, Peterson KS, Alba PR, DuVall SL, Patterson OV. Detecting adverse drug events with rapidly trained classification models. *Drug Saf* (2019) **42**:147–56. doi:10.1007/s40264-018-0763-y

19. He L, Yang Z, Lin H, Li Y. Drug name recognition in biomedical texts: a machine-learning-based method. *Drug Discov Today* (2014) **19**:610–7. doi:10.1016/j.drudis.2013.10.006

20. Sampathkumar H, Chen XW, Luo B. Mining adverse drug reactions from online healthcare forums using hidden Markov model. *BMC Med Inform Decis Mak* (2014) **14**:91. doi:10.1186/1472-6947-14-91

21. Le H, Chen R, Harris S, Fang H, Lyn-Cook B, Hong H, et al. RxNorm for drug name normalization: a case study of prescription opioids in the FDA adverse events reporting system. *Front Bioinformatics* (2023) **3**:1328613. doi:10.3389/fbinf.2023.1328613

22. Hamon T, Grabar N. Linguistic approach for identification of medication names and related information in clinical narratives. *J Am Med Inform Assoc* (2010) **17**:549–54. doi:10.1136/jamia.2010.004036

23. Segura-Bedmar I, Martínez P, Segura-Bedmar M. Drug name recognition and classification in biomedical texts: a case study outlining approaches underpinning automated systems. *Drug Discov Today* (2008) **13**:816–23. doi:10.1016/j.drudis.2008.06.001

24. Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Phys Doklady* (1965) **10**:707–10.

25. Tan P-N, Steinbach M, Kumar V. *Introduction to data mining*. 1st ed. Addison-Wesley Longman Publishing Co., Inc. (2005).

26. Chen R, Ho JC, Lin J-MS. Extracting medication information from unstructured public health data: a demonstration on data from population-based and tertiary-based samples. *BMC Med Res Methodol* (2020) **20**:258. doi:10.1186/s12874-020-01131-7

27. Peters L, Kapusnik-Uner JE, Nguyen T, Bodenreider O. An approximate matching method for clinical drug names. *AMIA Annu Symp Proc* (2011) **2011**: 1117–26.

28. Rocktäschel T, Weidlich M, Leser U. ChemSpot: a hybrid system for chemical named entity recognition. *Bioinformatics* (2012) **28**:1633–40. doi:10.1093/bioinformatics/bts183

29. Lafferty JD, McCallum A, Pereira FCN. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: *Proceedings of the eighteenth international conference on machine learning*. San Francisco, CA: Morgan Kaufmann Publishers Inc. (2001). p. 282–9.

30. Liu Z, Yang M, Wang X, Chen Q, Tang B, Wang Z, et al. Entity recognition from clinical texts via recurrent neural network. *BMC Med Inform Decis Mak* (2017) **17**:67. doi:10.1186/s12911-017-0468-7

31. Jagannatha AN, Yu H. Structured prediction models for RNN based sequence labeling in clinical text. *Proc Conf Empir Methods Nat Lang Process* (2016) **2016**: 856–65. doi:10.18653/v1/d16-1082

32. Habibi M, Weber L, Neves M, Wiegandt DL, Leser U. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics* (2017) **33**:i37–i48. doi:10.1093/bioinformatics/btx228

33. Wei Q, Ji Z, Li Z, Du J, Wang J, Xu J, et al. A study of deep learning approaches for medication and adverse drug event extraction from clinical text. *J Am Med Inform Assoc* (2020) **27**:13–21. doi:10.1093/jamia/ocz063

34. Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6 years. *J Am Med Inform Assoc* (2011) **18**:441–8. doi:10.1136/amiajnl-2011-000116

35. Freimuth RR, Wix K, Zhu Q, Siska M, Chute CG. Evaluation of RxNorm for medication clinical decision support. *AMIA Annu Symp Proc* (2014) **2014**:554–63.

36. RxNorm Files. RxNorm (2024). Available online at: https://www.nlm.nih.gov/research/umls/rxnorm/docs/rxnormfiles.html (Accessed January 8, 2024).

37. RxNorm. RxNorm technical documentation (2024). Available online at: https://www.nlm.nih.gov/research/umls/rxnorm/docs/techdoc.html (Accessed January 8, 2024).

38. Sayers E. The E-utilities in-depth: parameters, syntax and more. In: *Entrez programming Utilities help*. Bethesda (MD): National Center for Biotechnology Information US (2009). Available online at: https://www.ncbi.nlm.nih.gov/books/NBK25499/ (Accessed November 30, 2022).

39. Honnibal M, Johnson M. *An improved non-monotonic transition system for dependency parsing*. Lisbon, Portugal: Association for Computational Linguistics (2015). p. 1373–8.

40. Li Z, Huang R, Xia M, Patterson TA, Hong H. Fingerprinting interactions between proteins and ligands for facilitating machine learning in drug discovery. *Biomolecules* (2024) **14**:72. doi:10.3390/biom14010072

41. Liu J, Khan MKH, Guo W, Dong F, Ge W, Zhang C, et al. Machine learning and deep learning approaches for enhanced prediction of hERG blockade: a comprehensive QSAR modeling study. *Expert Opin Drug Metab and Toxicol* (2024) **20**:665–84. doi:10.1080/17425255.2024.2377593

42. Tang W, Zhang X, Hong H, Chen J, Zhao Q, Wu F. Computational nanotoxicology models for environmental risk assessment of engineered nanomaterials. *Nanomaterials* (2024) **14**:155. doi:10.3390/nano14020155

43. Guo W, Liu J, Dong F, Hong H. Unlocking the potential of AI: machine learning and deep learning models for predicting carcinogenicity of chemicals. *J Environ Sci Health C* (2024) **43**:23–50. doi:10.1080/26896583.2024.2396731

44. Huang L, Song M, Shen H, Hong H, Gong P, Deng H-W, et al. Deep learning methods for omics data imputation. *Biology* (2023) **12**:1313. doi:10.3390/biology12101313

45. Khan MKH, Guo W, Liu J, Dong F, Li Z, Patterson TA, et al. Machine learning and deep learning for brain tumor MRI image segmentation. *Exp Biol Med (Maywood)* (2023) **248**:1974–92. doi:10.1177/15353702231214259

46. Guo W, Liu J, Dong F, Song M, Li Z, Khan MKH, et al. Review of machine learning and deep learning models for toxicity prediction. *Exp Biol Med (Maywood)* (2023) **248**:1952–73. doi:10.1177/15353702231209421

47. Liu J, Xu L, Guo W, Li Z, Khan MKH, Ge W, et al. Developing a SARS-CoV-2 main protease binding prediction random forest model for drug repurposing for COVID-19 treatment. *Exp Biol Med (Maywood)* (2023) **248**:1927–36. doi:10.1177/15353702231209413

48. Ji Z, Guo W, Wood EL, Liu J, Sakkiah S, Xu X, et al. Machine learning models for predicting cytotoxicity of nanomaterials. *Chem Res Toxicol* (2022) **35**:125–39. doi:10.1021/acs.chemrestox.1c00310

49. Liu J, Guo W, Sakkiah S, Ji Z, Yavas G, Zou W, et al. Machine learning models for predicting liver toxicity. *Methods Mol Biol* (2022) **2425**:393–415. doi:10.1007/978-1-0716-1960-5_15

50. Liu J, Guo W, Dong F, Aungst J, Fitzpatrick S, Patterson TA, et al. Machine learning models for rat multigeneration reproductive toxicity prediction. *Front Pharmacol* (2022) **13**:1018226. doi:10.3389/fphar.2022.1018226

51. Guo W, Liu J, Dong F, Chen R, Das J, Ge W, et al. Deep learning models for predicting gas adsorption capacity of nanomaterials. *Nanomaterials* (2022) **12**:3376. doi:10.3390/nano12193376

52. Idakwo G, Thangapandian S, Luttrell J, Li Y, Wang N, Zhou Z, et al. Structure–activity relationship-based chemical classification of highly imbalanced Tox21 datasets. *J Cheminformatics* (2020) **12**:66. doi:10.1186/s13321-020-00468-x

53. Tan H, Wang X, Hong H, Benfenati E, Giesy JP, Gini GC, et al. Structures of endocrine-disrupting chemicals determine binding to and activation of the estrogen receptor α and androgen receptor. *Environ Sci Technol* (2020) **54**:11424–33. doi:10.1021/acs.est.0c02639

EBM Experimental Biology and Medicine

TYPE Original Research
PUBLISHED 28 February 2025
DOI 10.3389/ebm.2025.10389

OPEN ACCESS

*CORRESPONDENCE
Wen Zou,
wen.zou@fda.hhs.gov
Ningning Wu,
nxwu@ualr.edu

RECEIVED 28 September 2024
ACCEPTED 16 January 2025
PUBLISHED 28 February 2025

CITATION
Ma L, Chen R, Ge W, Rogers P, Lyn-Cook B, Hong H, Tong W, Wu N and Zou W (2025) AI-powered topic modeling: comparing LDA and BERTopic in analyzing opioid-related cardiovascular risks in women. *Exp. Biol. Med.* 250:10389. doi: 10.3389/ebm.2025.10389

AI-powered topic modeling: comparing LDA and BERTopic in analyzing opioid-related cardiovascular risks in women

Li Ma[1,2], Ru Chen[3], Weigong Ge[1], Paul Rogers[1], Beverly Lyn-Cook[4], Huixiao Hong[1], Weida Tong[1], Ningning Wu[2]* and Wen Zou[1]*

[1]Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, AR, United States, [2]Department of Information Science, University of Arkansas at Little Rock, Little Rock, AR, United States, [3]Office of New Drug, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD, United States, [4]Division of Biochemical Toxicology, National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, AR, United States

Abstract

Topic modeling is a crucial technique in natural language processing (NLP), enabling the extraction of latent themes from large text corpora. Traditional topic modeling, such as Latent Dirichlet Allocation (LDA), faces limitations in capturing the semantic relationships in the text document although it has been widely applied in text mining. BERTopic, created in 2022, leveraged advances in deep learning and can capture the contextual relationships between words. In this work, we integrated Artificial Intelligence (AI) modules to LDA and BERTopic and provided a comprehensive comparison on the analysis of prescription opioid-related cardiovascular risks in women. Opioid use can increase the risk of cardiovascular problems in women such as arrhythmia, hypotension etc. 1,837 abstracts were retrieved and downloaded from PubMed as of April 2024 using three Medical Subject Headings (MeSH) words: "opioid," "cardiovascular," and "women." Machine Learning of Language Toolkit (MALLET) was employed for the implementation of LDA. BioBERT was used for document embedding in BERTopic. Eighteen was selected as the optimal topic number for MALLET and 23 for BERTopic. ChatGPT-4-Turbo was integrated to interpret and compare the results. The short descriptions created by ChatGPT for each topic from LDA and BERTopic were highly correlated, and the performance accuracies of LDA and BERTopic were similar as determined by expert manual reviews of the abstracts grouped by their predominant topics. The results of the t-SNE (t-distributed Stochastic Neighbor Embedding) plots showed that the clusters created from BERTopic were more compact and well-separated, representing improved coherence and distinctiveness between the topics. Our findings indicated that AI algorithms could augment both traditional and contemporary topic modeling techniques. In addition, BERTopic has the connection port for ChatGPT-4-Turbo or other large language models in its algorithm for automatic interpretation, while with LDA interpretation must be manually,

and needs special procedures for data pre-processing and stop words exclusion. Therefore, while LDA remains valuable for large-scale text analysis with resource constraints, AI-assisted BERTopic offers significant advantages in providing the enhanced interpretability and the improved semantic coherence for extracting valuable insights from textual data.

KEYWORDS

AI, BERTopic, topic modeling, opioid, cardiovascular risks

Impact statement

This study provides a comparative analysis of LDA and BERTopic in the context of AI-driven topic modeling to analyze opioid-related cardiovascular risks in women. While both methods were capable of effectively identifying topics within text corpora, our findings reveal that BERTopic offers obvious advantage due to its seamless integration with AI techniques and improved semantic coherence in text documents. In addition, it uncovered themes related to opioid-associated health risks and outcomes in specialized patient groups, including pregnant patients and those undergoing coronary surgery. BERTopic's ability to automatically incorporate contextual information through transformer-based models makes it particularly well-suited for AI generation tasks, where adaptability and precision are critical. In comparison, LDA, although performing well, requires data pre-processing and manual adjustments to achieve similar levels of AI integration. These results underscore the potential importance of AI integration into topic modeling techniques in the analysis of large-scale biomedical text data to achieve more accurate and meaningful insights. This integration not only enhances the precision of topic modeling but also accelerates the modeling and output interpretation, potentially empowering researchers and practitioners with varying levels of expertise to derive valuable insights from unstructured text data.

Introduction

The opioid epidemic has become a serious national crisis in the United States, with far-reaching consequences across various populations [1]. In 2023, nearly 8.6 million Americans 12 years and older reported misusing prescription opioids and over 5 million reported a prescription use disorder in the past year [2]. It was reported that approximately 294,000 people died from overdoses involving prescription opioids from 1999 to 2022 [3]. Healthcare systems bear substantial costs due to increased hospitalizations and emergency department visits associated with opioid overdoses [4]. Women have seen a marked rise in opioid-related issues, for example, since 1999 deaths from prescription opioid overdoses increased 642% among women compared with a 439% increase among men [5]. Among these, women face unique cardiovascular risks associated with opioid use, necessitating targeted research to understand these complex relationships [6].

Natural language processing (NLP) has become an essential tool for extracting meaningful insights from vast amounts of biomedical literature, such as PubMed abstracts. Topic modeling has emerged as a foundational technique within the domain of NLP and text mining, providing an essential methodology for extracting insightful patterns from extensive and intricate text datasets. As an unsupervised machine learning approach, topic modeling discerns latent themes or topics within a corpus of documents, thereby facilitating the systematic organization, comprehension, and extraction of meaningful patterns from vast amounts of unstructured text data, where manual analysis is both impractical and infeasible [7–10].

The utilization of topic modeling extends across various fields and applications. It is broadly implemented in document classification and organization, text summarization, customer feedback analysis, sentiment analysis, trend analysis, bioinformatics analysis, and even biological and biomedical research [11–18]. Traditional topic modeling methods such as Latent Dirichlet Allocation (LDA) have established a robust framework for understanding and organizing unstructured text data. While these conventional techniques have proven effective across various contexts, they face evident limitations in capturing the intricate semantic relationships within increasingly complex and voluminous datasets [19], such as medical literature. Additionally, the outputs produced by these traditional methods often pose interpretability challenges, particularly for individuals who lack domain-specific expertise [11, 13, 20, 21].

BERTopic was developed by Grootendorst in 2022 by leveraging advances in deep learning, especially those in transformer-based models such as Bidirectional Encoder Representations from Transformers (BERT) which excels in generating deep contextualized word embeddings by considering the full context of words in a sentence, as read from both the left and the right [22]. While LDA (which employs the "bag-of-words" method) and TF-IDF etc. ignore the order and context of words [23], BERTopic uses BERT embeddings which can capture the contextual relationships between words [24]. BERTopic consists of four major modularized steps: document embedding, dimensionality reduction, clustering,

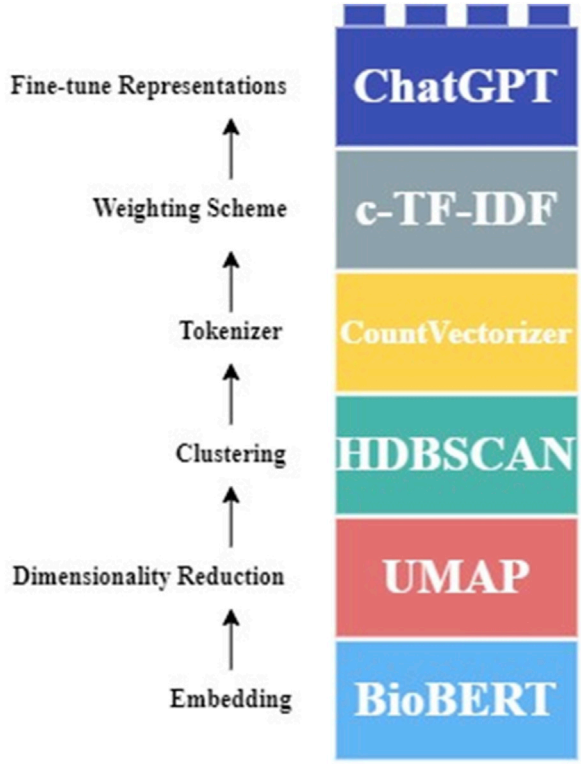


FIGURE 1
Customized BERTopic pipeline used for this study. The abstracts were first embedded using BioBERT. Then UMAP was employed for dimensionality reduction, followed by HDBSCAN for clustering. CountVectorizer was then used for tokenization and handling the stop words. c -TF-IDF was applied for a weighting scheme. Lastly, fine-tuning representations was achieved with ChatGPT, enhancing the interpretability and quality of the topic modeling outputs.

and topic representation [22]. These steps are independent from each other. Each of these steps can be modified or replaced without affecting the others. This allows users to customize and tailor the topic modeling processing by integrating different algorithms or techniques at different stages of the pipeline. [25]. Therefore, BERTopic has an important advantage in its modularity. This flexibility is able to extend integrating an Artificial Intelligence (AI) module for enhancing result interpretability.

Multiple studies have applied topic modeling to analyze text data from social media platforms, electronic health records, and adverse event reporting systems to understand patterns of opioid misuse and its societal impact. For instance, we have applied LDA to perform text mining on prescription opioids-related literatures in PubMed to capture the research themes and to explore the prevalent topic dynamics in the literatures [6]. LDA has also been utilized to examine X (previously Twitter data) for trends in public sentiment and discourse around opioid use, highlighting a range of topics including how opioids are administered, opioid use affecting life and withdrawal symptoms due to trying to quit opioids [26]. Although BERTopic modeling is a relatively new algorithm, it has been applied in the text mining in many fields [27, 28]. These approaches underscore the potential of topic modeling to inform policy, improve surveillance systems, and enhance targeted interventions addressing the opioid crisis.

AI has significantly evolved over the past few years and has found applications in an increasing number of fields. This study delves into the innovative integration of AI to aid in the interpretation of results derived from traditional topic modeling approaches like LDA and contemporary methods such as BERTopic. To comprehensively compare the performance of AI-integrated LDA and BERTopic, we used a curated dataset of biomedical abstracts retrieved from PubMed for prescription opioids-related cardiovascular issues in women. By applying these two AI-integrated models to this specific dataset, we aimed to evaluate their effectiveness and accuracies in uncovering the intricate themes within the literature and their ability to handle the complexities inherent in biomedical texts. The comparison focused on the coherence, contextual relevance, and ease of use of each model, providing insights into their respective strengths and limitations.

Materials and methods

Data collection and preprocessing

The PubMed abstract collection was pursued in April 2024 by utilizing the PubMed file retrieval tool easyPubMed v2.13. All the available abstracts which were published before April 15th, 2024, were obtained by searching PubMed with three mesh words: “opioid”, “cardiovascular”, and “women” based on the following criteria: a). language: only English-language abstracts were included; b). availability: abstracts that are freely accessible and available in full text; and c). species: the search scope was restricted to human related research. The search query was set up as “cardiovascular AND opioid AND humans [mh] AND english [la] AND [(women) OR (female)] NOT exclude preprints [Publication Type]”.

Datasets collected directly from PubMed may contain noisy information that can compromise the relevance of the results [18]. The curated abstract dataset was preprocessed firstly by removing numbers, punctuations, special characters, html tags and URLs from the dataset using sed (stream editor in Linux). Next, the Stanford NLP tool Stanza [29] was employed to lemmatize the pre-processed text data to remove inflectional endings and to convert a word back to its root form (e.g., running to run). Stop words were excluded by the *remove-stopwords* function in MALLET (Machine Learning for Language Toolkit) [30].

Unlike LDA, BERTopic does not need data preprocessing and uses the original sentences in the curated dataset as the original structures of the texts play vital roles for BERTopic’s transformer models.

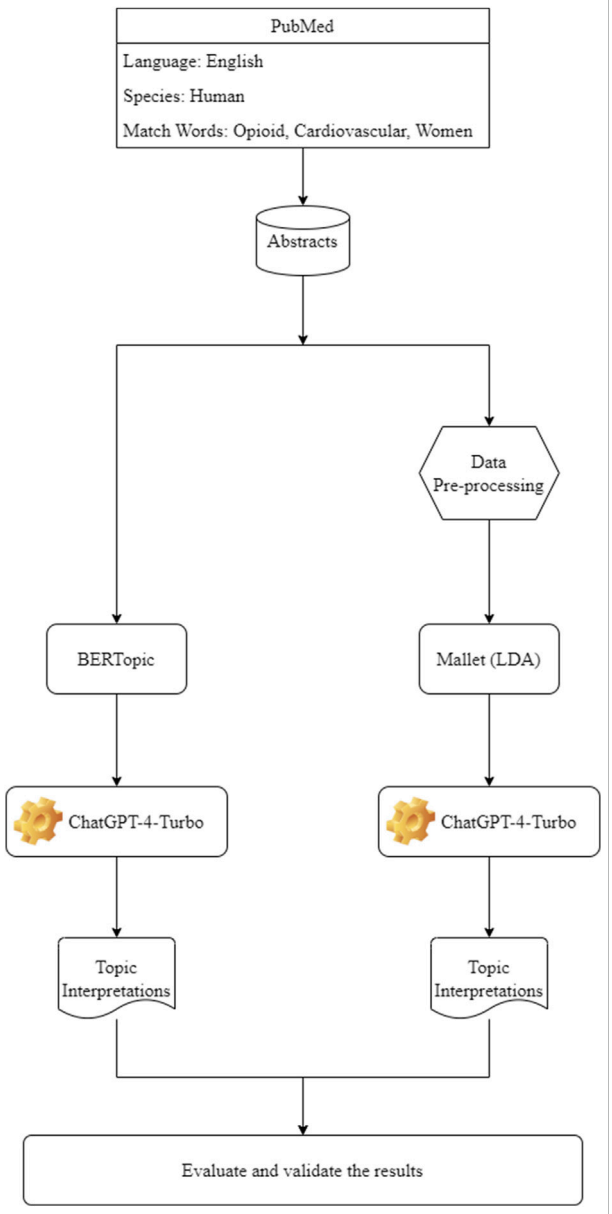


FIGURE 2
Research workflow.

Implementation of topic models

Model 1: ChatGPT powered LDA

The latest version, MALLET v2.0.8, was installed and used under openJDK 11.0.22. to implement LDA. MALLET is a Java-based package designed for statistical natural language processing, document classification, topic modeling, information extraction, and other machine learning applications regarding textual data [30].

As a crucial step in topic modeling, determining the optimal number of topics for a LDA model can significantly influence the quality and interpretability of the results [31]. It is often done through time-consuming trial-and-error or using perplexity-based methods which may not always yield stable results. In this study, the Rate of Perplexity Change (RPC)-based approach [32] was adopted to determine the optimal number of topics in LDA. This method aims to find the first change point in the RPC which implies the most appropriate number of topics for a given dataset.

MALLET was then re-run on the same dataset with all the same settings except changing the parameter for topic number to the most fitting one. The top 100 words of each topic along with their corresponding probabilities and the abstracts clustered in the topic were uploaded to ChatGPT-4-Turbo to generate a one-sentence description for each topic.

Model 2: BERTopic

As shown in Figure 1, BioBERT was selected for the document embedding as the first step [33]. For dimensionality reduction and clustering, the default algorithms of Uniform Manifold Approximation & Projection (UMAP) [34] and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [35] were used, respectively. Although not necessary to remove stop words from the data, CountVectorizer from the sklearn package [36] was utilized to handle the stop words. ChatGPT-4-Turbo was employed as the last step.

Performance comparison

Accuracy of relevance by manual expert review

To validate the accuracies of the outcomes from the two AI-integrated topic modeling approaches, we randomly picked one topic from each group of topics generated by the two approaches. The abstracts which were clustered into the two topics were read by domain experts to manually evaluate if the abstracts properly aligned with the respective topics.

Abstract clusters by visualization analysis

Each abstract was labeled with the topics and their corresponding probabilities, derived from LDA and BERTopic, respectively. t-SNE (t-distributed Stochastic Neighbor Embedding) [37] was utilized to reduce the 18-dimensional (LDA MALLET)/21-dimensional (BERTopic) topic probabilities to 3 dimensions prior to clustering the abstracts. Each cluster was then evaluated to see if the abstracts in a cluster had the same topic number.

Coherence comparison

In addition to comparing the topic results manually, UMass coherence score [38] was utilized to evaluate the performance of the two methods, BERTopic and LDA (MALLET). UMass coherence scores for each topic were calculated by using the Python package gensim v3.8.3 [39]. An acceptable coherence score should be a value between −14 and 14 according to the genism documentation [40].

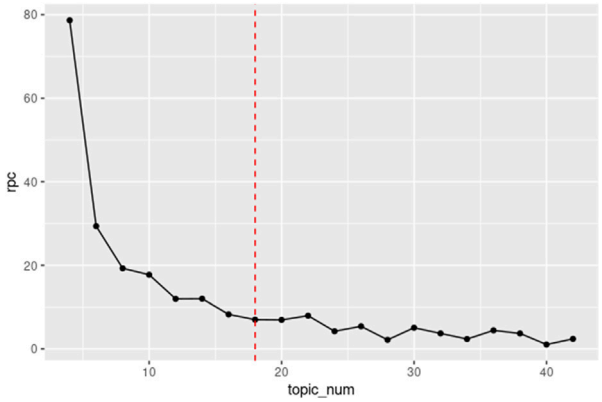


FIGURE 3
The Rate of Perplexity Change (RPC) Plot. Illustration of the RPC as a function of the number of topics (topic_num) for LDA modeling. The y-axis represents the RPC, while the x-axis indicates the number of topics.

Results and discussion

Currently, the development of AI has raised an interesting question on how to integrate the AI/large language models with the traditional NLP to enhance the capabilities of language processing systems, enabling more accurate, context-aware, and sophisticated analysis of text data. Traditional NLP methods often rely on rule-based approaches and statistical models, while AI, particularly through machine learning and deep learning, brings a more dynamic and context-aware understanding of language. In this study, we integrated ChatGPT4-Turbo to traditional MALLET-based LDA and BioBERT-embedded BERTopic and underscore the transformative potential of AI in enhancing topic modeling techniques. A dataset of PubMed abstracts focused on opioid-related cardiovascular risks in women was used as a case study. The objective was to compare and evaluate the performance of these two topic modeling techniques leveraged by the advances of AI algorithms in uncovering meaningful themes within a specialized biomedical dataset. The proposed approach can be applied to any topic modeling algorithm coupled with an AI system in the downstream pipeline through either manually implementation or automatic connection with programming language interface. Figure 2 illustrates the workflow for this study. 1,837 abstracts were retrieved and downloaded from PubMed through April 2024 using three MeSH words: "opioid," "cardiovascular," and "women," followed by processing by MALLET topic modeling or BERTopic. The outputs of the two topic modeling approaches were compared from several aspects following integration with ChatGPT-4-Turbo manually or automatically.

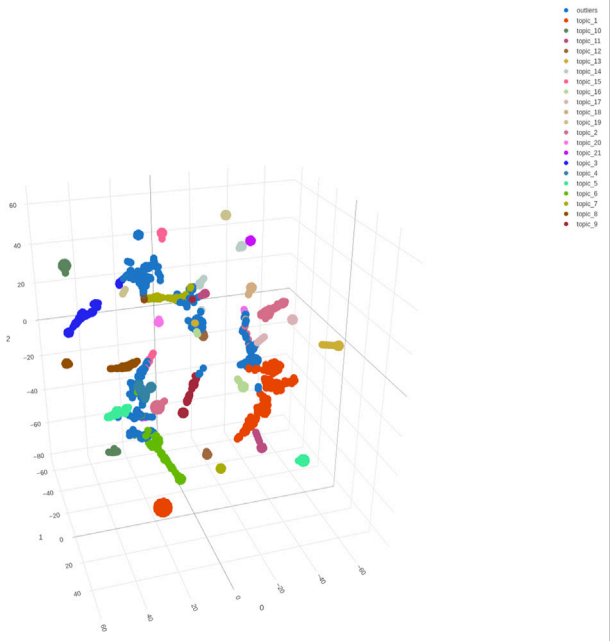


FIGURE 4
A 3D scatter plot visualization of the distribution of abstracts based on the topics identified by the BERTopic model. Each point represents an abstract and the color represents the abstracts assigned topic (Table 1). The axes represent the reduced dimensions obtained through t-SNE, which projects the high-dimensional topic space into three dimensions for better visualization.

Ease of large language model integration and use

As shown in Figure 2, the integration of ChatGPT-4-Turbo or other large languages with LDA required additional steps, including data pre-processing, tuning hyperparameters and manually uploading the LDA outputs to ChatGPT-4-Turbo for refining the generated topics. The RPC-based method was applied to calculate the rate of change in statistical perplexity, and using this approach, 18 was identified as the optimal topic number for the LDA topic modeling (Figure 3). These additional data processing steps were time-consuming and labor-intensive, and required a deep understanding of both the model and the dataset to achieve optimal results, especially when dealing with a large and complicated dataset.

BERTopic provided a simpler and streamlined experience. Since BERTopic technically determines the optimal topic number and the other parameters by the algorithm itself, it is not necessary to spend extra time tuning parameters. Consequently, BERTopic generated 21 topics from this dataset. The integration of ChatGPT-4-Turbo is an inherent part of the BERTopic, therefore, the ease of use and automatic

TABLE 1 The topics generated by ChatGPT-integrated BERTopic and LDA.

BERTopic	
Topic	**Topic Labels by ChatGPT-4-Turbo**
1	Effects of Remifentanil on Cardiovascular Response during Anesthesia Induction and Intubation
2	Cardiovascular Effects of Various Anesthetics in Coronary Surgery
3	Anesthesia Management in Pregnant Patients with Cardiac and Pulmonary Complications
4	Opioid Use and Associated Health Aspects in Various Patient Populations
5	Opioid Use and Associated Health Aspects in Various Patient Populations
6	Postoperative Pain Management in Cardiovascular Surgeries
7	Comparative Analyses of Opioid and Sedative Efficacy in Postoperative and Intensive Care Settings
8	Methadone Use and Associated Cardiovascular Risks
9	Opioid and Drug Use Disorders in Hospitalized Patients
10	Opioid-Associated Risks and Mortality in Medical Settings
11	Opioids and Health Outcomes in Specific Patient Populations
12	Cardiovascular Effects and Analgesia in Anesthesia Procedures
13	Opioid Use, Migraine Management, and Associated Health Outcomes
14	Opioid Effects and Management in Postoperative and Cardiac Care
15	Opioid and Opium Use Impact on Health and Mortality in Iran
16	Cardiovascular Effects of Anesthesia in Surgical Patients
17	Opioid and Anesthetic Effects on Cardiovascular and Hemostatic Responses in Clinical Settings
18	Management of Pain and Opioid Use in Clinical Settings
19	Effects of Salvinorin and Related Compounds on Opioid Receptors and Vascular Responses in Humans and Animal Models
20	Opioid Receptor Activation and Its Effects on Cellular and Physiological Responses
21	Drug Use and Health Outcomes in Special Populations
MALLET	
Topic	**Topic Labels by ChatGPT-4-Turbo**
1	Topic_1 primarily explores the interactions and effects of opioids on cardiovascular and autonomic nervous system responses, particularly focusing on changes in blood pressure, heart rate, and sympathetic activity during various medical procedures and conditions
2	The topic primarily focuses on the association between opioid use and various health outcomes, particularly cardiovascular diseases, in different populations
3	This topic primarily focuses on the medical complications and management associated with intravenous drug use, particularly in relation to infective endocarditis, vascular injuries, and the administration of anesthesia and analgesia in surgical settings
4	The topic primarily focuses on the clinical efficacy, safety, and comparative analysis of various pharmacological treatments, including opioids, NSAIDs, and other analgesics, in managing pain and related symptoms across different medical conditions and patient populations
5	The topic focuses on the effects and management of analgesia and anesthesia during labor and delivery, examining their impact on maternal and neonatal outcomes, including fetal heart rate patterns and drug exposure
6	The topic focuses on the comparison and evaluation of various sedative and analgesic drugs, their combinations, and administration techniques in managing pain and sedation during medical procedures, emphasizing their effects on cardiovascular and respiratory systems, efficacy, safety, and patient recovery outcomes
7	This topic focuses on the clinical outcomes, interventions, and trials related to acute myocardial infarction (AMI), including the use of various medications and their effects on patient recovery and complications

(Continued on following page)

TABLE 1 (*Continued*) The topics generated by ChatGPT-integrated BERTopic and LDA.

MALLET	
Topic	**Topic Labels by ChatGPT-4-Turbo**
8	Topic 8 focuses on the comparison and evaluation of different surgical and anesthetic techniques in managing postoperative outcomes, pain, and complications in various medical procedures, emphasizing the effectiveness, safety, and impact on recovery times and hospital stays
9	This topic focuses on the study of opioid effects and interactions, particularly in relation to pain, stress responses, and various physiological and behavioral outcomes
10	The effects of various anesthetic agents and techniques on cardiovascular and hemodynamic responses during surgical procedures
11	Topic_11 focuses on the comparison and evaluation of different analgesic techniques and medications for postoperative pain management, particularly in relation to their efficacy, side effects, and impact on recovery in various surgical settings
12	The topic primarily focuses on the health impacts, particularly cardiovascular and pulmonary issues, associated with substance abuse and overdose, including the effects of various drugs like opioids, cocaine, and methadone on the human body
13	Topic 13 focuses on the effects and management of anesthesia, particularly in relation to cardiovascular stability, respiratory effects, and recovery times, with a specific emphasis on various anesthetic agents and techniques used during surgical procedures
14	Pharmacokinetics and pharmacodynamics of opioids and other drugs, focusing on their distribution, clearance, and effects within the body, particularly in relation to the blood-brain barrier and various bodily fluids
15	Polypharmacy and drug interactions in elderly patients, focusing on adverse drug reactions, prescribing patterns, and the impact of specific medications on health outcomes
16	Topic_16 focuses on the management and outcomes of opioid use in various medical settings, including treatment programs and emergency interventions, with an emphasis on the impact of opioid-related complications and mortality rates
17	Topic 17 primarily explores the cardiovascular effects and electrophysiological properties of various substances, including opioids and anesthetics, on the human heart, particularly focusing on QT interval prolongation, heart rate, and blood pressure responses
18	This topic involves the study of opioid receptors and their interactions with various peptides and antagonists, focusing on their effects on cardiovascular and neuroendocrine responses, as well as their potential therapeutic applications in conditions like pain management, shock, and cardiac function

handling of context made BERTopic more accessible and efficient, reducing the need for extensive manual intervention.

Topic coherence and interpretability

Table 1 shows the output topics from AI-integrated LDA model (18 topics) and BERTopic model (21 topics), that were subsequently interpreted by ChatGPT-4-Turbo. The MALLET-based LDA model generated 18 topics across the dataset (Table 1). The topics generated by LDA were coherent and interpretable, for example, Topic 1, explored the "interactions and effects of opioids on cardiovascular and autonomic nervous system responses," with a broad focus on "blood pressure, heart rate, and sympathetic activity." The manual integration of AI techniques improved coherence to an extent and lowered dependency on expertise. While relevant themes were captured, the interpretations from ChatGPT-4-Turbo seemed to lack contextual specificity, potentially due to the indirect manual integration of the LDA outputs into ChatGPT-4-Turbo.

Comparatively, the abstracts in the same dataset were first embedded using BioBERT (Figure 2). The use of transformer-based contextual embeddings enabled BERTopic to distinguish between closely related concepts, such as different cardiovascular effects of various anesthetic agents and opioid use during specific medical procedures. Consequently, BERTopic generated 21 distinct topics, with clear, contextually rich labels (Table 1). For instance, Topic 1 focused on the "Effects of Remifentanil on Cardiovascular Response during Anesthesia Induction and Intubation," while Topic 12 highlighted "Cardiovascular Effects and Analgesia in Anesthesia Procedures." These topics demonstrated BERTopic's ability to generate specific and clinically relevant themes directly tied to the nuances of medical procedures and conditions. In addition, BERTopic excelled in maintaining contextual relevance. For instance, BERTopic differentiated between opioid-induced complications in various medical settings, such as postoperative care and labor (Topic 6, 7, 14), without the need for extensive manual adjustments.

We also calculated the UMass coherence scores to evaluate the performance of the topic modeling (Table 2). The calculation of the UMass coherence score provided a quantitative measure of topic quality based on word co-occurrence patterns within the dataset. More negative values represent that the words rarely co-occur, while values closer to zero indicate a higher tendency for words to co-occur [38, 41]. The UMass coherence scores of the 18 topics generated by LDA MALLET fall within the interval of −3.1 and −1 (Table 2a); and the coherence scores of the 21 topics from BERTtopic fall within the interval of −1.1 and 0

TABLE 2 Coherence scores of the topics generated by LDA (MALLET) (a) and BERTopic (b).

(a)	
Topic	UMass coherence score
1	−1.5384
2	−2.0769
3	−1.4679
4	−1.9958
5	−2.1126
6	−1.8969
7	−2.1451
8	−1.9566
9	−2.0757
10	−1.7339
11	−3.0060
12	−2.2331
13	−1.3128
14	−1.9189
15	−1.5609
16	−1.7165
17	−1.8042
18	−1.3799
(b)	
Topic	UMass coherence score
1	−0.1023
2	−0.0798
3	−0.0593
4	−0.0531
5	−0.2831
6	−0.4271
7	−0.1564
8	−0.2707
9	−0.0783
10	−0.1169
11	−0.2800
12	−0.0983
13	−0.1544
14	−0.4083

(Continued in next column)

TABLE 2 (*Continued*) Coherence scores of the topics generated by LDA (MALLET) (a) and BERTopic (b).

(b)	
Topic	UMass coherence score
15	−0.1719
16	−1.0587
17	−0.6269
18	−0.6592
19	−0.7001
20	−0.9951
21	−0.1447

(Table 2b). All the coherence scores were between −14 and 14, and therefore considered as reasonable according to the genism documentation [40]. Comparatively, higher coherence scores for the topics from BERTopic (Table 2b) indicated that the words within a topic are more semantically related, reflecting better topic interpretability for the dataset.

Relative accuracy

Topic 7 generated by MALLET-based LDA, and Topic 10 generated by BERTopic were randomly selected and expert reviews were conducted to assess the accuracy of topic relevancies of the clustered abstracts. Supplementary Figure S1A, S1B are the word clouds generated from Topic 7 using LDA and from Topic 10 using BERTopic, respectively. Topic 7 generated using LDA is interpreted as "focuses on the clinical outcomes, interventions, and trials related to acute myocardial infarction (AMI), including the use of various medications and their effects on patient recovery and complications" (Table 1). Results from two experts' reviews revealed that 44 out of 49 (89.8%) abstracts were relevant to this topic (Supplementary Table S1A). Comparatively, Topic 10 from BERTopic was interpreted as "Opioid-Associated Risks and Mortality in Medical Settings" (Table 1), and the experts' reviews revealed that 45 out of 53 (84.9%) abstracts were relevant to the topic (Supplementary Table S1B). The reviews also revealed that both models produced topics with similar levels of accuracy in terms of relevance to the opioid-related cardiovascular issues present in the dataset. The overall alignment with expert-identified themes was consistent across both approaches.

Although the two topics were randomly selected for both topic modeling approaches, the relatively lower level of accuracy of relevance from the BERTopic-generated topic might be due to the insufficient tuning of the parameters after embedding with BioBERT. In this study, we adopted the default setting provided by the developer, and ignored the fact that the hyperparameters,

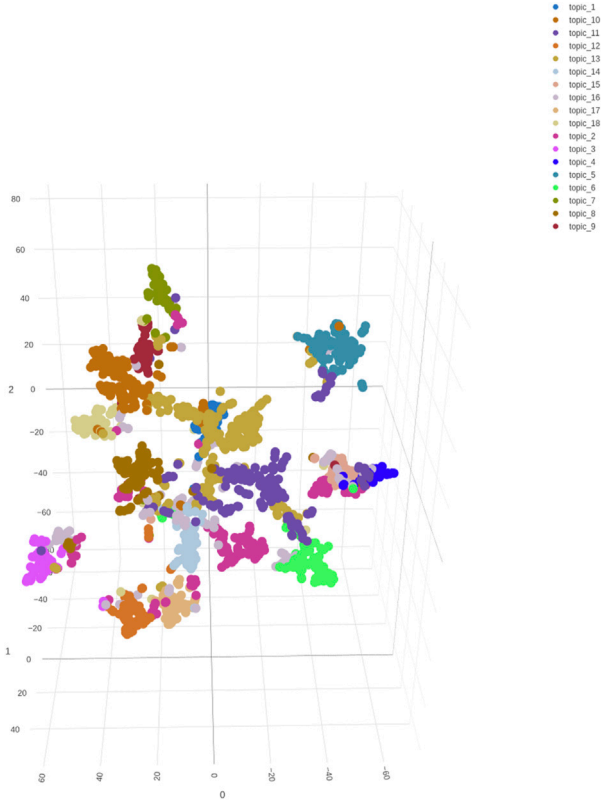


FIGURE 5
A 3D scatter plot visualization of the distribution of abstracts based on the topics identified by the MALLET-based LDA model. Each point represents an abstract and the color represents the abstracts assigned topic. The axes represent the reduced dimensions obtained through t-SNE, which projects the high-dimensional topic space into three dimensions for better visualization.

such as n_components, n_neighbors of UMAP, and the parameter min_cluster_size of HDBSCAN, may directly affect the decision for the number of topics after embedding, thereby affecting the topic outputs.

Specificity by visualization analysis

Hypothetically, each abstract is represented as a point in an 18-dimensional (LDA)/21-dimensional (BERTopic) space where each coordinate represents the probability that a corresponding topic occurring in the document. The high dimensionality of the data usually poses challenges for visualization. As such, t-SNE [37] was used to reduce the 18-dimensional/21-dimensional topic probabilities to 3 dimensions, respectively. This transformation provided an intuitive visualization of abstract distribution by plotting the data in 3-dimensional space. Figures 4, 5 show the distributions of abstracts labeled by BERTopic and Mallet, respectively. Each point represents an abstract, and each color represents a different topic to which the abstract was clustered. The 3D scatter plot provides a powerful tool for visually assessing the performance of the LDA and BERTopic models and understanding the structure of the topics within the dataset. Comparing the separation of clusters in both plots, BERTopic (Figure 4) shows a larger distance between the clusters and more distinct separation between the clusters as compared to LDA (Figure 5), which suggests better topic coherence and differentiation. Meanwhile, the LDA scatter plot (Figure 5) reveals more overlapping clusters than the BERTopic scatter plot (Figure 4), indicating that LDA topics are less distinct or more generalized, which is in agreement with the results in Table 1. Moreover, since the outliers may represent abstracts that don't fit well into any of the topics, fewer outliers in Figure 4 generally suggests that BERTopic is better at capturing the underlying structure of the dataset. In summary, for biomedical applications, particularly those involving detailed and context-sensitive information like opioid-related cardiovascular risks in women, BERTopic offers some advantages over traditional LDA models, and better captures the structure and themes of this dataset, especially when using AI-enhancement.

The results of the case study: opioids-related cardiovascular risks in women

Both AI-integrated LDA and BERTopic models generated meaningful topics from the case study on opioid-related cardiovascular risks in women, highlighting key areas of concern and clinical relevance. The BERTopic model identified 21 distinct topics, emphasizing specific clinical contexts such as the effects of remifentanil on cardiovascular response during anesthesia, opioid use in different patient populations, and postoperative pain management following cardiovascular surgeries. This model also uncovered themes related to opioid-associated health risks and outcomes in specialized patient groups, including pregnant patients and those undergoing coronary surgery.

In comparison, the LDA model generated 18 topics with a broader focus on the interactions between opioids with cardiovascular and autonomic nervous system responses. This analysis highlighted the association between opioid use and cardiovascular diseases across different populations and detailed the medical complications related to intravenous drug use, such as infective endocarditis and vascular injuries. The LDA model also explored the efficacy and safety of various pharmacological treatments, including opioids, in managing pain and related symptoms, particularly in surgical and postoperative settings.

We will apply the results obtained from the two topic modeling approaches and explore the detailed information for opioid-related cardiovascular risks in women in a future study.

In conclusion, this study demonstrated the effectiveness of both AI-integrated LDA and BERTopic in the text mining of opioid-related cardiovascular risks in women, with BERTopic offering more granular insights, context-specific topics, and a user-friendly working stream

through its AI integration. In this study, we did not try to change the traditional/advanced topic modeling algorithms, but to integrate AI tools to enhance or empower the performance of the models. The findings highlighted the importance of AI integration with traditional NLP techniques, which reveal potentially promising directions for future research advancements. By combining the strengths of traditional methods with the advanced pattern recognition and contextual understanding of AI, researchers and developers can build more powerful tools for applications in many fields. As AI continues to evolve, its integration with NLP will likely drive further innovations in how we understand and interact with language. It is reasonable to expect that integrating AI into currently available computational algorithms is a highly promising approach that enhances efficiency, adaptability, and accuracy across various domains. Meanwhile, there exist some challenges to consider. Based on our limited experiences, thoughtful design and validation (such as domain expertise integration) are essential to maximize its benefits.

Author contributions

LM performed all the calculations and data analysis and wrote the first draft of the manuscript. This work was established primarily by WZ in developing the methods, conceiving the original idea, and guiding the data analysis and presentation of results. LM, NW, and WZ participated in the data set construction and the resulting figures. WT contributed to project management and interpreting the results. All authors contributed to data verification, approach evaluation, and assisted with writing the manuscript. All authors contributed to the article and approved the submitted version.

Author disclaimer

The views presented in this article do not necessarily reflect those of the U.S. Food and Drug Administration. Any mention of commercial products is for clarification and is not intended as an endorsement.

Data availability

Publicly available datasets were analyzed in this study. This data can be found here: https://pubmed.ncbi.nlm.nih.gov/.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was funded by the Center for Drug Evaluation and Research, and the National Center for Toxicological Research of US Food and Drug Administration. LM acknowledges the support of a fellowship from the Oak Ridge Institute for Science and Education (ORISE), administered through an interagency agreement between the US Department of Energy and the US Food and Drug Administration.

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.ebm-journal.org/articles/10.3389/ebm.2025.10389/full#supplementary-material

References

1. CDC. *Understanding the opioid overdose epidemic* (2024). Available from: https://www.cdc.gov/overdose-prevention/about/understanding-the-opioid-overdose-epidemic.html (Accessed August 21, 2024).

2. SAMHSA. *Key substance use and mental health indicators in the United States: results from the 2023 national survey on drug use and health*. HHS, editor: HHS Publication No. PEP24-07-021, NSDUH Series H-59 (2024).

3. CDC. *Wide-ranging online data for epidemiologic research (WONDER)* (2018). Available from: https://wonder.cdc.gov/ (Accessed September 1, 2024).

4. Florence CS, Zhou C, Luo F, Xu L. The economic burden of prescription opioid overdose, abuse, and dependence in the United States, 2013. *Med Care* (2016) **54**: 901–6. doi:10.1097/mlr.0000000000000625

5. Abuse NNIoD. *Drug overdose deaths: facts and figures* (2019).

6. Le H, Hong H, Ge W, Francis H, Lyn-Cook B, Hwang YT, et al. A systematic analysis and data mining of opioid-related adverse events submitted to the FAERS database. *Exp Biol Med (Maywood)* (2023) **248**:1944–51. doi:10.1177/15353702231211860

7. Snyder RM. An introduction to topic modeling as an unsupervised machine learning way to organize text information. *Assoc Supporting Computer Users Education* (2015).

8. Mohr JW, Bogdanov P. *Introduction—topic models: what they are and why they matter*. Elsevier (2013). p. 545–69.

9. George LE, Birla L. A study of topic modeling methods. In: *Second international conference on intelligent computing and control systems (ICICCS)*. IEEE (2018). p. 109–13.

10. Alghamdi R, Alfalqi K. A survey of topic modeling in text mining. *Int J Adv Computer Sci Appl* (2015) **6**. doi:10.14569/ijacsa.2015.060121

11. Abdelrazek A, Eid Y, Gawish E, Medhat W, Hassan A. Topic modeling algorithms and applications: a survey. *Inf Syst* (2023) **112**:102131. doi:10.1016/j.is.2022.102131

12. Boyd-Graber J, Hu Y, Mimno D. Applications of topic models. *Foundations Trends® Inf Retrieval* (2017) **11**:143–296. doi:10.1561/1500000030

13. Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia tools Appl* (2019) **78**:15169–211. doi:10.1007/s11042-018-6894-4

14. Liu L, Tang L, Dong W, Yao S, Zhou W. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus* (2016) **5**:1608–22. doi:10.1186/s40064-016-3252-8

15. Zhao W, Chen JJ, Perkins R, Wang Y, Liu Z, Hong H, et al. A novel procedure on next generation sequencing data analysis using text mining algorithm. *BMC Bioinformatics* (2016) **17**:213. doi:10.1186/s12859-016-1075-9

16. Zhao W, Chen JJ, Foley S, Wang Y, Zhao S, Basinger J, et al. Biomarker identification from next-generation sequencing data for pathogen bacteria characterization and surveillance. *Biomark Med* (2015) **9**:1253–64. doi:10.2217/bmm.15.88

17. Wang SH, Ding Y, Zhao W, Huang YH, Perkins R, Zou W, et al. Text mining for identifying topics in the literatures about adolescent substance use and depression. *BMC Public Health* (2016) **16**:279. doi:10.1186/s12889-016-2932-1

18. Le H, Zhou J, Zhao W, Perkins R, Ge W, Lyn-Cook B, et al. Text fingerprinting and topic mining in the prescription opioid use literature. *2021 IEEE Int Conf Bioinformatics Biomed (Bibm)* (2021) 2741–8. doi:10.1109/bibm52615.2021.9669550

19. Vayansky I, Kumar SA. A review of topic modeling methods. *Inf Syst* (2020) **94**:101582. doi:10.1016/j.is.2020.101582

20. Kherwa P, Bansal P. Topic modeling: a comprehensive review. *EAI Endorsed Trans scalable Inf Syst* (2019) **7**. doi:10.4108/eai.13-7-2018.159623

21. Jiang Y, Fu M, Fang J, Rossi M. Applying topic modeling with prior domain-knowledge in information systems research. In: *Pacific asia conference on information systems*. Nanchang, China: Association for Information Systems (2023). p. 1582.

22. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv preprint. *arXiv:220305794* (2022).

23. Wallach HM. Topic modeling: beyond bag-of-words. In: *Proceedings of the 23rd international conference on Machine learning* (2006). p. 977–84.

24. Turton J, Vinson D, Smith RE. Deriving contextualised semantic features from bert (and other transformer model) embeddings. arXiv preprint. *arXiv:201215353* (2020).

25. Grootendorst M. *The algorithm* (2024). Available from: https://maartengr.github.io/BERTopic/algorithm/algorithm.html (Accessed August 15, 2024).

26. Pandrekar S, Chen X, Gopalkrishna G, Srivastava A, Saltz M, Saltz J, et al. Social media based analysis of opioid epidemic using reddit. *AMIA Annu Symp Proc* (2018) **2018**:867–76.

27. Baird A, Xia Y, Cheng Y. Consumer perceptions of telehealth for mental health or substance abuse: a Twitter-based topic modeling analysis. *JAMIA Open* (2022) **5**:ooac028. doi:10.1093/jamiaopen/ooac028

28. Raza S, Schwartz B, Lakamana S, Ge Y, Sarker A. A framework for multi-faceted content analysis of social media chatter regarding non-medical use of prescription medications. *BMC Digit Health* (2023) **1**:29. doi:10.1186/s44247-023-00029-w

29. Qi P, Zhang Y, Zhang Y, Bolton J, Manning CD. Stanza: a Python natural language processing toolkit for many human languages. *arXiv preprint arXiv:200307082* (2020). doi:10.48550/arXiv.2003.07082

30. McCallun AK. *MALLET: a machine learning for Language Toolkit* (2002). Available from: http://mallet.cs.umass.edu (Accessed August 17, 2024).

31. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *J machine Learn Res* (2003) **3**:993–1022. doi:10.7551/mitpress/1120.003.0082

32. Zhao W, Chen JJ, Perkins R, Liu Z, Ge W, Ding Y, et al. A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC bioinformatics* (2015) **16**:S8–10. doi:10.1186/1471-2105-16-s13-s8

33. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* (2020) **36**:1234–40. doi:10.1093/bioinformatics/btz682

34. McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:180203426* (2018). doi:10.48550/arXiv.1802.03426

35. Campello RJ, Moulavi D, Sander J. Density-based clustering based on hierarchical density estimates. In: *Pacific-Asia conference on knowledge discovery and data mining*. Springer (2013). p. 160–72.

36. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J machine Learn Res* (2011) **12**:2825–30. doi:10.5555/1953048.2078195

37. Van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Machine Learn Res* (2008) **9**:2579–2605.

38. Mimno D, Wallach H, Talley E, Leenders M, McCallum A. Optimizing semantic coherence in topic models. In: *Proceedings of the 2011 conference on empirical methods in natural language processing2011* (2024). p. 262–72.

39. Řehůřek R, Sojka P. *Software framework for topic modelling with large corpora* (2010).

40. Coder O-O. *Data science topics* (2019). Available from: https://datascience.oneoffcoder.com (Accessed November 15, 2024).

41. MALLET. *Topic model diagnostics* (2018). Available from: https://mallet.cs.umass.edu/diagnostics.php (Accessed November 15, 2024).

EBM Experimental Biology and Medicine

TYPE Original Research
PUBLISHED 03 December 2024
DOI 10.3389/ebm.2024.10393

OPEN ACCESS

*CORRESPONDENCE
Joshua Xu,
joshua.xu@fda.hhs.gov
Dan Li,
dan.li@fda.hhs.gov

RECEIVED 02 October 2024
ACCEPTED 18 November 2024
PUBLISHED 03 December 2024

CITATION
Li D, Wu L, Lin Y-C, Huang H-Y, Cotton E, Liu Q, Chen R, Huang R, Zhang Y and Xu J (2024) Enhancing pharmacogenomic data accessibility and drug safety with large language models: a case study with Llama3.1. *Exp. Biol. Med.* 249:10393. doi: 10.3389/ebm.2024.10393

Enhancing pharmacogenomic data accessibility and drug safety with large language models: a case study with Llama3.1

Dan Li[1]*, Leihong Wu[1], Ying-Chi Lin[2,3], Ho-Yin Huang[2,4], Ebony Cotton[1], Qi Liu[5], Ru Chen[6], Ruihao Huang[5], Yifan Zhang[1] and Joshua Xu[1]*

[1]Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, AR, United States, [2]School of Pharmacy, College of Pharmacy, Kaohsiung Medical University, Kaohsiung, Taiwan, [3]Master/Doctoral Degree Program in Toxicology, College of Pharmacy, Kaohsiung Medical University, Kaohsiung, Taiwan, [4]Department of Pharmacy, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Kaohsiung, Taiwan, [5]Center for Drug Evaluation and Research (CDER), U.S. Food and Drug Administration (FDA), Silver Spring, MD, United States, [6]Immediate Office, Office of Translational Sciences, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, United States

Abstract

Pharmacogenomics (PGx) holds the promise of personalizing medical treatments based on individual genetic profiles, thereby enhancing drug efficacy and safety. However, the current landscape of PGx research is hindered by fragmented data sources, time-consuming manual data extraction processes, and the need for comprehensive and up-to-date information. This study aims to address these challenges by evaluating the ability of Large Language Models (LLMs), specifically Llama3.1-70B, to automate and improve the accuracy of PGx information extraction from the FDA Table of Pharmacogenomic Biomarkers in Drug Labeling (FDA PGx Biomarker table), which is well-structured with drug names, biomarkers, therapeutic area, and related labeling texts. Our primary goal was to test the feasibility of LLMs in streamlining PGx data extraction, as an alternative to traditional, labor-intensive approaches. Llama3.1-70B achieved 91.4% accuracy in identifying drug-biomarker pairs from single labeling texts and 82% from mixed texts, with over 85% consistency in aligning extracted PGx categories from FDA PGx Biomarker table and relevant scientific abstracts, demonstrating its effectiveness for PGx data extraction. By integrating data from diverse sources, including scientific abstracts, this approach can support pharmacologists, regulatory bodies, and healthcare researchers in updating PGx resources more efficiently, making critical information more accessible for applications in personalized medicine. In addition, this approach shows potential of discovering novel PGx information, particularly of underrepresented minority ethnic groups. This study highlights the ability of LLMs to

enhance the efficiency and completeness of PGx research, thus laying a foundation for advancements in personalized medicine by ensuring that drug therapies are tailored to the genetic profiles of diverse populations.

KEYWORDS

pharmacogenomics, large language models, LLMs, biomarker, minority ethnic groups

Impact statement

This study demonstrates the utility of Large Language Models (LLMs), specifically Llama3.1-70B, in automating pharmacogenomic (PGx) data extraction, addressing the limitations of traditional manual methods that are labor-intensive and slow to update. By achieving high accuracy in identifying drug-biomarker pairs and integrating diverse data sources, this work offers a practical solution for pharmacologists, regulatory agencies, and healthcare professionals to streamline PGx database updates. With automated extraction processes, LLMs reduce the time and effort required to incorporate new PGx insights, potentially enabling updates at a frequency and scale that were previously unfeasible. This capability is critical for translating PGx research into actionable, personalized treatment guidelines that reflect the genetic diversity of patient populations, ultimately advancing equity in personalized medicine.

Introduction

Pharmacogenomics (PGx) represents a pivotal advancement in personalized medicine, tailoring drug therapies based on an individual's genetic profile [1, 2]. By understanding how genetic variations influence drug response, PGx enables healthcare providers to optimize treatment efficacy and minimize adverse drug reactions [3, 4]. This personalized approach holds the potential to significantly enhance patient outcomes, especially in the management of complex diseases such as cancer, cardiovascular disorders, and mental health conditions [5]. The importance of PGx lies in its ability to provide more precise and effective treatments. For instance, variations in genes encoding drug-metabolizing enzymes, drug transporters, and drug targets can greatly influence a patient's response to certain medications. These genetic differences can determine whether a patient will benefit from a particular drug, experience no effect, or suffer from adverse reactions [6, 7].

Despite its promise, the clinical implementation of PGx has been slower than anticipated, partly due to the complexity of drug-gene interactions and the need for extensive empirical evidence [8]. As our understanding of genetic factors in drug response continues to grow, PGx is poised to become a standard component of healthcare, revolutionizing the way treatments are tailored to individual patients. Various databases and resources for PGx information have been established to improve the accessibility and utility of this data. Key resources include the Pharmacogenomics Knowledgebase (PharmGKB), which curates information about how genetic variations affect drug response [9]. The pharmacogenomics database (PGxDB) database offers a comprehensive platform for integrating PGx data, allowing researchers to explore drug, target, and disease relationships [10]. Additionally, the FDA has released the Table of Pharmacogenomic Biomarkers in Drug Labeling (Table of Pharmacogenomic Biomarkers in Drug Labeling | FDA), providing drug and PGx biomarker pairs found in given drug labeling sections which serves as the primary data source for this study. Meanwhile, PGx-related research articles containing new findings and conclusions are crucial for timely updating of current PGx information. For instance, relevant abstracts can be retrieved from PubMed or other resources. These resources are essential for advancing the field of PGx and ensuring that clinicians have the necessary tools to apply genetic insights to patient care.

Large Language Models (LLMs) like Llama3.1 represent a significant advancement in natural language processing, offering powerful capabilities for extracting and analyzing complex data from diverse sources. These models, trained on vast amounts of text, can understand and generate human-like language, making them highly effective for tasks such as data extraction, summarization, and information synthesis [11, 12]. Recent studies have demonstrated the potential of LLMs in various fields, including PGx. For instance, LLMs have been shown to significantly improve the efficiency and accuracy of data extraction processes, and AI assistant showed improved efficacy in answering user questions [13]. By leveraging these models, researchers can automate the extraction of PGx information, overcoming challenges related to the time-consuming and labor-intensive nature of manual data processing.

In this study, we focused on evaluating the capabilities of LLMs, particularly Llama3.1-70B [14, 15], for PGx information extraction from various data sources. Our goal was to enhance the current PGx information collection by improving its accuracy and incorporating recent studies to fill in gaps and ensure the data is comprehensive. It was essential to ensure that the model could reliably identify and extract key PGx data, such as drugs and related biomarkers, from diverse sources with a remarkable degree of precision. The model demonstrated a high accuracy rate of 91.4% when extracting information from structured texts in the FDA PGx Biomarker table and 82% from the mixed texts, underscoring its effectiveness in handling different types of data.

A key aspect of our study was the integration of diverse resources, including well-structured databases like the FDA PGx Biomarker table, alongside relevant scientific abstracts. By combining these sources, we were able to cross-validate and enrich the PGx data, ensuring a more comprehensive, accurate, and up-to-date dataset, particularly with insights related to underrepresented populations and novel drug-biomarker interactions. The results can better support personalized medicine initiatives and enhance the overall effectiveness of pharmacogenomic applications.

Materials and methods

Data processing for the FDA PGx biomarker table

The FDA PGx Biomarker table (06/2023 version) was downloaded in PDF format and converted into one Excel table. All the special characters were then removed from the texts. Biomarkers with multiple gene names or aliases were further processed to ensure all the entries were retained. For instance, for the listed biomarker ERBB2 (HER2), either ERBB2 or HER2 identified by the model was considered a correct identification. To ensure there was sufficient content from which the model could extract information, labeling texts in the FDA PGx Biomarker table with fewer than 300 words were removed from the analysis.

Prompt and model settings

The Llama3.1-70B-Instruct model [14, 15] was employed in this study for the PGx information extraction and summarization. The model was run using its default settings. We utilized the "client.chat.completions.create" function to interact with the model and obtain the responses. To guide the model effectively, we set the system context as: "You are an expert in pharmacogenetics and assist me in extracting information from texts." This context was designed to align the model's responses with the specialized nature of the task. The PGx texts from the PGx Biomarker table that required information extraction, along with specific questions, were provided in the prompt as user content. For example, a typical prompt would be: "*Please review this labeling text and identify the pairs of drug and biomarker clearly mentioned. Output the pairs in 'drug-biomarker' format. Please try to give me both the generic name and brand name of the drug.*" As a result, the model may identify multiple drug-biomarker pairs from the query texts, and we consider the extraction correct if the listed pair is included in the results.

The prompt we used to extract PGx information from the label texts was "*Based on this content [texts for information extraction], answer the following questions step-by-step in short answers, only about the drug [drug name] and biomarker [biomarker name] as a pair. Then please generate a horizontal form table with the following items: Phenotypes/Genotypes: Identify the phenotypes (drug response influenced) or genotypes (genetic variants) associated with the biomarker. Frequency by Ethnicity: Provide the frequencies of the identified phenotypes or genotypes by ethnicity. Reason for PGx Labeling: State the reason for pharmacogenomic labeling of the biomarker. ADRs Associated with Biomarker: Identify adverse drug reactions related to the biomarker. Gender Differences: Indicate whether the biomarker is influenced by gender (Yes/No). Ethnicity Differences: Indicate whether drug response differs by ethnicity (Yes/No). Asian Stats: Provide the phenotype or genotype frequency of the biomarker in the Asian population. If no data is available, write 'No data.' Black/AA Stats: Provide the phenotype or genotype frequency of the biomarker in the Black population. If no data is available, write 'No data.' Hispanic Stats: Provide the phenotype or genotype frequency of the biomarker in the Hispanic population. If no data is available, write 'No data.' Polymorphism: Identify the genotype of the biomarker that influences drug response. Summary: Categorize the information using one or more keywords from 'Therapeutic Use,' 'Dosing,' 'Drug Response,' 'Metabolism,' and 'Ethnicity-Specific'.*"

Generation of mixed texts

To mimic the real-world scientific texts, which often discuss multiple drugs and biomarkers, we generated mixed texts by combining the labeling texts associated with two different drug-biomarker pairs from the FDA PGx Biomarker table. Each labeling text record was divided into five groups by randomly determining where to break the text, always ensuring the breaks occurred at the end of a sentence. This approach preserved the original sequence of sentences within each group. To create a mixed text, we selected these ten groups, five from each of two different segmented records, and merged them. This process allowed us to generate new, coherent mixed texts while blending information from two distinct drug-biomarker pairs (Supplementary Figure 1).

PubMed abstracts query

The PubMed API and Entrez library [16] were used to retrieve relevant abstracts based on a given drug-biomarker pair. We requested the title or abstract of one publication to contain both the drug and biomarker. To further narrow down the candidates to ensure the relevance of the collected abstracts, we also required that one of the keyworks, including PGx, pharmacogenomics, minority, variants, mutations, and population, be presented in either the title or abstract. Additionally, if no abstract could be found based on the initial query, we then searched for those abstracts that mentioned only the drug and biomarker. Considering the limitation of prompt

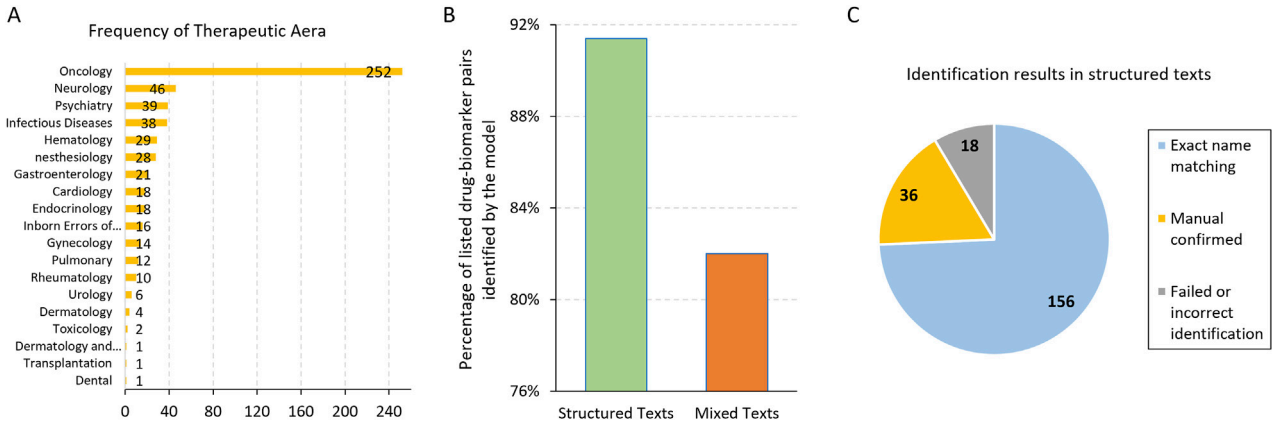


FIGURE 1
(A) The frequency of the Therapeutic Area in the FDA PGx Biomarker table. Majority of the records were related to Oncology. **(B)** The percentage of listed drug-biomarker pairs identified correctly by the model in structured and mixed texts, respectively. **(C)** The number and partition of the drug-biomarker identification results in structured texts.

length of the Llama3.1-70B model, we collected up to five abstracts for each PGx labeling record. The same prompt was used to extract PGx information from abstracts and from labeling texts.

Calculation of concordance rate

In this study, we used the concordance rate to measure the extent to which PGx categories (Therapeutic Use, Dosing, Drug Response, Metabolism, and Ethnicity-Specific) identified from the PGx labeling texts were also represented in the relevant abstracts for the same drug-biomarker pair. The concordance rate was calculated using the following formula:

$$\textit{Concordance rate} = \frac{\#\text{of PGx categories common to both PGx labeling texts and relevant abstracts}}{\#\text{of PGx categories identified in PGx labeling texts}}$$

This metric provided a clear and quantitative assessment of the overlap between the information in the PGx labeling texts and the scientific abstracts, allowing us to evaluate the consistency and completeness of the extracted data across different sources.

Results

High accuracy achieved with structured labeling texts in the FDA PGx biomarker table

We first evaluated the model's ability to identify drug and biomarker pairs from the labelling texts in the FDA PGx Biomarker table. Each entry contains the drug name, associated biomarker, therapeutic area, and labeling texts. Our analysis focused on the therapeutic area of Oncology, which had the largest number of records in the table (Figure 1A). We excluded records with non-gene biomarkers such as chromosome alterations or hormone receptors. As a result, out of 210 drug-biomarker pairs, the model successfully identified 192 pairs, achieving an identification accuracy of 91.4% (Figure 1B). Among these, 36 pairs required manual review and confirmation due to discrepancies arising from variations in nomenclature, such as the use of generic versus brand names of drugs or biomarker aliases. For example, the model identified the biomarker MKI67 as Ki-67, where MKI67 refers to the gene encoding the Ki-67 protein, indicating both terms represent the same entity. After manual validation, these 36 pairs missed by exact name matching were confirmed as correctly identified, contributing to the overall count of 192 accurate predictions (Figure 1C).

By manually reviewing the 18 records where the model failed to identify the drug-biomarker pairs, we found that most of them had short labeling texts in the FDA PGx Biomarker table, sometimes without the drug or biomarker even mentioned, leaving no way for the model to extract them. Another example was the drug brand name LONSURF, which was mentioned in the labeling text column of the PGx Biomarker table, but the listed drug names were tipiracil and trifluridine, the generic names of this drug. For this particular record, the model failed to identify either the brand or generic names.

Challenges with mixed texts

As Llama3.1-70B demonstrated high accuracy in identifying drug-biomarker pairs from a section of labeling text, we further challenged the model with mixed texts from two records. This

approach aimed to mimic the complex content often encountered in scientific studies, where discussions typically involve multiple drugs and biomarkers. To create a mixture testing set, we selected two records, each related to different drugs, and split them by sentences. These sentences were then merged to form a single paragraph, which was subsequently fed to the model (Methods, Supplementary Figure 1). This setup was designed to evaluate the model's ability to accurately extract relevant drug-biomarker pairs from a less structured and more intricate text, closely resembling real-world scientific documentation.

From the 156 records where the model correctly identified the drug-biomarker pairs without manual confirmation, we generated 50 mixture texts for testing (Methods). Using the same prompt and manual validation, we observed that the model could accurately identify at least one drug-biomarker pair for the testing records in 41 out of 50 (82%) cases (Figure 1B). Specifically, the model identified all the two drug-biomarker pairs in 32 records (64%), indicating a relatively high level of accuracy even with mixed and more complex text inputs. However, some cases posed significant challenges for the model. For instance, fusion names like BCR-ABL1 were occasionally difficult for the model to identify correctly. Additionally, there were instances where the model misidentified drugs due to the complexity of the text. In one particular case, a record included two drugs: ALIMTA (the brand name for pemetrexed) and pembrolizumab, which was mentioned as a comparator drug in the study. The primary drug for this record was pemetrexed, but the model incorrectly identified pembrolizumab as the paired drug. Notably, the drug-biomarker pair for this challenging case had been correctly identified in previous assessments without the interference of another record.

We further evaluated the mis-identified drug-biomarker pairs in the mixture texts by examining cases where the model incorrectly linked the drug and biomarker from two different records. As a result, ten mis-linked drug-biomarker pairs were identified from nine records. The results suggest that the presence of unrelated content may confuse the model, highlighting the need for careful consideration when handling complex and mixed information in texts.

Extraction of PGx information related to minority groups

Pharmacogenomics information is crucial for understanding how genetic variations influence drug responses across different population groups. Many PGx studies highlight the role that ethnic differences may play in drug efficacy and safety, with some of these findings reflected in labeling documents. Unique genetic profiles that may significantly impact responses to medications have been observed among minority groups, though these profiles remain underexplored. Despite growing awareness of genetic diversity, many minority populations continue to be underrepresented in PGx research, contributing to gaps in personalized medicine.

In this study, we collected 178 records from the FDA PGx Biomarker table containing terms such as "American," "Asian," "Caucasian." For each labeling text, we tasked the model to extract PGx information related to race or ethnicity. Key information extracted included the presence of ethnicity differences, frequency of genetic variants by ethnicity, reasons for PGx labeling, and adverse drug reactions (ADRs) associated with biomarkers (as detailed in Table 1). The model demonstrated its effectiveness by accurately identifying crucial details, such as the phenotypes of Poor Metabolizers (PM) and Extensive Metabolizers (EM) for the tolterodine-CYP2D6 pair. It correctly highlighted that the tolterodine labeling indicates approximately 7% of Caucasians and 2% of African Americans were poor metabolizers in that study. It is important to acknowledge that this labeling uses outdated terminology. The terms "White" and "Black/African American" are now preferred. This differentiation is vital for understanding the potential risks of adverse reactions, like QT prolongation, in specific populations (Table 1).

We assessed the model's accuracy in determining whether there were "ethnicity differences" in the labeling text column. The model was asked to answer a Yes/No question (Table 1) based on whether any information on ethnicity difference was found in the texts (Methods). Of the 178 records analyzed, 94 contained information explicitly stating ethnicity differences. However, some records mentioned the inclusion of diverse minority groups in studies but did not discuss or conclude any differences among these groups. For example, a labeling might state "*56 of the subjects were male, 61 were White, 20 were Black or African American, 8 were Hispanic or Latino*" but if no comparisons or outcomes were discussed, it should be marked as having no ethnicity difference.

We then manually reviewed the records classified by the model as having no ethnicity difference, identifying any false negatives. Impressively, the model achieved 100% accuracy in correctly identifying records that explicitly stated ethnicity differences. This finding underscores the model's reliability in detecting ethnicity-related PGx information and highlights the importance of ensuring accurate representation and consideration of minority groups in PGx research. This work illustrates the value of using LLMs to systematically and accurately identify PGx information across diverse populations. With appropriate data, LLMs have the potential to retrieve important PGx insights for minority groups from diverse published sources, contributing to more inclusive and equitable healthcare practices.

Validation of extracted PGx information

The extracted data, encompassing details about drug-biomarker pairs, genetic variations, and ethnicity-specific

TABLE 1 An example of the PGx information extracted from the FDA PGx Biomarker table related to the give drug-biomarker pair of Tolterodine-CYP2D6.

Pair	Tolterodine-CYP2D6
Phenotypes/Genotypes	Poor metabolizers (PM), Extensive metabolizers (EM)
Frequency by Ethnicity	Approximately 7% of Caucasians, approximately 2% of African Americans
Reason for PGx Labeling	Increased risk of QT prolongation and higher serum concentrations of tolterodine in poor metabolizers
ADRs Associated with Biomarker	QT prolongation, increased risk of cardiac arrhythmias
Gender Differences	No
Ethnicity Differences	Yes
Asian Stats	No data
Black/AA Stats	Approximately 2%
Hispanic Stats	No data
Polymorphism	CYP2D6 poor metabolizers have a slower rate of tolterodine metabolism, resulting in higher serum concentrations
Summary	Metabolism, Dosing, Drug Response, Ethnicity-Specific

information, plays a vital role in personalized medicine, which requires high accuracy. While verifying straightforward elements identified by the model, such as the presence or absence of ethnicity differences, is relatively easy, evaluating the detailed PGx information extracted from the texts is challenging due to its complexity. The intricacies involved in interpreting genetic data and its clinical implications require careful consideration. Manually verifying the extracted information would be impractical given the large volume and complexity of the data. Therefore, we implemented a systematic validation process using predefined PGx categories to evaluate the accuracy and consistency of the extracted information. This approach ensured a thorough and efficient assessment, allowing us to confirm the reliability of the model's outputs.

Particularly, when we tasked the model with extracting PGx information from the labeling texts in the FDA PGx Biomarker table, we also required a summary of each record using predefined keywords, including Therapeutic Use, Dosing, Drug Response, Metabolism, and Ethnicity-Specific (Table 1). For each ethnic PGx record, we collected up to five PubMed abstracts that contained the drug-biomarker pair in the title or abstract. To address concerns that abstracts might focus on different aspects and to narrow down the search to more relevant studies, we included additional keywords such as pharmacogenomics, PGx, and minority, in the PubMed query (Methods). This approach increased the chances of retrieving abstracts that provided the necessary PGx details, ensuring a thorough and focused validation process.

As a result, 137 out of 178 ethnic records had at least one abstract found in PubMed that contained the drug-biomarker pairs. The Llama3.1-70B model was then tasked again to tag each individual abstract with the predefined PGx information categories. By comparing the categories from the FDA PGx Biomarker table with those from the relevant abstracts, we evaluated the accuracy and consistency of the extracted information, ensuring alignment with external authoritative sources. A matched PGx category indicates that the particular drug-biomarker pair was studied by different research groups and that similar findings were concluded in the PGx field.

Among the 178 ethnic records in the FDA PGx Biomarker table, 125 discussed Drug Response, making it the most frequently mentioned category (Figure 2A). Additionally, we found a high consistency in that 78 out of 94 records (83%) identified with Ethnicity Differences were categorized as Ethnicity-Specific. In contrast, only 29 records were related to Dosing. However, the abstracts we collected, which involved the same drugs and biomarkers, exhibited different frequency patterns for these PGx categories (Figure 2A). The lower frequency of ethnicity-specific data in the abstracts suggests that this aspect may not be a major focus in the studies we collected.

We then calculated the PGx categories concordance rate, defined as the percentage of the categories identified in PGx labeling that were also covered by those from relevant abstracts. To assess the consistency of the extracted information, we compared the highest concordance rate based on a single abstract and the rate based on the aggregated abstract set. The median consistency was over 85% (Figure 2B), indicating high accuracy of the PGx information extracted by the LLM. This cross-validation not only confirms the reliability of the model's extraction capabilities but also highlights the robustness of our methodology in integrating and validating pharmacogenomic data across diverse sources.

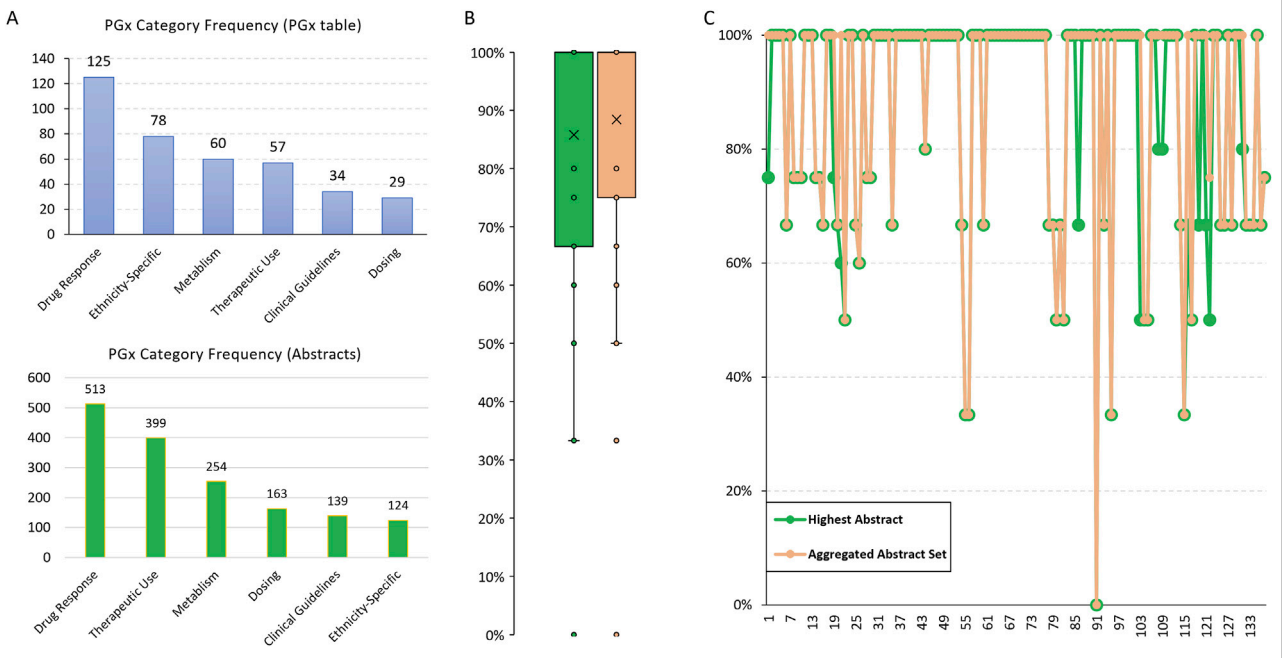


FIGURE 2
PGx categories summarized from the FDA PGx Biomarker table and relevant scientific abstracts. **(A)** The frequency of predefined PGx categories summarized by Llama3.1-70B for the 178 ethnic records from the FDA PGx Biomarker table. **(B)** The concordance rate of PGx categories between the FDA PGx Biomarker table and abstracts. The highest rate based on a single abstract and the rate based on an aggregated abstract set were compared. **(C)** A comparison of the highest and the aggregated concordance rate for each individual record.

TABLE 2 An example of the PGx information extracted from the FDA PGx Biomarker table and relevant abstracts collected from PubMed for Tolterodine-CYP2D6.

Pair	Tolterodine-CYP2D6
Phenotypes/Genotypes	Extensive metabolizers (EM), Poor metabolizers (PM), Intermediate metabolizers (IM), Ultra-rapid metabolizers (UM), Variants: CYP2D6*2, CYP2D6*10, CYP2D6*92, CYP2D6*93, CYP2D6*94, CYP2D6*95, CYP2D6*96, F164L, F219S, D336N, E215K
Frequency by Ethnicity	CYP2D6 poor metabolizers: 5-10% in Caucasians, 1-2% in Asians, No data for other ethnicities
Reason for PGx Labeling	Variability in metabolism leading to differences in drug response and adverse reactions
ADRs Associated with Biomarker	Increased risk of adverse reactions in PMs due to higher plasma concentrations. Decreased efficacy in UMs due to lower plasma concentrations
Gender Differences	No
Ethnicity Differences	Yes, CYP2D6 allele frequencies vary among ethnicities
Asian Stats	CYP2D6*10: 51.4% (Japanese), 40.9% (Chinese) * CYP2D6 poor metabolizers: 1-2%
Black/AA Stats	No data
Hispanic Stats	No data
Polymorphism	Variants: CYP2D6*2, CYP2D6*10, CYP2D6*92, CYP2D6*93, CYP2D6*94, CYP2D6*95, CYP2D6*96, F164L, F219S, D336N, E215K
Summary	Metabolism, Drug Response, Ethnicity-Specific

No big difference was observed between the highest and aggregated concordance rates (Figure 2C), suggesting that individual abstracts are sufficiently comprehensive in covering the relevant PGx categories. The approach we used successfully retrieved abstracts that were well-aligned with the information we were interested in from the FDA PGx Biomarker table,

ensuring that the abstracts are relevant and valuable for validating the PGx information.

The findings indicate that we can use these abstracts complementarily with the labeling texts to potentially extract additional PGx information for certain drug-biomarker pairs. As shown in Table 2, we asked the model the same questions based on the integrated texts of the four relevant abstracts (PMIDs: 28087463, 24619889, 22277677, 14606931). Additional PGx information associated with ethnic groups of Japanese and Chinese were found in these abstracts.

Discussions

While it is relatively straightforward to validate the extraction of certain PGx items from structured texts, such as drug and biomarker names from labeling sections, assessing the overall quality and completeness of the extracted information from more variable sources poses significant challenges. Unlike structured data, where predefined formats facilitate comparison and validation, publications and reports vary widely in focus and detail, complicating direct comparison of PGx information across different sources. To address this challenge, we employed a strategy where the model was instructed to tag the extracted texts with predefined categories, enabling a more systematic comparison. This tagging approach offers an initial method for aligning information across sources; however, we recognize that these categories may require further refinement or customization based on the specific content and objectives of different studies. Our results demonstrated that Llama3.1-70B achieved high accuracy in extracting drug and biomarker pairs from structured labeling texts, particularly when biomarkers were listed as gene or protein names. However, the model encountered difficulties when extracting less common biomarker names, such as "hormone receptors," which were excluded from the main analysis due to lower extraction accuracy. This limitation highlights the importance of prompt engineering and model tuning for specific use cases. Tailoring prompts to explicitly account for uncommon biomarkers or providing additional context within the prompt could improve the model's ability to accurately identify and extract these entities, an approach that warrants further exploration.

Identifying drug-biomarker pairs in mixed texts, where multiple records are combined, presents a more complex challenge for LLMs. Our study found that while Llama3.1-70B performed well with structured labeling texts, its accuracy decreased when processing mixed texts, likely due to the increased ambiguity and variety of content. This challenge would likely increase further with full-text publications, where drug-biomarker relationships are not always clearly delineated. To address these complexities, future studies could be benefit from a targeted approach, such as instructing the model to focus on specific drug-biomarker pairs to enhance extraction accuracy. In preliminary tests, the model was able to accurately identify relevant information from mixed texts when a specific drug-biomarker pair was targeted, suggesting that targeted prompts could improve accuracy in more complex texts.

Our findings demonstrate that LLMs like Llama3.1-70B can efficiently support the extraction of PGx information from structured sources, such as the FDA PGx Biomarker table, providing a foundation for integrating valuable data from scientific abstracts and potentially, with further refinement, from more complex sources like full-text publications. This automated approach can reduce the time and effort required for initial data extraction, improving the completeness of PGx databases by streamlining the process. However, we recognize that integrating LLM-extracted data directly into regulatory or clinical decision-making frameworks would require extensive validation and quality control, including human oversight, to ensure accuracy and relevance.

Implementing a structured workflow that leverages LLMs for routine extraction of PGx data could support the initial stages of database updates. Such a process would involve combining LLM-extracted insights with manual review and verification steps, enhancing the accessibility and usability of PGx data for non-regulatory applications, such as research and exploratory analyses in pharmacogenomics. This framework can be refined to incorporate more sophisticated validation methods, advancing the field of personalized medicine incrementally through a combination of automated and manual processes. Future work will focus on evaluating and refining this workflow to ensure reliability and utility in various PGx contexts.

While our study utilizes the Llama3.1-70B model, the primary focus of this work is the development of a generalizable framework for pharmacogenomic (PGx) data extraction. Our approach, which involves structured prompts, data integration techniques, and strategies for handling complex, mixed-text data, is designed to be adaptable to future advancements in LLM technology. As LLMs continue to improve, this framework can be applied to newer models, enabling consistent, automated PGx data extraction and updating without reliance on a specific LLM version. This flexibility makes the framework suitable for various applications in PGx research, supporting the evolving needs of pharmacologists, regulatory bodies, and healthcare researchers.

Author contributions

DL and JX conceived the project. DL, LW, and JX devised the experiments. DL, LW, Y-CL, H-YH, EC, QL, RC, RH, YZ, and JX contributed to the data analysis. Y-CL, H-YH, and EC manually

verified some of the results. DL and JX prepared the manuscript with the support from all co-authors. All authors read and approved the final manuscript.

Data availability

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding authors.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was funded by the U.S. Food and Drug Administration.

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.ebm-journal.org/articles/10.3389/ebm.2024.10393/full#supplementary-material

References

1. Haga SB, Burke W. Using pharmacogenetics to improve drug safety and efficacy. *Jama* (2004) **291**(23):2869–71. doi:10.1001/jama.291.23.2869

2. Johnson JA, Cavallari LH. Pharmacogenetics and cardiovascular disease—implications for personalized medicine. *Pharmacol Rev* (2013) **65**(3): 987–1009. doi:10.1124/pr.112.007252

3. Kalow W. Pharmacogenetics and pharmacogenomics: origin, status, and the hope for personalized medicine. *Pharmacogenomics J* (2006) **6**(3):162–5. doi:10.1038/sj.tpj.6500361

4. Micaglio E, Locati ET, Monasky MM, Romani F, Heilbron F, Pappone C. Role of pharmacogenetics in adverse drug reactions: an update towards personalized medicine. *Front Pharmacol* (2021) **12**:651720. doi:10.3389/fphar.2021.651720

5. Miteva-Marcheva NN, Ivanov HY, Dimitrov DK, Stoyanova VK. Application of pharmacogenetics in oncology. *Biomarker Res* (2020) **8**(1):32. doi:10.1186/s40364-020-00213-4

6. Ingelman-Sundberg M, Rodriguez-Antona C. Pharmacogenetics of drug-metabolizing enzymes: implications for a safer and more effective drug therapy. *Philosophical Trans R Soc B: Biol Sci* (2005) **360**(1460):1563–70. doi:10.1098/rstb.2005.1685

7. Meyer UA. Pharmacogenetics and adverse drug reactions. *The Lancet* (2000) **356**(9242):1667–71. doi:10.1016/s0140-6736(00)03167-6

8. Bienfait K, Chhibber A, Marshall JC, Armstrong M, Cox C, Shaw PM, et al. Current challenges and opportunities for pharmacogenomics: perspective of the industry pharmacogenomics working group (I-PWG). *Hum Genet* (2022) **141**(6): 1165–73. doi:10.1007/s00439-021-02282-3

9. Barbarino JM, Whirl-Carrillo M, Altman RB, Klein TE. PharmGKB: a worldwide resource for pharmacogenomic information. *WIREs Syst Biol Med* (2018) **10**(4):e1417. doi:10.1002/wsbm.1417

10. Nguyen Trinh Trung DH, Alexander S, Tanoli Z, Gloriam DE, Kooistra AJ, Caroli J, et al. *Pgxdb Figshare Softw* (2024). doi:10.6084/m9.figshare.26538574.v1

11. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A survey of large language models (2023). doi:10.48550/arXiv.2303.18223

12. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. *NPJ digital Med* (2022) **5**(1):194. doi:10.1038/s41746-022-00742-2

13. Murugan M, Yuan B, Venner E, Ballantyne CM, Robinson KM, Coons JC, et al. Empowering personalized pharmacogenomics with generative AI solutions. *J Am Med Inform Assoc* (2024) **31**(6):1356–66. doi:10.1093/jamia/ocae039

14. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, et al. Llama: open and efficient foundation language models (2023). doi:10.48550/arXiv.2302.13971

15. Meta AI. Llama3.1-70B model. Available from: https://ai.meta.com/blog/meta-llama-3-1/ and https://github.com/meta-llama/llama-models (Accessed September 1, 2024).

16. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* (2009) **25**(11):1422–3. doi:10.1093/bioinformatics/btp163

EBM Experimental Biology and Medicine

TYPE Original Research
PUBLISHED 21 May 2025
DOI 10.3389/ebm.2025.10445

OPEN ACCESS

*CORRESPONDENCE
Li Shen,
li.shen@pennmedicine.upenn.edu

†These authors have contributed equally to this work

RECEIVED 22 November 2024
ACCEPTED 23 April 2025
PUBLISHED 21 May 2025

CITATION
Wang Z, Chen J, Ionita M, Zhan Q, Zhou Z and Shen L (2025) Optimal transport reveals immune perturbation and fingerprints over time in COVID-19 vaccination.
Exp. Biol. Med. 250:10445.
doi: 10.3389/ebm.2025.10445

Optimal transport reveals immune perturbation and fingerprints over time in COVID-19 vaccination

Zexuan Wang[1†], Jiong Chen[2†], Matei Ionita[3†], Qipeng Zhan[1], Zhuoping Zhou[1] and Li Shen[4*]

[1]Graduate Group in Applied Mathematics and Computational Science, University of Pennsylvania, Philadelphia, PA, United States, [2]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA, United States, [3]Institute for Immunology and Immune Health, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, United States, [4]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, United States

Abstract

Mass cytometry enables high-throughput characterization of heterogeneous cell populations at single-cell resolution, using metal isotopes to capture cellular signals and avoiding the spectral overlap common in flow cytometry. Despite advancements, conventional data analysis often focuses on manual gating or clustering within specific samples, overlooking disparities across subjects or biological samples. To address this gap, we propose a novel framework that treats the cell-by-protein matrix as a high-dimensional distribution, using Quantized Optimal Transport (QOT) to quantify distances between samples based on their cellular protein expression profiles. This approach allows for a direct comparison of distributions without relying on predefined gating strategies, capturing subtle variations in the data. We validated our method through two experiments using real-world time-series Coronavirus Disease 2019 (COVID-19) cytometry data. First, we conducted a leave-one-out analysis to identify immunologically unstable proteins over time, revealing CD3 and CD45 as the proteins changing the most during the vaccine response. Second, we aimed to capture individual immune fingerprints over time by calculating pairwise Wasserstein distances between samples and applying hierarchical clustering. Using silhouette scores to evaluate clustering effectiveness, we identified optimal combinations of immunological markers that effectively grouped samples from the same participant across different time points. Our findings demonstrate that the QOT framework provides a robust and flexible tool for cohort-level analysis of mass cytometry data, enabling the identification of unstable immunological markers and capturing immune response heterogeneity among vaccinated cohorts.

KEYWORDS

optimal transport, COVID-19 vaccination, immunity, mass cytometry, fingerprint

Impact statement

Mass cytometry enables high-throughput characterization of cellular heterogeneity, but conventional analysis often focuses on manual gating or clustering within specific samples. We propose a novel quantitative framework that directly compares the high-dimensional protein expression distributions between samples using Quantized Optimal Transport. This approach captures subtle differences without relying on predefined gating strategies. Experiments on real-world COVID-19 cytometry data identified CD3 and CD45 as the most unstable proteins during the vaccine response. Furthermore, by calculating pairwise distances and applying hierarchical clustering, we determined optimal protein combinations that effectively grouped samples from the same individual over time, reflecting unique immune fingerprints. Our findings showcase the power of this framework for cohort-level mass cytometry analysis, enabling the discovery of key immunological changes and individual response patterns.

Introduction

Mass Cytometry (Cytometry by Time-Of-Flight) is a high-throughput technology to characterize heterogeneous cell populations in a single cell resolution [1]. As an advancement over traditional flow cytometry, mass cytometry utilizes isotopes instead of fluorophores to capture cellular signals, making a broader range of features available and avoiding the experimental difficulties related to spectral overlap [2]. In comparison with conventional single-cell RNA-seq experiments, mass cytometry also provides a higher throughput, which is capable of handling millions of cells along with a lower dimension of the cellular features derived from surface antigens, thus allowing more accurate capture of precise cell subpopulations [3]. Moreover, mass cytometry uses antibodies labeled with elemental heavy metal ions via chelating polymers to measure target proteins on single cells directly. In this method, stained cells are nebulized, vaporized, and ionized; the resulting ion cloud is mass-filtered to remove low-mass ions and then analyzed by time-of-flight mass spectrometry, precisely quantifying bound antibodies and revealing the expression of markers of interest, making it an ideal technique for monitoring the human immune system [4, 5]. The primary data analysis of mass cytometry experiments usually involves either manually separating cell subpopulations on a bivariate setting where the process is referred to as “manual gating” [6], supervised cell annotation trained by manual label [7–9], or via unsupervised clustering algorithms to group cells together [10–12]. However, these approaches often do not learn the disparities across subjects or biological samples, but they try to interpret the relationships of cells within a specific sample. Comparing the mass cytometry profile in a systematic resolution will also provide benefits in investigating global variation and differences [13, 14]. Characterizing and tracing the entire cell population would not only enable a more comprehensive understanding of how immune response varies systematically but also differentiate between samples and various cell subtypes in different diseases [15–18].

Optimal Transport (OT) is a mathematical framework originally proposed by Monge [19] and later reformulated by Kantorovich into a computationally tractable form [20]. OT addresses the challenge of comparing empirical distributions by finding the most efficient way to transform one distribution into another, ensuring mass preservation while minimizing an associated cost function. Recently, OT has been applied to mass cytometry data for automatic gating [21, 22].

Despite numerous algorithms developed for manual gating, practical methods for downstream analysis are still lacking. Traditional studies often compare disease states by focusing on the proportions of gated cell populations among cohorts to infer protein importance and disease-related protein expression [5, 23]. This approach may overlook differences in protein expression levels within cell populations. This work proposes a novel framework that treats the cell-by-protein matrix as a high-dimensional distribution, with each protein representing a dimension. By representing each sample as a distribution of cells across these protein dimensions, we can directly compare the distributions between cohorts using Optimal Transport. This allows us to quantify differences in protein expression profiles without relying on predefined gating strategies, capturing more nuanced variations in the data. Our main contributions are:

1. Quantifying Subject Differences via Quantized Optimal Transport: We introduce a method that utilizes Quantized Optimal Transport (QOT) to quantify the distance between subjects, viewing each cohort as a distribution of cells in high-dimensional protein expression space. This strategy can be applied with or without prior gating, providing flexibility in analysis.
2. Demonstrating Effectiveness on Coronavirus Disease 2019 (COVID-19) Cytometry Data: We validate our method through two experiments using real-world time-series COVID-19 cytometry data (Figure 1A). Specifically, we focus on (i) identifying immunologically unstable proteins over time (Figure 1B) and (ii) identifying informative proteins that contribute to fingerprint differentiation (Figure 1C). These case studies highlight the utility of our approach in revealing immune stability and heterogeneity of immune responses among vaccinated cohort.

Materials and methods

We evaluated our approach on a synthetic dataset—with three cohorts, each sampled at three-time points—and a real-

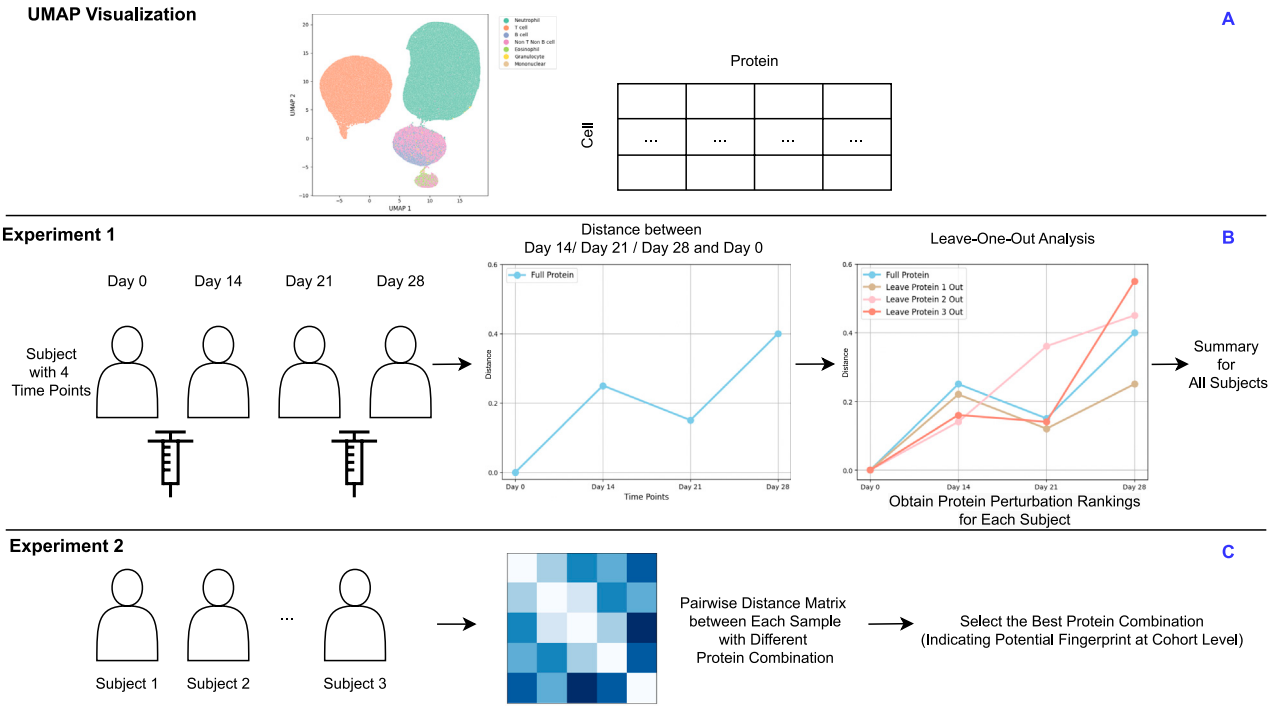


FIGURE 1
(A) UMAP visualization of a representative sample. **(B)** Experiment 1: For subjects measured at multiple time points, proteins are ranked by their level of perturbation within each subject. These rankings are then aggregated to identify proteins showing the greatest instability across the cohort. **(C)** Experiment 2: Each subject measurement at a specific time point is treated as an individual sample. Pairwise distance matrices are computed using various protein combinations to identify the optimal combination that effectively clusters samples. This optimal protein combination reflects the cohort-level fingerprint. **(A)** is based on real data, while **(B,C)** are derived from a synthetic design.

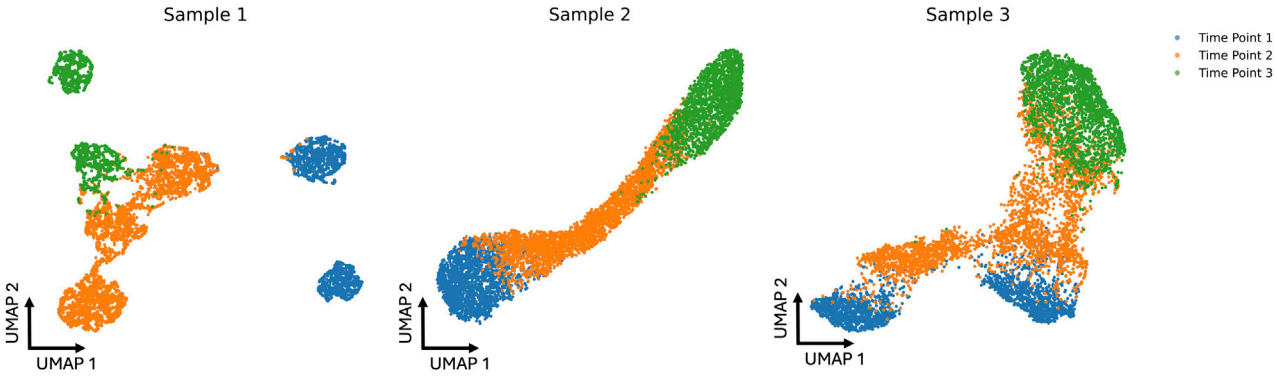


FIGURE 2
UMAP projections of three simulated cohorts at three different time points (blue, orange, and green). Each sample's distribution reveals distinct cell patterns that change over time, showing the dynamic shifts in the data across the simulated time course.

world mass cytometry dataset of single-cell protein expression in immune cells [24]. Additionally, in this paper, the term "cohort" refers to the entire dataset under study. To avoid confusion, we use the term "subject" to denote all time points belonging to the same individual. The term "sample" is used to refer to an individual file within the dataset, representing a specific time point for a individual.

In our synthetic dataset, we introduced distinct evolutionary patterns to capture heterogeneity within each subject. Specifically, Subject 1 follows a branched trajectory,

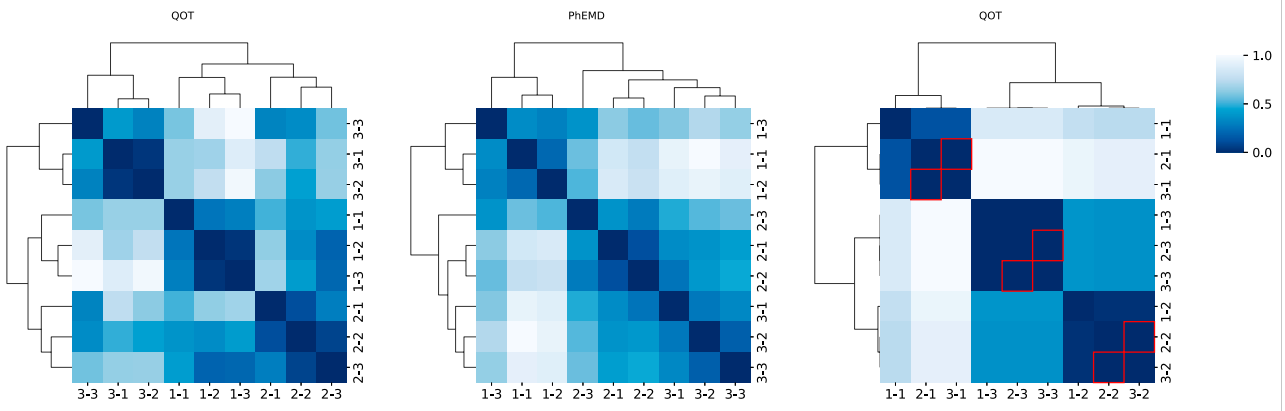


FIGURE 3
Heatmaps of pairwise distance matrices for three cohorts, computed by three methods (QOT, PhEMD, and PILOT). The label "1-1" indicates cohort 1 at time point 1. Each heatmap is hierarchically clustered, with darker shades signifying higher distances. In the PILOT panel, red boxes highlight entries where the distance is zero (excluding the diagonal).

TABLE 1 Performance comparison of QOT, PhEMD, and PULOT based on the Silhouette Score, Adjusted Rand Index (ARI), and Runtime.

Method	Silhouette score	ARI	Runtime (s)
QOT	0.529	1.00	5.26
PhEMD	0.429	1.00	1,440
PILOT	−0.340	−0.333	1.20

with separate cell populations diverging from Time 1 to Time 2, and again from Time 2 to Time 3. Subject 2 evolves along a smooth, curved progression, while Subject 3 exhibits a Y-shaped branching pattern, where all cells transition to new states from a common lineage. Each subject is characterized by 2, 3, and 2 cell types at Time 1, Time 2, and Time 3, respectively. Subject 1 has disproportionately sized cell types but a total of 7,000 cells across all time points. By contrast, Subjects 2 and 3 each maintain 1,000 cells per cell type, also yielding 7,000 cells in total. Further details on the cell-type proportions for Subject 1 can be found in Supplementary Appendix Table S1. The UMAP projection of the cohorts is shown in Figure 2.

For the real-word datasets, Whole blood was profiled from a cohort of 37 healthy subjects at multiple time points during two-dose mRNA vaccination against SARS-CoV-2. Each sample contains approximately 321 k cells. Most blood draws occurred at four standardized time points: a baseline draw before the first dose (T1), 2 weeks after the first dose (T2), before the second dose (T3), and a week after the second dose (T4). A few subjects had extra blood draws between T1 and T4 at intermediate time points. This yielded a total of 150 blood samples since not all subjects were available for each time point. The whole blood samples were stained with the Maxpar Direct Immunophenotyping Assay, a standardized panel for broad immunophenotyping of immune cell types. Finally, data was collected on a CyTOF2 instrument. Demographic and vaccination details is shown in Table S.2, S.3, S.4, S.5.

Quantized optimal transport

In this section, we briefly explain the Quantized Optimal Transport (QOT) method [25] for calculating distances at the sample level based on high-dimensional mass cytometry data.

Given a collection of P samples, denoted as $\mathcal{G} = \{G_1, G_2, \ldots, G_P\}$, each sample G_k is represented by an $n_k \times m$ matrix, where n_k is the number of cells in sample k, and m is the number of features (proteins). Our framework aims to compute the distance between two samples based on their cellular protein expression profiles.

We first model each sample as a distribution defined by its protein expression levels to compute the distance between two samples. This involves two main steps: (1) fitting a Gaussian mixture model (GMM) to each sample's data (Equations 1, 2) and (2) calculating the distance between the samples using their corresponding GMMs (Equations 3-9). For simplicity, we will use GMM as a short abbreviation for the Gaussian mixture model throughout the rest of this manuscript.

Each sample G_k is modeled as a GMM:

$$\omega_k = \sum_{h=1}^{H_k} \alpha_{k,h} \mathcal{N}\left(\mu_{k,h}, \Sigma_{k,h}\right), \tag{1}$$

where H_k is the number of Gaussian components in the GMM for sample k, $\alpha_{k,h}$ are the mixture weights satisfying

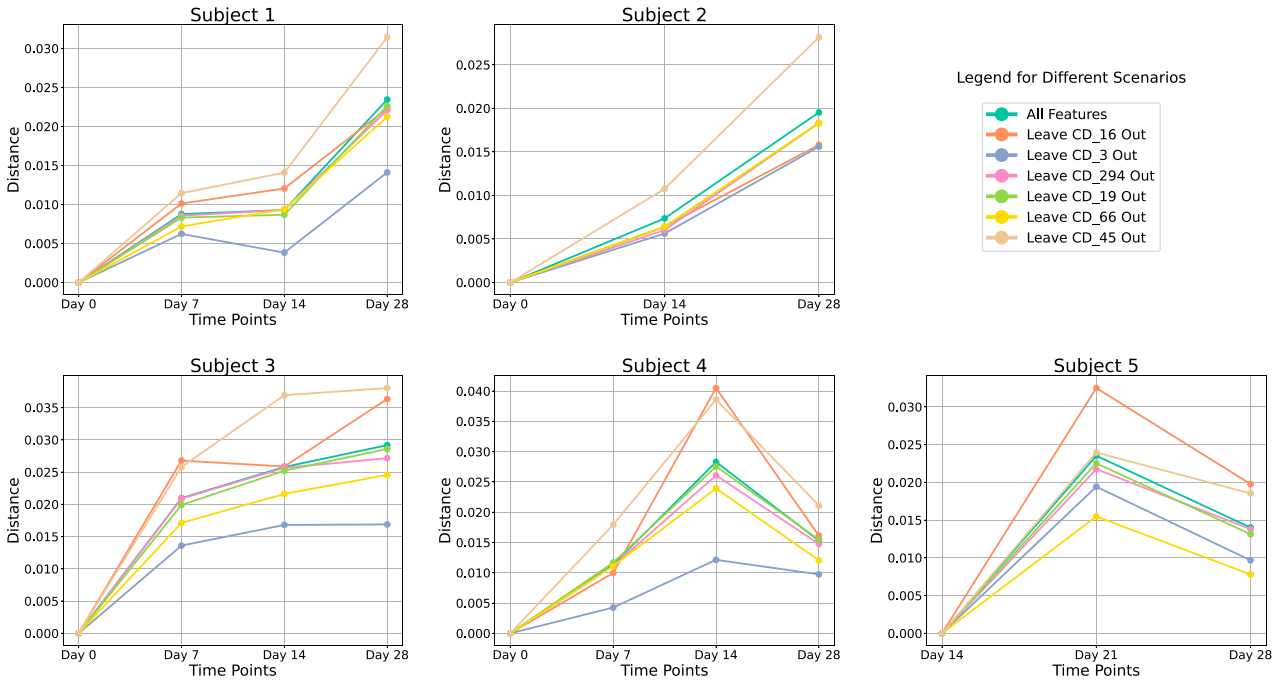


FIGURE 4
Leave-One-Out analysis at subject-level. Each line represents the perturbation in distance relative to the baseline at Day 0 over various time points. Different colored lines indicate the results using all proteins *versus* excluding a specific protein.

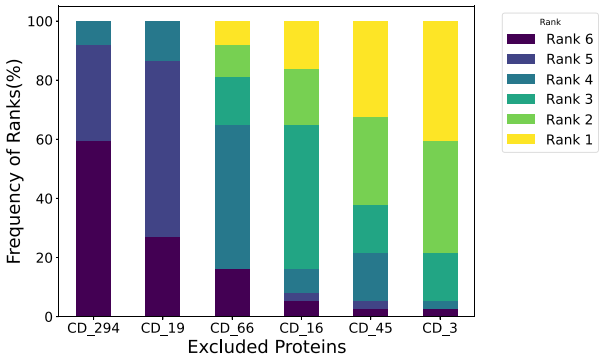


FIGURE 5
Leave-One-Out analysis at cohort-level. We assessed the impact of the subject when excluding each immunological protein and then summarized it at the cohort level. For instance, a 30% frequency at Rank 1 for CD 3 indicates that when measuring perturbation, perturbation caused by removing CD 3 ranks first in 30% of the subjects.

$$\sum_{h=1}^{H_k} \alpha_{k,h} = 1 \quad \text{and} \quad \alpha_{k,h} \geq 0, \tag{2}$$

$\mu_{k,h} \in \mathbb{R}^m$ are the mean vectors, and $\Sigma_{k,h} \in \mathbb{R}^{m \times m}$ are the covariance matrices of the Gaussian components. This approach allows the GMM to effectively encapsulate the distribution of the high-dimensional cytometry data for each sample.

Distances between cohorts are computed using the Wasserstein distance, quantifying the minimal cost of transporting one probability distribution into another. Specifically, we compute the Wasserstein distance between the GMMs representing the samples.

The distance between two samples, represented by their respective GMMs ω_i and ω_j, is computed by solving the following optimal transport problem:

$$\min_{T \in \mathbb{R}^{H_i \times H_j}} \sum_{p=1}^{H_i} \sum_{q=1}^{H_j} T_{pq} C_{pq}, \tag{3}$$

subject to the constraints:

$$\sum_{q=1}^{H_j} T_{pq} = \alpha_{i,p}, \quad \forall p = 1, \ldots, H_i, \tag{4}$$

$$\sum_{p=1}^{H_i} T_{pq} = \alpha_{j,q}, \quad \forall q = 1, \ldots, H_j, \tag{5}$$

$$T_{pq} \geq 0, \quad \forall p = 1, \ldots, H_i; \quad \forall q = 1, \ldots, H_j, \tag{6}$$

where T_{pq} represents the amount of mass transported from the p-th Gaussian component of ω_i to the q-th Gaussian component of ω_j, and C_{pq} is the cost of transporting unit mass between these components.

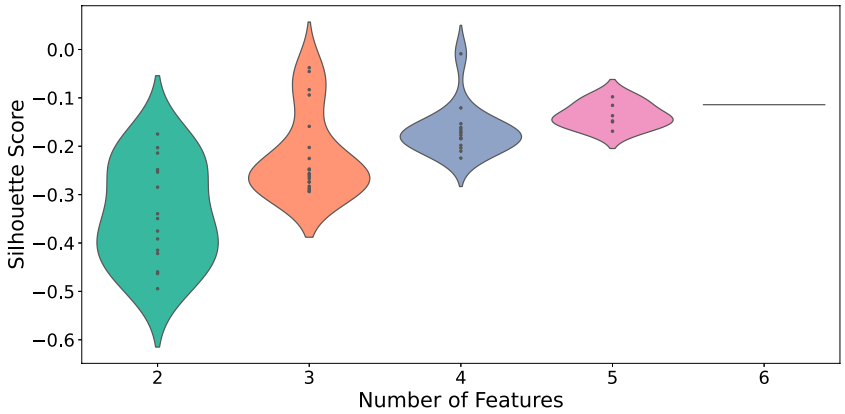


FIGURE 6
Distribution of silhouette scores across different feature counts for CD protein combina-tions. Each violin's width represents the density of silhouette scores for that feature count.

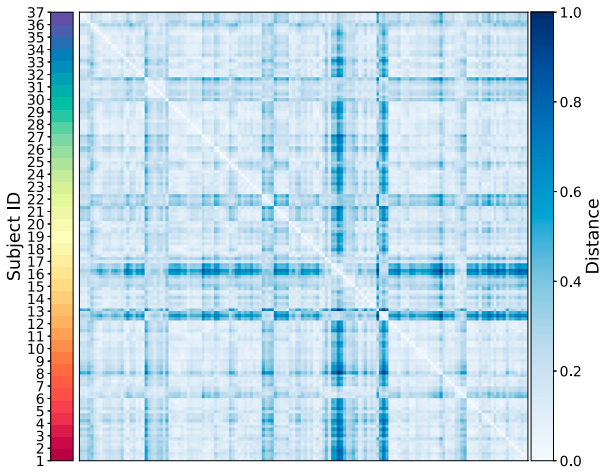


FIGURE 7
Heatmap of the sample-level distance matrix. Distance values are color-coded, with lighter shades of blue indicating closer proximity and darker shades representing greater distances. The color bar on the right provides the distance scale, while a second color bar on the left annotates subject IDs.

The cost matrix $C \in \mathbb{R}^{H_i \times H_j}$ has entries defined as:

$$C_{pq} = W_2^2\left(\mathcal{N}\left(\mu_{i,p}, \Sigma_{i,p}\right), \mathcal{N}\left(\mu_{j,q}, \Sigma_{j,q}\right)\right), \tag{7}$$

where W_2^2 denotes the squared Wasserstein distance between two Gaussian distributions. The squared Wasserstein distance between the Gaussian components is given by:

$$W_2^2\left(\mathcal{N}\left(\mu_{i,p}, \Sigma_{i,p}\right), \mathcal{N}\left(\mu_{j,q}, \Sigma_{j,q}\right)\right) = \|\mu_{i,p} - \mu_{j,q}\|^2 + \mathrm{Tr}\left(\Sigma_{i,p} + \Sigma_{j,q} - 2\left(\Sigma_{i,p}^{1/2}\Sigma_{j,q}\Sigma_{i,p}^{1/2}\right)^{1/2}\right), \tag{8}$$

where $\|\cdot\|$ denotes the Euclidean norm, $\mathrm{Tr}(\cdot)$ is the trace operator, and $\Sigma^{1/2}$ denotes the matrix square root of Σ. An alternative approach is to consider GMMs as point clouds instead of distribution, which provides scalability for larger-scale datasets. This approach involves calculating the cost matrix using the cosine distance between the centroids of Gaussian Mixture Models (GMMs):

$$C(p,q) = 1 - \frac{\mu_{i,p} \cdot \mu_{j,q}}{|\mu_{i,p}|2|\mu j, q|_2} \tag{9}$$

Experimental designs

Stability of cohorts across time

To identify immunologically unstable proteins across the subjects, we conducted a leave-one-out analysis to determine which proteins, when excluded, would result in the most perturbation in the immune profiles over time. This approach allowed us to assess the stability of each protein by measuring its impact on the temporal distributional similarity of immune profiles.

For each time point measurement within each subject, we first calculated the Wasserstein distance between the baseline time point (T1) and each subsequent time point (T2, T3, T4) using the full set of immunological proteins. This provided a reference measure of distributional change with all proteins included over time. Mathematically, the Wasserstein distance D_t^{full} between T1 and time point t (where $t \in \{\text{T2, T3, T4}\}$) was calculated as:

$$D_t^{\text{full}} = \mathrm{Dist}\left(G_{\text{T1}}, G_t\right)$$

where *Dist* denotes the QOT distance calculated in previous section, and G_{T1} and G_t represent the protein distribution profiles at T1 and time point t, respectively.

We then systematically excluded one protein at a time from the dataset. After excluding a protein, we recalculated the

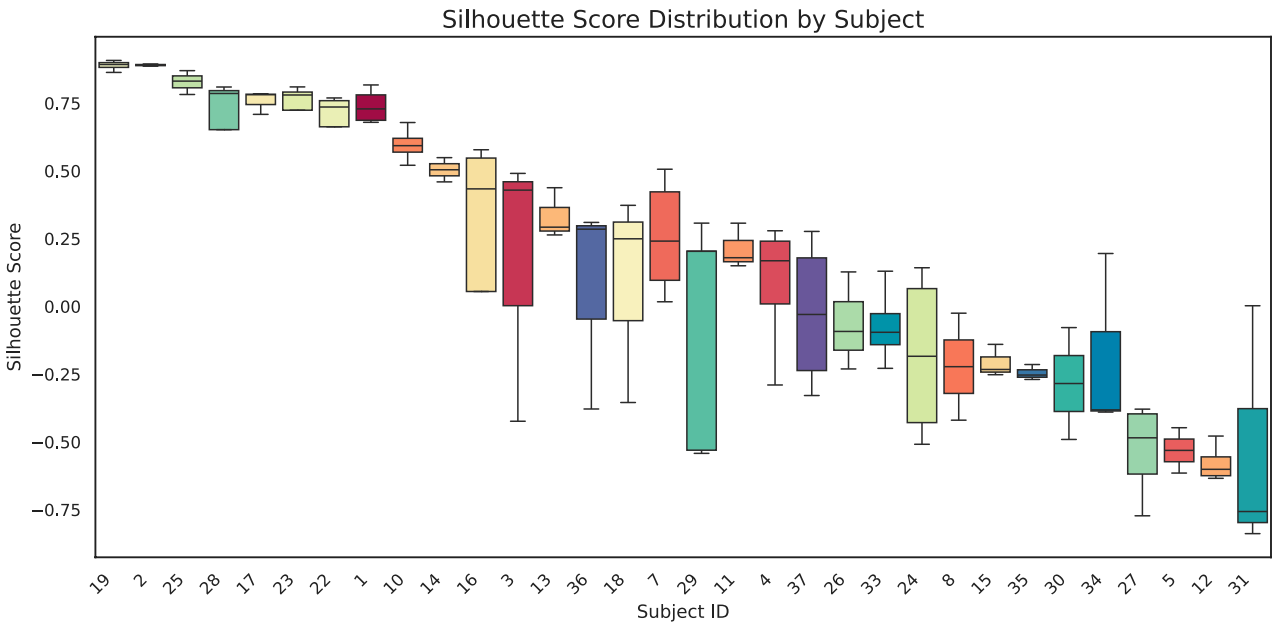


FIGURE 8
Box-and-whisker plots show the distribution of silhouette scores for each subject. Each box represents the interquartile range for that subject, with the black horizontal line indicating the median silhouette score. Higher (positive) values suggest more cohesive clustering, whereas lower (or negative) scores indicate overlap or unclear structure among clusters.

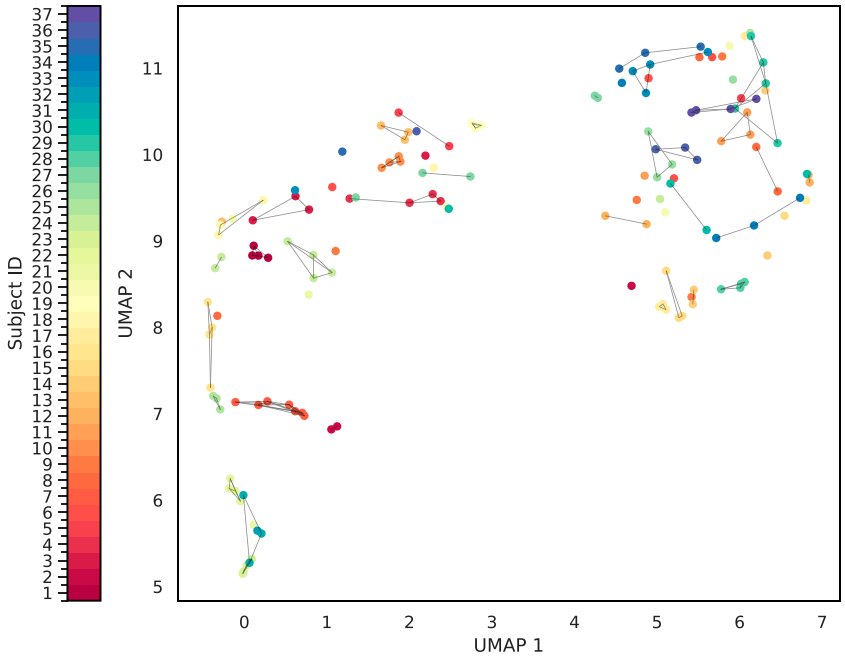


FIGURE 9
Cluster Analysis of the sample-level distance matrix. The UMAP representation of the sample level distance matrix. Points are connected based on whether belong to same sample and whether it is closed enough.

Wasserstein distances between T1 and each subsequent time point for each participant, obtaining D_t^{excl}. This process was repeated for every protein in the dataset, resulting in a set of perturbed Wasserstein distances corresponding to each excluded protein.

To quantify the perturbation caused by the exclusion of each protein, we calculated the absolute difference between the Wasserstein distances with all proteins included and with one protein excluded for each time point:

$$\Delta D_t = \left| D_t^{\text{full}} - D_t^{\text{excl}} \right|$$

We then summed these absolute differences across all time points to obtain a total perturbation score for each protein:

$$\Delta D_{\text{total}} = \sum_t \Delta D_t = \sum_t \left| D_t^{\text{full}} - D_t^{\text{excl}} \right|$$

a larger ΔD_{total} indicated that the excluded protein had a significant impact on the temporal stability of the immune profile, suggesting it is unstable protein over time.

We ranked the proteins for each subject based on the magnitude of ΔD_{total} their exclusion caused, from the least to the most perturbing. By aggregating these rankings across all subjects, we identified proteins that consistently resulted in the most perturbation when excluded. Proteins frequently ranked as causing the maximum perturbation across subjects were considered the most immunologically unstable over time.

Finger print of cohorts

In addition to identifying stable immunological features, we conducted a second experiment to capture subject immune fingerprints over time. The goal was to determine the optimal combination of immunological markers to effectively cluster samples from the same participant across different time points despite natural variations due to vaccination or immune fluctuations.

To achieve this, we calculated pairwise Wasserstein distances between all samples based on their immunological marker distributions, providing a quantitative measure of dissimilarity between samples. We then evaluate the effectiveness of different combinations of immunological protein expression with the silhouette score. The silhouette score assesses how well each sample fits within its assigned cluster compared to other clusters, offering a metric for the quality of the clustering solution. By testing various combinations of immunological protein expression and calculating the corresponding silhouette scores, we identified the feature sets that most effectively clustered samples from the same subject. In addition, we employed a UMAP visualization in which samples are connected if they meet two criteria: (1) they belong to the same group, and (2) their Euclidean distance is below a specified threshold (0.8).

Results

Cohort-level analysis of simulation dataset

In the cohort analysis of our simulation dataset, we compared QOT with two state-of-the-art approaches, PhEMD and PILOT, and examined their respective cohort-level distance matrices (Figure 3). Ideally, a well-structured distance matrix should exhibit a block diagonal pattern, where each block represents the same subject measured at different time points. Both QOT and PhEMD reveal these per-subject relationships clearly. In contrast, PILOT produces a mixed pattern: its hierarchical clustering intermingles different subjects, indicating it does not preserve the per-cohort structure. Moreover, PILOT assigns zero distances (highlighted by red boxes) for certain entries, suggesting identical samples. This misleading result arises from the methodology of PILOT. Specifically, PILOT first creates a uniform mask across all subjects and then considers only the proportions of cell types when computing pairwise distances. As a result, if two samples (e.g., Cohort2, Time1 and Cohort3, Time1) both contain the same set of cell types in identical proportions, PILOT assigns a zero distance, even if their expression levels differ substantially. Consequently, the uniform mask obscures critical differences in the data, failing to capture the true biological variability.

We quantitatively evaluated each distance matrix using the Silhouette score, Adjusted Rand Index (ARI), and runtime, as shown in Table 1. The Silhouette score assesses how well each sample is grouped within its own cluster and separated from others, while the ARI quantifies the agreement between true and predicted cluster assignments (with 1.0 indicating perfect alignment). Both QOT and PhEMD correctly distinguish different cohorts, achieving an ARI of 1.0. However, QOT produces a more pronounced cluster structure, reflected in a higher Silhouette score. In terms of computational efficiency, QOT completes in 5.26 s, compared to PhEMD's 1,440 s, demonstrating superior scalability for large-scale analyses. By contrast, PILOT fails to cluster cohorts correctly, often yielding misleading zero distances and not preserving the expected block-diagonal structure.

Cohort-level analysis of COVID-19 reveals immunologically unstable protein

In our cohort-level analysis of COVID-19, we aimed to identify immunologically unstable proteins across 37 healthy subjects. We employed a leave-one-out (LOO) approach, systematically excluding each protein to evaluate its contribution to immune perturbations over time. Figure 4 illustrates the subject-level LOO results, where each line traces the distance of a subject's sample at Day 7, 14, or 21 from its baseline (Day 0) under two conditions: using all available

features *versus* excluding a specific protein. The horizontal gap between these lines shows how strongly the excluded protein influences the observed perturbation. For instance, if removing CD16 produces a significant shift in distance relative to baseline, it implies that CD16 is a key driver of the subject's immune response over time; conversely, a negligible gap suggests that removing a protein has minimal effect and is more stable. Complete subject-level analyses are provided in the (Supplementary Appendix Figure SA1–SA3). From these LOO assessments, we found that subjects 1 through 4 showed CD3 and CD45 as their most unstable proteins, whereas subject 5's data highlighted CD16 and CD66 as the most variable over time. We then aggregated these subject-level findings to derive cohort-level insights, presented in Figure 5. Consistently, CD3 emerged as the most unstable protein across the overall cohort, followed closely by CD45.

Furthermore, our analysis indicates that removing CD45 leads to a higher distance from baseline. In other words, when CD45 is present, it helps keep the measured distance lower, suggesting a regulatory or stabilizing role. This finding aligns with the work of Hermiston et al., who showed that CD45 modulates signals from integrins and cytokine receptors [26], as well as Priest et al., who reported that CD45 expression on B cells shapes functional memory subsets post-vaccination [27]. By contrast, removing CD3 causes the distance from baseline to decrease, implying that including CD3 consistently drives the distance upward. This indicates that CD3 is a more perturbed protein in our dataset. Supporting this observation, Sattler et al. found that following SARS-CoV-2 vaccination, high-avidity spike-specific CD4 T cells lost surface CD3 expression after *in vitro* antigen restimulation, reflecting dynamic changes in T cell activation [28]. Similarly, Jaber et al. documented heightened CD3 T-helper cell responses in COVID-19 vaccine recipients [29], underscoring the pivotal role of CD3 in mediating immune perturbations in this setting.

Immune biomarkers for temporal fingerprint clustering

To identify Temporal Fingerprint Clusters across subjects, we treated each visit (timepoint) as an individual sample. Consequently, data from 37 healthy subjects resulted in 147 total samples for this analysis. Our working hypothesis is that, in an ideal scenario, samples originating from the same subject would naturally cluster together, reflecting each individual's inherent characteristics. We then evaluated combinations of proteins to determine which set yields the most informative clustering, as shown in Figure 6. We find combination of CD19, CD16, CD294, CD66b yiels highest silhouette score. We calculated distance matrices using subsets of these proteins—ranging from two to six proteins per subset. The most effective protein combination results, as indicated by the highest silhouette score, are illustrated in Figure 7. For visualization, we employed UMAP to project the distance matrix corresponding to the optimal silhouette score.

We quantitatively assessed clustering quality using the silhouette score, a well-established metric that compares each data point's average distance to others in the same cluster against its average distance to points in different clusters. Overall, we obtained a mean silhouette score of 0.156, suggesting that, while some structure is present, the clusters are not strongly separated on average. To explore subject-level variations, we also plotted the distribution of silhouette scores for each subject (Figure 8). Approximately one-third of subjects exhibit well-separated clusters, another third show moderately acceptable clustering, and the remaining subjects have less well-defined structures. Notably, although the low-dimensional representation in Figure 9 shows that different time points from the same subject can appear spatially grouped, the clusters themselves are not well separated across subjects. This observation aligns with the slightly lower silhouette score, which reflects both intra-cluster cohesion and inter-cluster separation.

Discussion

This study applied Quantized Optimal Transport (QOT) to analyze mass cytometry data from COVID-19 vaccinated cohort. Our approach uniquely avoids the biases of traditional gating by treating cell profiles as high-dimensional distributions. We demonstrated this method's utility in identifying unstable proteins like CD3 and CD45, which varied significantly over time, indicating their active roles in the immune response to vaccination. Additionally, our study demonstrates the use of optimal protein combinations to find immune fingerprints for subjects. By using silhouette scores for clustering optimization, we identified protein sets that consistently group samples from the same individual across different time points, highlighting its potential for personalized medicine.

For future work, we aim to refine our analytical framework for high-dimensional mass cytometry data, enhancing its capability to handle large-scale datasets effectively. In our initial experiment, we employed an exclusion analysis to assess protein importance. Integrating methods such as Shapley values with Wasserstein distances could significantly enhance interpretability. Additionally, our current analysis does not account for subclusters within the distance matrices. Investigating these subclusters could reveal new phenotypic subtypes related to vaccination responses, providing insights into immune system dynamics.

Author contributions

All authors participated in the conceptualization, methodology, validation, visualization, investigation, writing - original draft, writing - review and editing and formal analysis. ZW, JC, and LS Contributed to software. MI and LS contributed to resources and data curation. LS contributed to Supervision and funding acquisition. All authors contributed to the article and approved the submitted version.

Data availability

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving humans were approved by Immune Health, Perelman School of Medicine at the University of Pennsylvania. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported in part by NIH Grants R01 AG071470, U01 AG066833, and U01 AG068057.

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Generative AI statement

The authors declare that Gen AI was used in the creation of this manuscript. While preparing this work, the authors used ChatGPT 4 to help check the grammar.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.ebm-journal.org/articles/10.3389/ebm.2025.10445/full#supplementary-material

References

1. Tanner SD, Baranov VI, Ornatsky OI, Bandura DR, George TC. An introduction to mass cytometry: fundamentals and applications. *Cancer Immunol Immunother* (2013) **62**:955–65. doi:10.1007/s00262-013-1416-8

2. Spitzer M, Nolan G. Mass cytometry: single cells, many features. *Cell* (2016) **165**:780–91. doi:10.1016/j.cell.2016.04.019

3. Liu X, Song W, Wong BY, Zhang T, Yu S, Lin GN, et al. A comparison framework and guideline of clustering methods for mass cytometry data. *Genome Biol* (2019) **20**(1):297–18. doi:10.1186/s13059-019-1917-7

4. Wang W, Su B, Pang L, Qiao L, Feng Y, Ouyang Y, et al. High-dimensional immune profiling by mass cytometry revealed immunosuppression and dysfunction of immunity in COVID-19 patients. *Cell and Mol Immunol* (2020) **17**(6):650–2. doi:10.1038/s41423-020-0447-2

5. Rubin SJS, Bai L, Haileselassie Y, Garay G, Yun C, Becker L, et al. Mass cytometry reveals systemic and local immune signatures that distinguish inflammatory bowel diseases. *Nat Commun* (2019) **10**(1):2686. doi:10.1038/s41467-019-10387-7

6. Hartmann FJ, Bendall SC. Immune monitoring using mass cytometry and related high-dimensional imaging approaches. *Nat Rev Rheumatol* (2020) **16**(2):87–99. doi:10.1038/s41584-019-0338-z

7. Li H, Shaham U, Stanton KP, Yao Y, Montgomery RR, Kluger Y. Gating mass cytometry data by deep learning. *Bioinformatics* (2017) **33**:3423–30. doi:10.1093/bioinformatics/btx448

8. Cheng L, Karkhanis P, Gokbag B, Liu Y, Li L. DGCyTOF: deep learning with graphic cluster visualization to predict cell types of single cell mass cytometry data. *PLoS Comput Biol* (2022) **18**(4):e1008885. doi:10.1371/journal.pcbi.1008885

9. Chen J, Ionita M, Feng Y, Lu Y, Orzechowski P, Garai S, et al. Automated cytometric gating with human-level performance using bivariate segmentation. *bioRxiv* (2024):2024.05.06.592739. doi:10.1101/2024.05.06.592739

10. Bagwell CB, Inokuma M, Hunsberger B, Herbert D, Bray C, Hill B, et al. Automated data cleanup for mass cytometry. *Cytometry A* (2020) **97**(2):184–98. doi:10.1002/cyto.a.23926

11. Qiu P, Simonds EF, Bendall SC, Gibbs KD, Jr, Bruggner RV, Linderman MD, et al. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat Biotechnol* (2011) **29**:886–91. doi:10.1038/nbt.1991

12. Van Gassen S, Callebaut B, Van Helden MJ, Lambrecht BN, Demeester P, Dhaene T, et al. FlowSOM: using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry Part A* (2015) **87**(7):636–45. doi:10.1002/cyto.a.22625

13. Carr EJ, Dooley J, Garcia-Perez JE, Lagou V, Lee JC, Wouters C, et al. The cellular composition of the human immune system is shaped by age and cohabitation. *Nat Immunol* (2016) **17**(4):461–8. doi:10.1038/ni.3371

14. Tsang J, Schwartzberg P, Kotliarov Y, Biancotto A, Xie Z, Germain R, et al. Global analyses of human immune variation reveal baseline predictors of postvaccination responses. *Cell* (2014) **157**:499–513. doi:10.1016/j.cell.2014.03.031

15. Behbehani GK, Bendall SC, Clutter MR, Fantl WJ, Nolan GP. Single-cell mass cytometry adapted to measurements of the cell cycle. *Cytometry Part A* (2012) **81A**:552–66. doi:10.1002/cyto.a.22075

16. Diggins KE, Ferrell PB, Jr, Irish JM. Methods for discovery and characterization of cell subsets in high dimensional mass cytometry data. *Methods* (2015) **82**:55–63. doi:10.1016/j.ymeth.2015.05.008

17. Greenplate AR, McClanahan DD, Oberholtzer BK, Doxie DB, Roe CE, Diggins KE, et al. Computational immune monitoring reveals abnormal double-negative T cells present across human tumor types. *Cancer Immunol Res* (2019) **7**(1):86–99. doi:10.1158/2326-6066.cir-17-0692

18. Spitzer MH, Carmi Y, Reticker-Flynn NE, Kwek SS, Madhireddy D, Martins MM, et al. Systemic immunity is required for effective cancer immunotherapy. *Cell* (2017) **168**:487–502.e15. doi:10.1016/j.cell.2016.12.022

19. Monge G. Mémoire sur la théorie des déblais et des remblais. *Proc Lond Math Soc* (1781) **s1-14**:666–704. Available online at: https://scholar.google.com/scholar?hl=en&as_sdt=0%2C39&q=The+geometry+of+optimal+transportation&btnG=

20. Kantorovitch L. On the translocation of masses. *Management Sci* (1958) **5**(1):1–4. doi:10.1287/mnsc.5.1.1

21. Del Barrio E, Inouzhe H, Loubes JM, Matrán C, Mayo-Íscar A. optimalFlow: optimal transport approach to flow cytometry gating and population matching. *BMC bioinformatics* (2020) **21**:479–25. doi:10.1186/s12859-020-03795-w

22. Freulon P, Bigot J, Hejblum BP. CytOpT: optimal transport with domain adaptation for interpreting flow cytometry data. *The Ann Appl Stat* (2023) **17**(2): 1086–104. doi:10.1214/22-aoas1660

23. Hata K, Yanagihara T, Matsubara K, Kunimura K, Suzuki K, Tsubouchi K, et al. Mass cytometry identifies characteristic immune cell subsets in bronchoalveolar lavage fluid from interstitial lung diseases. *Front Immunol* (2023) **14**:1145814. doi:10.3389/fimmu.2023.1145814

24. Ionita M, Chen J, Greenplate A, Shen A. *Mass cytometry data with 5 independent manual annotations (Version 1)*. Philadelphia, PA: Pennsieve Discover (2024). doi:10.26275/864R-DV00

25. Wang Z, Zhan Q, Yang S, Mu S, Chen J, Garai S, et al. QOT: efficient computation of sample level distance matrix from single-cell omics data through quantized optimal transport. *bioRxiv* (2024):2024.02.06.578032. doi:10.1101/2024.02.06.578032

26. Hermiston ML, Xu Z, Weiss A. CD45: a critical regulator of signaling thresholds in immune cells. *Annu Rev Immunol* (2003) **21**:107–37. doi:10.1146/annurev.immunol.21.120601.140946

27. Priest DG, Ebihara T, Tulyeu J, Søndergaard JN, Sakakibara S, Sugihara F, et al. Atypical and non-classical CD45RBlo memory B cells are the majority of circulating SARS-CoV-2 specific B cells following mRNA vaccination or COVID-19. *Nat Commun* (2024) **15**(1):6811. doi:10.1038/s41467-024-50997-4

28. Sattler A, Gamradt S, Proß V, Thole LML, He A, Schrezenmeier EV, et al. CD3 downregulation identifies high-avidity, multipotent SARS-CoV-2 vaccine–and recall antigen–specific Th cells with distinct metabolism. *JCI insight* (2024) **9**:e166833. doi:10.1172/jci.insight.166833

29. Jaber HM, Ebdah S, Al Haj Mahmoud SA, Abu-Qatouseh L, Jaber YH. Comparison of T cells mediated immunity and side effects of mRNA vaccine and conventional COVID-19 vaccines administrated in Jordan. *Hum Vaccin and Immunother* (2024) **20**(1):2333104. doi:10.1080/21645515.2024.2333104

EBM Experimental Biology and Medicine

TYPE Original Research
PUBLISHED 21 May 2025
DOI 10.3389/ebm.2025.10468

OPEN ACCESS

*CORRESPONDENCE
Sangeeta Khare,
sangeeta.khare@fda.hhs.gov

†PRESENT ADDRESS
Vicki Sutherland,
J&J Innovative Medicine, Spring House, PA, USA

‡These authors have contributed equally to this work

RECEIVED 12 December 2024
ACCEPTED 28 April 2025
PUBLISHED 21 May 2025

CITATION
Muthumula CMR, Yanamadala Y, Gokulan K, Karn K, Cunny H, Sutherland V, Santos JH and Khare S (2025) Effect of in utero and lactational exposure to antiretroviral therapy on the gut microbial composition and metabolic function in aged rat offspring. *Exp. Biol. Med.* 250:10468.
doi: 10.3389/ebm.2025.10468

Effect of in utero and lactational exposure to antiretroviral therapy on the gut microbial composition and metabolic function in aged rat offspring

Chandra Mohan Reddy Muthumula[1‡], Yaswanthi Yanamadala[1‡], Kuppan Gokulan[1], Kumari Karn[1], Helen Cunny[2], Vicki Sutherland[2†], Janine H. Santos[2] and Sangeeta Khare[1*]

[1]Division of Microbiology, National Center for Toxicological Research, US Food and Drug Administration, Jefferson, AR, United States, [2]Division of Translational Toxicology, National Institute of Environmental Health Sciences, Research Triangle Park, NC, United States

Abstract

Despite the highly effective impact of antiretroviral therapy (ART) in reducing mother-to-child transmission of human immunodeficiency virus (HIV), there are concerns of long-term impacts of ART on the health of the offspring. The implications of perinatal exposure to antiviral drugs on the gut bacterial population and metabolic function in the offspring is unclear but may influence health outcomes given the various reported effects of the microbiome in human health. This study aims to gain insight into the potential effect of *in utero* and lactational exposure to ART on gut microbiota populations and short-chain fatty acids (SCFAs) production in aged rat offspring. Pregnant rats were administered a combination of antiretroviral drugs (abacavir/dolutegravir/lamivudine) at two different dose levels during gestation and throughout lactation, and the fecal bacterial abundance and SCFA levels of the offspring were analyzed when they reached 12 months of age. Our results showed dose-dependent and sex-based differences in fecal microbial abundance at various taxonomic levels. Specifically, we found a decline in *Firmicutes* in males, and an increase in *Actinobacteria* among males and females. Furthermore, a sex-specific distribution reorganization of *Lactobacillus*, *Bifidobacterium*, and *Akkermansia* was identified. No significant difference in the concentration of prominent SCFAs and IgA levels were identified. These findings provide preliminary information indicating the need to evaluate perinatal effects of ART more comprehensively on the gut bacterial and metabolic function in future studies, and their potential role in offspring health outcomes.

KEYWORDS

antiretroviral therapy, intestine, fecal, microbiome, short-chain fatty acids, IgA, perinatal exposure

Impact statement

This work addresses a critical health challenge on understanding how preventive HIV medications given during pregnancy could affect complex community of gut bacteria in next (f1) generation using a rat model. The work advances our understanding by analyzing both gut bacterial communities and their products in aged rat offsprings including sex-specific responses. This study showed that early exposure to these medicines lead to changes in gut bacterial composition. In addition, this effect also differed between male and female rats. However, their metabolic products (short chain fatty acids) and immune factors (IgA) remained stable. These findings impact the field by highlighting the importance of inclusion of male and female as a biological factor. This study provides a foundation for understanding how early exposure to HIV medications might influence long-term development and suggest new directions for monitoring offspring health.

Introduction

The gut microbiome consists of a wide network of microorganisms (including various types of bacteria, fungi, archaea, and viruses) that live in the gastrointestinal tract [1]. This gut microbiome plays a major role in regulating the health of an individual. This complex system is involved in various essential physiological processes including nutrient metabolism, development of the immune system, and protection against pathogenic microorganisms [2]. The composition and function of the gut microbiome is influenced by a variety of factors, including diet, use of antibiotics, and the host genetics [3]. When an unhealthy imbalance occurs in the gut microbial composition, it can lead to various metabolic diseases and health problems such as obesity, type 2 diabetes, and inflammatory bowel disease [4].

Various studies have shown the role of maternal microbiome as a key determinant of the offspring's gut microbiome composition and function [5–9]. During pregnancy and childbirth, microbial populations are passed from mother-to-child through vertical transmission with the mode of delivery (vaginal birth vs. cesarean section) and feeding practices (breastfeeding vs. formula feeding) influencing the initial colonization of the infant gut [5–9]. This early-life colonization of the microbial communities influences the long-term development of the offspring's health outcomes, such as the maturation of the immune system, development of metabolic pathways, establishment of the gut-brain axis, etc. [10–12]. Studies have shown that disruptions to this early microbial colonization have been linked to an increased risk of allergies, asthma, and metabolic disorders later in life [13].

Abacavir, dolutegravir, and lamivudine (a combination of three antiretroviral drugs) have been used in the management of Human Immunodeficiency Virus (HIV) infection in both adult and pediatric patients. This tri-combination antiretroviral therapy (ART) consists of drugs from 2 different classes: nucleoside reverse transcriptase inhibitors (abacavir and lamivudine) and HIV integrase inhibitors (dolutegravir) [14–16]. In HIV-infected pregnant women, ART is essential for preventing transplacental (mother-to-fetus) transmission of HIV infection [17–20]. The long-term side effects of ART may be under rated if the clinical trials utilize very specific inclusion/exclusion criteria, and the follow-up duration is relatively short [21]. Numerous studies suggest that ART may affect the composition and diversity of the gut microbiome [22, 23]. Alteration in the maternal microbiome could potentially influence the vertical transmission of microbial communities to the offspring [24]. In addition, HIV infection itself has been shown to disrupt the normal balance of microorganisms in the gut, characterized by a decrease in beneficial bacteria (such as *Bacteroides*) and an increase in potentially pathogenic ones (such as Prevotella) [25, 26]. While ART is essential for managing HIV infection, its effect on gut microbial diversity is not clear. Some studies suggest it helps restore diversity while other studies indicate that ART further disrupts the gut microbial diversity. Since many factors like immune health, diet etc., influence the microbiome, it is unclear whether ART may worsen or alleviate the alterations in the gut microbial composition on both the maternal and infant microbiome [25, 26].

In addition to the various physiological functions, gut bacterial populations are also involved in the production of various short-chain fatty acids (SCFAs). These SCFAs (such as acetate, propionate, and butyrate, etc.) are key microbial metabolites produced from the fermentation of dietary fibers [27]. SCFAs have been shown to exert various beneficial effects on host health, including the regulation of immune function, energy metabolism, and gut barrier integrity [28–30]. Various studies have shown that alterations in SCFA production lead to the development of various metabolic diseases, including cardiovascular disease, obesity, and type 2 diabetes [31–36]. ART-induced disruption of the microbial ecosystem may alter SCFA production, potentially influencing the development and function of the infant gut microbiome and leading to long-lasting health consequences for the offspring. However, the specific impact of gestational ART exposure on SCFA production in the offspring remains largely unexplored.

Fecal bacterial population profiling has emerged as a powerful non-invasive tool for assessing the composition and function of the distal gut microbial community [37, 38]. Various studies reported that high-throughput sequencing technologies such as 16S rRNA gene sequencing and shotgun metagenomics, are useful to obtain a comprehensive snapshot of the microbiome composition and its functional potential [39, 40]. Furthermore, integrating metabolomic analysis such as SCFA quantification, with taxonomic profiling has provided valuable insights into the complex interplay between the gut microbiome and host physiology [41, 42]. However, limited studies have applied these multi-omics approaches to investigate the long-term impact of perinatal ART exposure on the offspring gut microbiome and metabolome.

In addition to the complex interplay between the gut microbiome and ART, it is also important to consider the role of the immune system in shaping the microbial composition. Immunoglobulin A (IgA), plays an important role in maintaining the delicate balance between the immune system and the gut microbiota of the host, ensuring a mutually beneficial relationship [43]. Investigating the interplay between gut microbiome, IgA, and SCFAs in the context of perinatal ART exposure could provide valuable insights into the balance between the immune system and microbial metabolites that could influence the offspring's health.

The current study tested whether indirect exposure through the Dams to Abacavir Sulfate (ABC)/Dolutegravir Sodium (DTG)/Lamivudine (3TC), hereon called TC-ART, led to changes in the gut microbiome when the offspring reached 1 year of age. In addition, changes in the abundance and activity of SCFA-producing bacteria resulting in altered SCFA profiles in the offspring, were examined.

Rationale for selection of antiretroviral treatment regimen

The drug regimen consisting of abacavir, dolutegravir, and lamivudine, was selected for this study based on being the current recommended TC-ART to be provided during pregnancy for patients that are naïve to ART or already on this combination[1]. For adults and children weighing 25 kg or more, this TC-ART is administered at a dosage of 600-50-300 mg in tablet form once daily [44, 45]. The selection of this once-daily combination therapy in the rat model is expected to mimic the dosing schedule used in human patients.

Materials and methods

Animal housing, care, treatment, euthanasia, and sample collection

Time mated Sprague Dawley rats (Hsd:SD) were obtained from Envigo (Indianapolis, IN). Pregnant rats and their male and female offspring were housed in the animal facility at Amplify Bio, West Jefferson, OH, an independent, scientific contract research organization. The facility's Institutional Animal Care and Use Committee (IACUC) reviewed the protocol and approved it. The IACUC number for this protocol is T06055. All rats were housed in polycarbonate cages with irradiated hardwood bedding chips (Sani Chips®; Envigo, Madison, WI). Natural crinkled kraft paper was provided during gestation and lactation for enrichment (Crink-l'nest™, The Andersons, Maumee, Ohio). Offspring remained with their respective dams until postnatal day (PND) 21. After the lactation period, first generation offspring were provided polycarbonate rectangular shelters (Rat Retreats™, Bio-Serve, Flemington, NJ) as enrichment and were group housed by sex, up to 5 per cage. Animals were fed irradiated NIH-07 pellets or wafers (Zeigler Bros., Gardners, PA) during gestation and lactation. After weaning, rats were fed NTP-2000 (Zeigler Bros., Gardners, PA). All animals were provided municipal water *ad libitum* from an automatic watering system. The water and feed were analyzed for known contaminants that could interfere with or affect the outcome of the study, and none were found. The animals used in this study of microbiome were part of a larger toxicology study that will be reported separately. The experimental design for the microbiome investigation is outlined in the Figure 1.

Pregnant Sprague Dawley rats (n = 5/group) were exposed via gavage to two different doses of the TC-ART (abacavir/dolutegravir/lamivudine) during gestation and lactation (GD6 - PND21). The doses of TC-ART used in this study were a low-dose of 150/12.5/75 mg/kg body weight and a high-dose of 300/25/150 mg/kg body weight. All animals, including the control group, were administered vehicle solution (0.2% methylcellulose/0.1%, Tween 80) at the same volume (5 mL/kg) and frequency as treatment groups. The offspring were indirectly exposed to the ART via the dam during their perinatal period only. The offspring (one male and female from each dosed or control dam) were aged to 12 months with no direct dosing. At the age of 12 months, these aged rats were sacrificed. Fecal samples were collected aseptically from the colon of animals and immediately transferred to liquid nitrogen and thereafter kept frozen at −80°C for the assessment of gut microbiota, SCFAs, and fecal IgA (bound and unbound) (Figure 1).

Fecal DNA extraction and long read sequencing

The DNA and RNA from the rat fecal samples were isolated using Zymo ZR-duet DNA/RNA Miniprep (Zymo Research, Tustin, CA, United States) as per manufacturer instructions. The quality and quantity of DNA was checked using a Cytation3 Cell Imaging Multimode Reader (BioTek, Winooski, VT, United States) and QubitTM Fluorometer (Thermo Fisher Scientific, Waltham, MA, United States).

Microbiome sequencing and analysis

To investigate the fecal microbiota composition of control, low-dose, and high-dose treated groups in both male and female

1 https://clinicalinfo.hiv.gov/en/guidelines/perinatal/whats-new

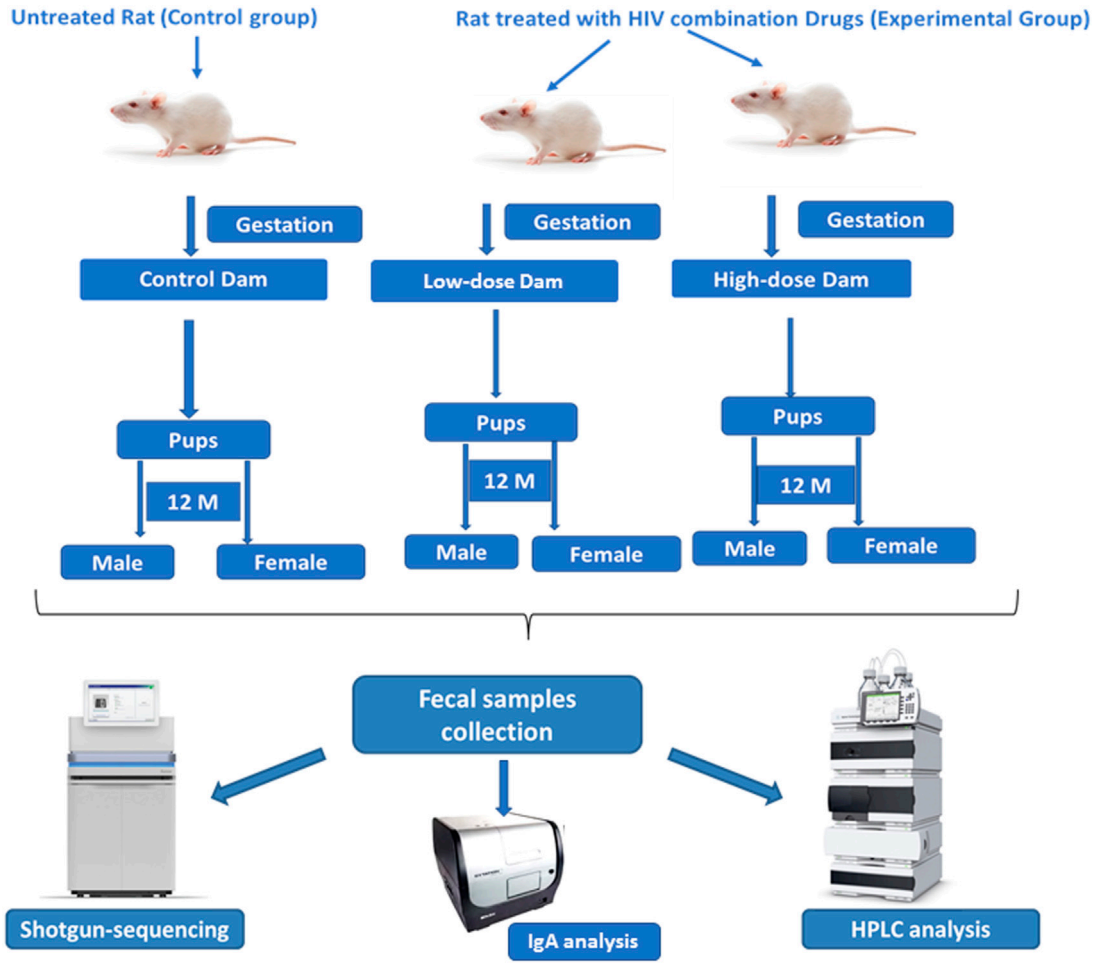


FIGURE 1
Experimental design for evaluating the effects of perinatal exposure of the HIV Tri-combination drug in a rat model (F1 generation). Schematic representation of the experimental design for 12-month-old F1 generation study shows that the pregnant female rats (dams) were divided into control (untreated) and experimental groups (receiving low- and high-dose HIV tri-combination drugs) during gestation till post-partum day 21. Male and female offspring were monitored for 12 months (F1 generation). Fecal samples were collected from all the offspring and analyzed using shotgun sequencing, HPLC, and IgA profiling to access microbiome composition, metabolite profiles, and mucosal immunity.

rats, Illumina NovoSeq sequencing technology was utilized. Operational Taxonomic Units (OTUs) were generated for each sample using a 97% sequence similarity threshold, and the number of sequences in each OTU was determined. The OTU representative sequences were compared against a microbial reference database to obtain classification information for each species corresponding to each OTU. Microbiome diversity and community structure were assessed via shotgun sequencing, using libraries prepared with a procedure adapted from the Nextera XT Kit (Illumina). Sequencing was performed on an Illumina NovaSeq 6000 platform, with paired-end 2 × 150 sequencing and a target depth of 20 million reads. DNA sequences were filtered for low quality (Q-Score <30) and length (<50 bp), and adapter sequences were trimmed using Cutadapt. Host sequences were removed using Bowtie2. Bacterial 16S rRNA gene sequences were extracted from the shotgun data and used for the microbiome analysis. Using the web-based platform MicrobiomeAnalyst [46, 47]. The Greengenes database was used for taxonomic classification. Data filtering included the removal of low-count features with a minimum count of 4 and a prevalence of 20% in samples, as well as low-variance features with a 10% cutoff. Data normalization was performed using Total Sum Scaling (TSS). Rarefaction curves were used to evaluate sequencing depth and Good's index was used to assess sequencing completeness. Alpha diversity was assessed using 2 metrics: Chao1 and Shannon index. Analysis of variance (ANOVA) was used to determine statistically significant differences in microbial community diversity in response to TC-ART treatment. Beta diversity was analyzed using Principal Coordinate Analysis (PCoA) based on Bray-Curtis

dissimilarity. Similarity analysis was conducted using Euclidean distance and the Ward hierarchical clustering algorithm, with results presented in a heatmap.

Fecal sample collection and processing for SCFA using HPLC

For SCFA profiling, 100 mg of frozen feces was weighed out into microcentrifuge tubes. One milliliter of chilled HPLC-grade water was added to the respective tubes. The samples were vortexed for 1 min and sonicated for 10 min until homogenized. The homogenized samples were then centrifuged at 18,000 × g for 10 min at 4°C. The supernatant was collected and filtered through a 0.22 µm syringe filter into amber HPLC vials.

Quantification of SCFA

SCFAs were quantified using high-performance liquid chromatography (HPLC) analysis. An Agilent Technology 1260 Infinity system coupled with an Agilent Technology Infinity Lab LC/MSD mass spectrophotometer and an auto sampler system was used for the analysis. Chromatographic separation and identification of SCFAs were performed using an Aminex HPX-87H column (300 mm × 7.8 mm, hydrogen form, 9 µm particle size, 8% cross-linkage; Bio-Rad) maintained at 65°C. A UV detector set at 210 nm using a spectral diode array system was employed for detection. The mobile phase consisted of freshly prepared 2.5 mM H_2SO_4 with a flow rate of 0.6 mL/min. The sample injection volume was set to 10 µL.

Calibration standards were prepared by diluting the respective reference standards for the following SCFAs: succinic acid, lactic acid, formic acid, acetic acid, propionic acid, isobutyric acid, butyric acid, isovaleric acid, valeric acid, hexanoic acid, and heptanoic acid in 2.5 mM H_2SO_4. Standards, samples, and spiked samples were analyzed by HPLC, and SCFAs were identified and quantified by retention time and peak area relative to the standards. The percentage recovery of the SCFAs from extraction ranged between 80.83 and 92.15%.

Quantification of bound and unbound IgA in rat feces

To quantify the level of bound and unbound IgA in rat feces, we followed the procedure described by Lahiani at al [48]. Rat fecal samples were weighed and diluted to prepare a concentration of 50 mg/mL PBS buffer containing 1mM phenylmethylsulphonyl fluoride (PMSF) and 1mM protease inhibitor cocktail solution. The samples were then vortexed rigorously and centrifuged at 4°C for 15 min at a speed of 900g. The resulting supernatant was collected and filtered through 0.22 um PTFE syringe filters to measure the unbound IgA level.

For the collection of bacterial-bound IgA, the same filter was washed with 0.05% tween 20, and the flow through was collected. A Rat IgA ELISA Kit (Bethyl Laboratories, Montgomery, TX, United States) containing pre-coated 96 well strip plate was used to assess the levels of IgA according to the manufacturers protocol. The absorbance was measured on a Cytation 3 plate reader (BioTek) at 450 nm. The standard curve was fitted into a 4-parameter curve fitting equation to calculate the analyte concentration in the original sample.

Statistical analysis of SCFAs and IgA levels

For comparing SCFAs and IgA levels between treatment groups, an unpaired two-sample t-test was performed. The significance was set at 5% ($p \leq 0.05$). The t-test assessed differences between the means of the data sets by calculating the variance from all animals in each group ($n = 5$). P-values below 0.05 were considered statistically significant.

Results

Diversity analysis of fecal microbial communities across treatment groups

Alpha diversity analysis

Alpha diversity measures the richness and diversity of species within a single sample using various indices, such as Chao1 and Shannon. The Chao1 index assesses species richness (i.e., the number of species), while the Shannon index evaluates species diversity, considering both richness and community evenness. In this study, the completeness of sequencing was tested using Good's coverage, which reached 100%, indicating that the majority of the bacterial species present in the samples had been detected.

The rarefaction curves of the observed OTUs (Figure 2) revealed that the number of OTUs increased with sequencing depth for all treatment groups. In females, the control (26–39 OTUs), low-dose (23–41 OTUs), and high-dose (17–47 OTUs) groups all showed substantial overlap, preventing clear discrimination of diversity changes due to treatment. Similar results were obtained for males: the control (32–51 OTUs), low-dose (25–38 OTUs), and high-dose (31–42 OTUs). The stabilization of the final curve indicates that the amount of sequencing data obtained was sufficient and representative.

The Chao1 index values (Figure 3A) ranged from 18 to 52 across all samples, with the highest value observed in the female high-dose group (HF_26) and the lowest in the female high-dose group (HF_30). No significant differences were

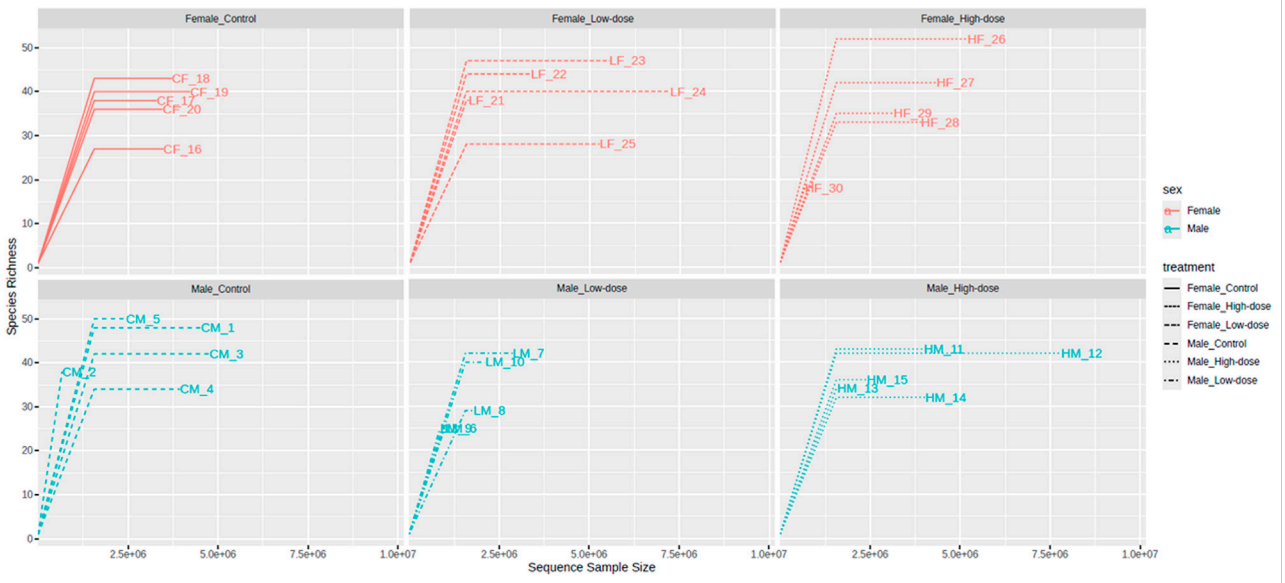


FIGURE 2
Rarefaction curves of the observed Operational Taxonomic Units (OTUs) in the fecal microbiome across different treatment groups (control, low-dose, and high-dose) in both male and female rats. The rarefaction curves plot the number of observed OTUs as a function of the number of sequences sampled, with a plateauing curve indicating that the majority of the bacterial species present have been captured.

observed when comparing the Chao1 index across groups, suggesting that the treatment did not significantly impact species richness in either male or female rats.

The Shannon index values (Figure 3B) ranged from 1.11 to 2.99, with the highest value found in the male control group (CM_5) and the lowest in the male control group (CM_4). Similar to the Chao1 index, no significant differences were observed in the Shannon index across treatment groups.

Collectively, these data demonstrate that TC-ART treatment did not alter species richness or diversity in aged male and female rats perinatally exposed to these drugs.

Comparison of fecal microflora across treatment groups

The identified bacteria were categorized into 7 phyla, 11 classes, 14 orders, 26 families, 39 genera, and 68 species across the animals. The composition of each sample community was calculated at every taxonomic level (phylum, class, order, family, genus, and species). Table 1 represents the taxonomic level classification of individual samples.

Phylum level analysis

At the phylum level (Figure 4), differences in relative abundance were observed between males and females in the control group (compare first and fourth bar on Figure 4), and further changes were observed upon treatment. Specifically, the relative abundance of both *Firmicutes* and *Bacteroidetes* exhibited a dose-dependent decrease in males, with more pronounced effect in the high-dose group. Conversely, a dose-dependent increase was observed for *Actinobacteria*. In females, changes in relative abundance did not follow dose-dependency, with *Firmicutes* decreasing in the low-dose but increasing in the high-dose group compared to the control. Similarly, *Actinobacteria* relative levels were higher in the low-dose than the high-dose group while *Verrucomicrobia* increased in the low-dose group but decreased in the high-dose group. *Bacteroidetes* decreased in the low-dose group but increased in the high-dose group.

Taken together, the phylum-level analysis suggests that perinatal TC-ART treatment has long-term influences in the gut microbiome composition, which is different between males and females.

To elucidate if the changes seen at the phyla level is also translated into the genus level, comparative analysis on the bacterial abundance at the genera level was conducted.

Genus level analysis

At the genus level (Figure 5), the relative abundance data revealed notable differences between control and TC-ART treated groups in both males and females. In the control groups, *Lactobacillus*, *Akkermansia*, *Bifidobacterium*, and *Bacteroides* were among the most abundant genera. However, the relative abundances of these genera were altered upon treatment with TC-ART.

In males, the relative abundance of *Lactobacillus* decreased in both low-dose and high-dose treatment groups compared to the

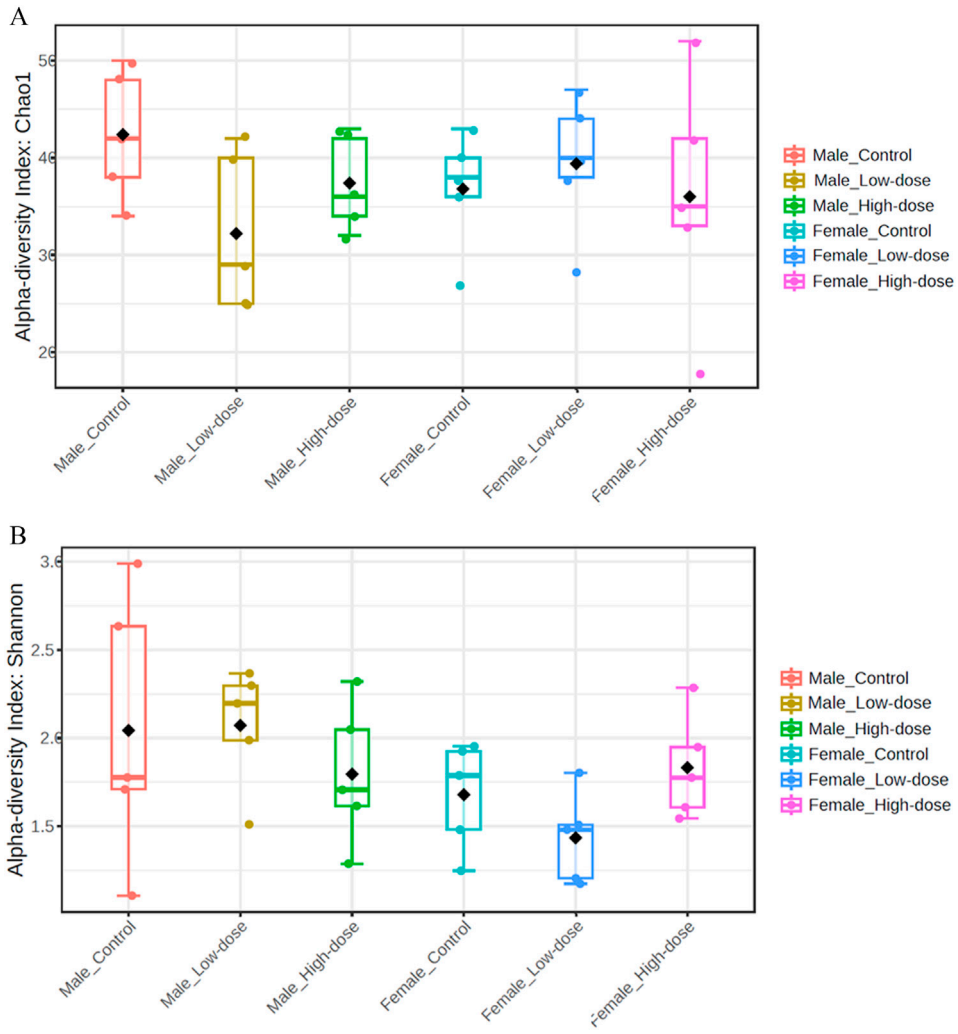


FIGURE 3
Fecal microbial population diversity across different treatment groups (control, low-dose, and high-dose) in both male and female rats. **(A)** Alpha diversity indices (Chao1) assess species richness, with higher values indicating a greater number of unique species within a sample. **(B)** Alpha diversity indices (Shannon) of the fecal microbiome across different treatment groups (control, low-dose, and high-dose) in both male and female rats assess species diversity, taking into account both richness and evenness, with higher values indicating greater diversity within a sample.

control. *Akkermansia* showed a slight increase in the low-dose group but decreased in the high-dose group. *Bifidobacterium* exhibited an increase in the treatment groups, with the high-dose group showing the highest relative abundance. *Bacteroides* and *Parabacteroides* decreased in both treatment groups.

In females, the relative abundance of *Lactobacillus* decreased in both low-dose and high-dose treatment groups, with the low-dose group showing a substantial decrease compared to controls. *Akkermansia* decreased in both treatment groups, with the high-dose group having the lowest relative abundance. *Bifidobacterium* increased in both treatment groups, with the low-dose group showing the highest relative abundance. *Bacteroides* and *Parabacteroides* increased in the high-dose group compared to the control.

The genera-level analysis reveals the differential impact of TC-ART on specific genus within the gut microbiome of males and females. The observed changes suggest that the drug modulates the relative abundances of key genera, such as *Lactobacillus*, *Akkermansia*, and *Bifidobacterium*, in a sex-specific manner. These alterations in genus-level composition contribute to the overall shifts observed at the phylum level.

To gain more granularity at the taxonomic level, the impact of TC-ART drug on different treatment groups was assessed by the heatmap analysis and Principal Coordinate Analysis (PCoA) at the species level.

TABLE 1 Operational Taxonomic Units (OTUs) species of samples on various Taxonomic levels.

Sample	Kingdom	Phylum	Class	Order	Family	Genus	Species
CM-1	1	6	9	11	20	27	48
CM-2	1	6	9	9	15	23	38
CM-3	1	6	8	10	19	26	43
CM-4	1	5	7	8	16	21	34
CM-5	1	6	10	11	21	29	53
LM-1	1	7	9	10	16	21	26
LM-2	1	6	10	11	19	27	42
LM-3	1	6	9	10	17	23	29
LM-4	1	5	7	8	14	20	25
LM-5	1	5	8	9	18	25	40
HM-1	1	6	10	11	20	27	43
HM-2	1	6	10	11	20	26	43
HM-3	1	6	9	10	18	24	34
HM-4	1	5	7	8	15	19	32
HM-5	1	5	8	9	17	23	37
CF-1	1	4	6	7	15	19	27
CF-2	1	6	8	9	18	24	38
CF-3	1	6	10	11	20	26	43
CF-4	1	5	7	8	19	24	41
CF-5	1	5	8	9	16	22	36
LF-1	1	5	9	10	18	23	38
LF-2	1	5	9	10	20	26	44
LF-3	1	5	9	10	21	29	47
LF-4	1	5	9	10	21	25	40
LF-5	1	5	7	8	15	18	28
HF-1	1	6	10	12	23	31	52
HF-2	1	6	10	11	21	25	42
HF-3	1	5	7	8	16	20	33
HF-4	1	5	8	9	17	23	35
HF-5	1	6	8	9	13	16	18

Table represents taxonomic diversity across control and experimental groups (CM, Control Male; LM, low-dose Male; HM, high-dose Male; CF, Control Female; LF, low-dose Female; HF, high-dose Female) followed by replicate numbers (1–5). Numbers indicate distinct taxonomic units detected at each classification levels from Kingdom to species.

Species level analysis

The heatmap in Figure 6 represents the average relative abundance of bacterial species in the fecal microbiome for each treatment group (control, low-dose, and high-dose; n = 5 per sex) in male and female rats (Supplementary Figure S1 displays the relative abundance for all animals in each group).

The control groups exhibit a distinct abundance profile compared to the treatment groups. The high-dose treated groups show an increase in the abundance of certain bacterial species, while the low-dose groups demonstrate an intermediate profile. These findings suggest that the treatment has a dose-dependent effect on the fecal microflora composition, with

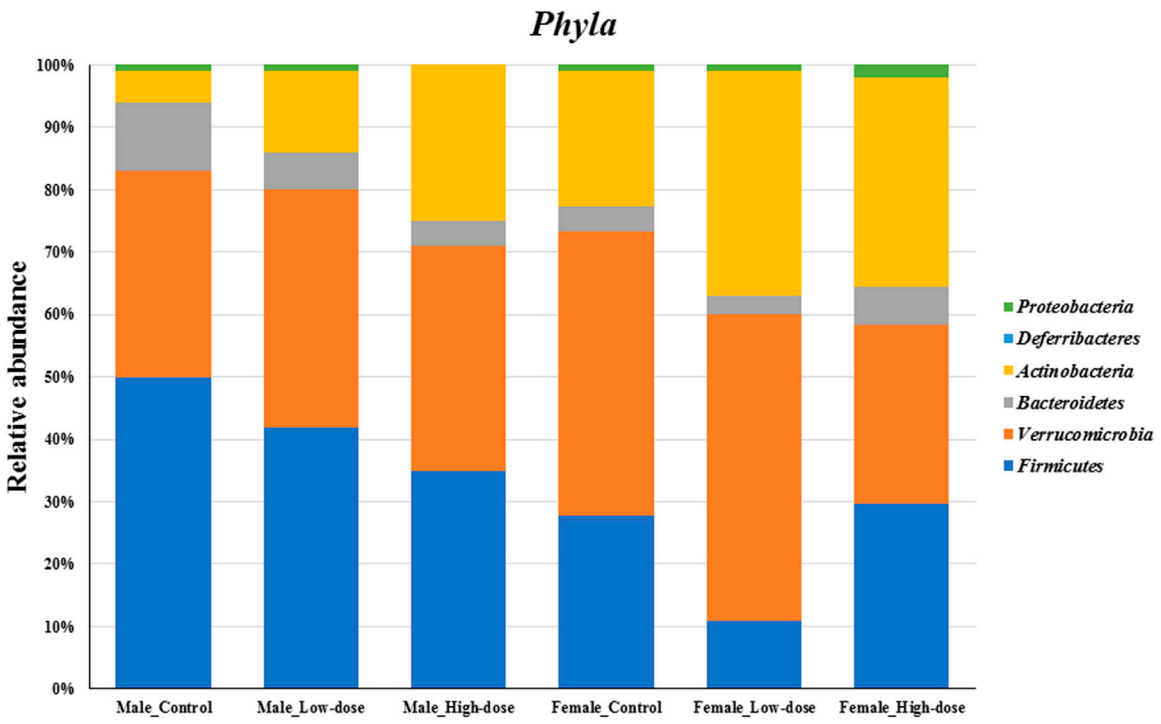


FIGURE 4
Changes in the fecal microbial composition at the *Phyla* level (top six phylum) across different treatment groups (control, low-dose, and high-dose) in both male and female rats (n = 5 in each group).

higher doses leading to more changes in the abundance of specific bacterial species.

The observed changes in microbial composition at the phylum, genus, and species levels are interconnected and reflect the taxonomic relationships among the affected bacteria. The decrease in *Firmicutes* at the phylum level might be primarily driven by the reduction in *Lactobacillus* species, which belong to this phylum. The increase in *Actinobacteria* can be largely attributed to the substantial increase in *Bifidobacterium pseudolongum*, a member of this phylum.

The sex-specific changes in *Akkermansia muciniphila*, the representative of the Verrucomicrobia phylum, directly contribute to the observed differences in *Verrucomicrobia* abundance between males and females. The increase in *Proteobacteria* in high-dose females can be linked to the slight increases in genera such as *Escherichia* and *Parasutterella*, which belong to this phylum.

The clustering patterns observed in the heatmap analysis further highlight the relationships among the affected species and their contribution to the overall changes in microbial composition. The co-clustering of various *Lactobacillus species* in males and their collective decrease with TC-ART underscore their shared response to the intervention. Similarly, the separate clustering of the control group in females emphasizes the impact of the drug on the female gut microbiome.

In conclusion, the perinatal exposure to TC-ART was associated with alterations in the gut microbial composition at multiple taxonomic levels, with sex-specific differences in adult rats. The changes observed at the phylum level are driven by the differential responses of specific genera and species, highlighting the intricate relationships within the gut microbiome.

Beta diversity analysis

Males

The PCoA plot for males (Figure 7A) reveals distinct clustering patterns related to the treatment groups. The control group (CM_1 to CM_5) forms two subclusters along Axis 1, indicating some within-group variation but overall separation from the treatment groups. This suggests that the untreated male samples have a distinct microbial community structure compared to those that received the treatments.

The low-dose group (LM_6 to LM_10) forms a relatively compact cluster, although it shows some variation along Axis 2. This cluster partially overlaps with both the control and high-dose groups, suggesting that the low-dose treatment induces a shift in the microbial community structure that is intermediate between the control and high-dose groups.

The high-dose group (HM_11 to HM_15) exhibits a separation from the control group, with samples spreading along both Axis 1 and Axis 2. This indicates that the high-

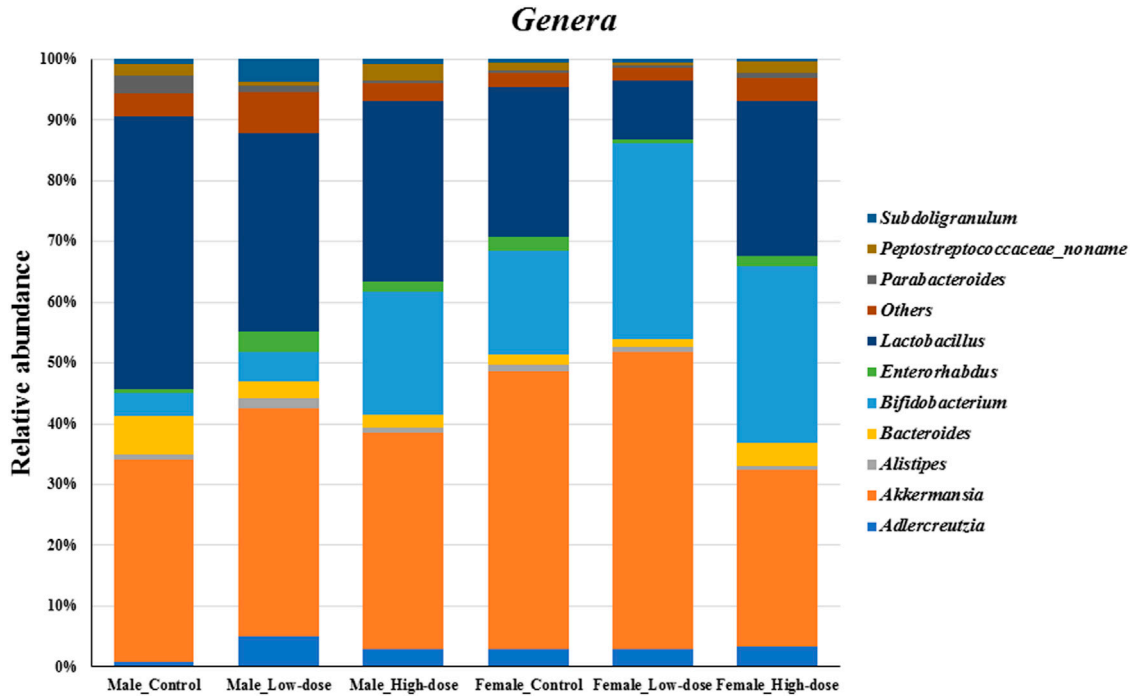


FIGURE 5
Changes in the fecal microbial composition at the *genera* level (top ten genus plus others) across different treatment groups (control, low-dose, and high-dose) in both male and female rats (n = 5 in each group).

dose treatment induces a shift in the microbial community structure compared to the untreated samples. However, the spread of the samples also suggests that there is considerable individual variation in the response to the high-dose treatment.

While there is a separation between the control group and the high-dose group, suggesting treatment-related changes, there is also some overlap, particularly between the low-dose group and the other groups. This overlap suggests that the treatment effect may not be as distinct for all individuals, and there could be other factors contributing to the variation within groups.

These results demonstrate that the treatments have an impact on the beta diversity of the male microbiome, with clustering patterns associated with each treatment group. The observed change across the treatment groups provides evidence for a dose-dependent response in the microbial community structure of males.

Females

The PCoA plot for females (Figure 7B) reveals a high degree of overlap among the control (CF_16 to CF_20), low-dose (LF_21 to LF_25), and high-dose (HF_26 to HF_30) groups. This overlap suggests that the treatments did not induce distinct shifts in the microbial community structure of females.

The control group samples tend to cluster towards the left side of the plot along Axis 1, but there is no clear separation between the control and treatment groups. This indicates that the untreated female samples do not have a markedly distinct microbial community structure compared to those that received the treatments.

The low-dose and high-dose groups are largely intermingled, with samples scattered throughout the plot. This lack of separation between the treatment groups suggests that increasing the treatment dose did not result in a consistent, dose-dependent shift in the microbiome composition of females.

The overall lack of clustering based on treatment groups in females contrasts with the patterns observed in males. While male samples showed distinct clustering and a gradient of change across treatment groups, female samples exhibit a high degree of overlap and no clear treatment-related patterns.

These results indicate that the treatments did not have a significant impact on the beta diversity of the female microbiome, as evidenced by the lack of distinct clustering patterns associated with the treatment groups. The overlap among the control and treatment groups suggests that factors other than the treatment itself may be driving the variation in the microbial community structure of females.

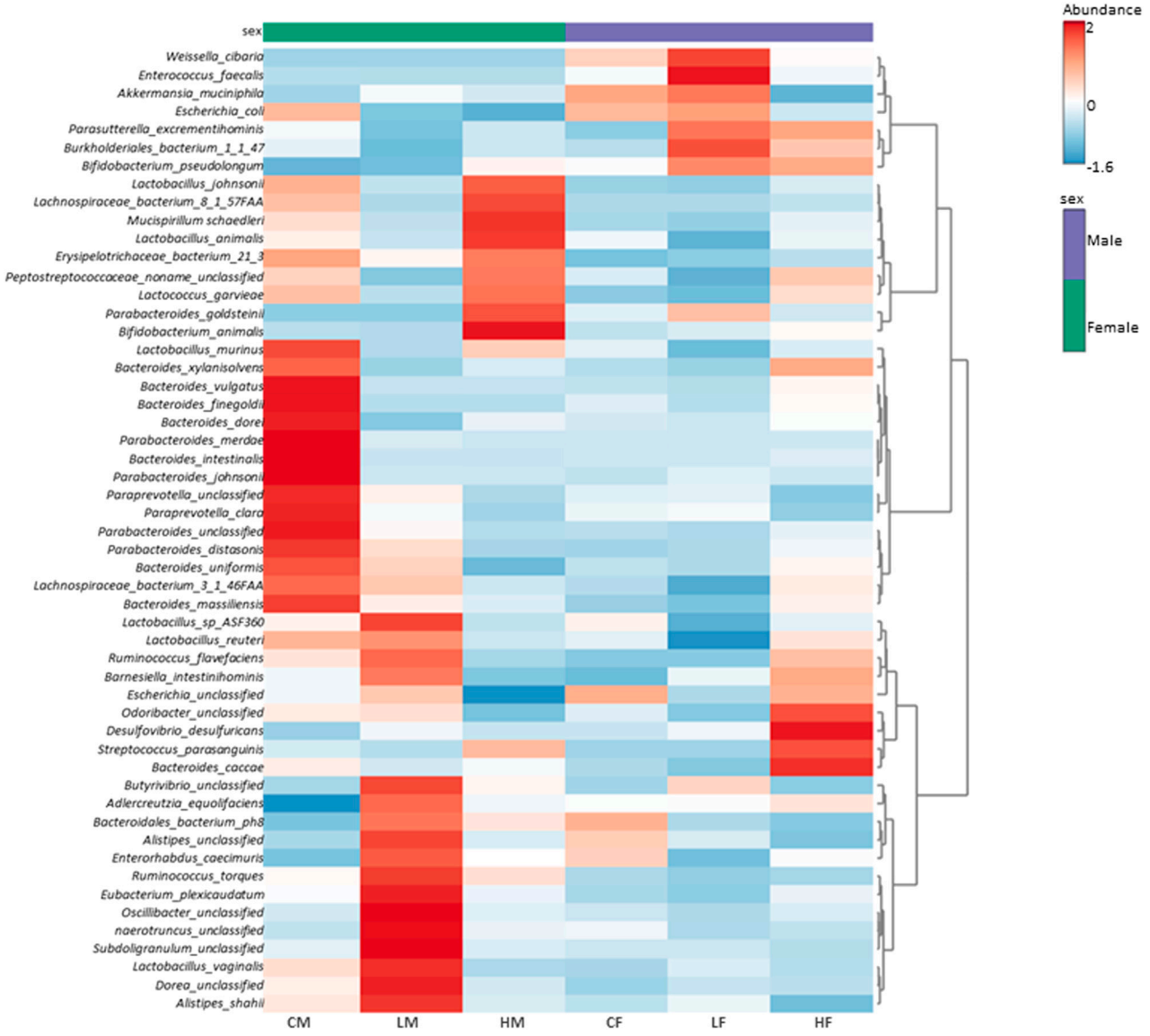


FIGURE 6
Heatmap illustrating the average (n = 5) relative abundance of bacterial species in the fecal microbiome for each treatment group (control, low-dose, and high-dose) in male and female rats. The color gradient from light to dark signifies low to high relative abundance. Vertical clustering represents the similarity in the abundance of different species among the treatment groups, with shorter branch lengths indicating greater similarity. Horizontal clustering shows the similarity of species abundance between treatment groups, with shorter branch lengths suggesting higher similarity between groups.

Effect of perinatal exposure to HIV TC-ART on SCFA levels in rat offspring

One prediction that the changes in microbiome abundance and composition as described above would be that their produced metabolites might also changed. Given that the metabolites derived from the microbiome can affect different biological processes, we then evaluated whether perinatal exposure to TC-ART would also impact the levels of SCFA produced by the bacteria. To this end, we compared the concentrations of various SCFAs between control, low-dose, and high-dose groups in both males and females. Figure 8 represents the average concentrations and standard deviations of each SCFA for the different treatment groups (control, low-dose, and high-dose) for male (A) and female (B) rats measured in fecal samples.

Data presented in Figure 8 showed no statistically significant differences between the groups. To better understand whether within the same sex the treatment had an effect, we next

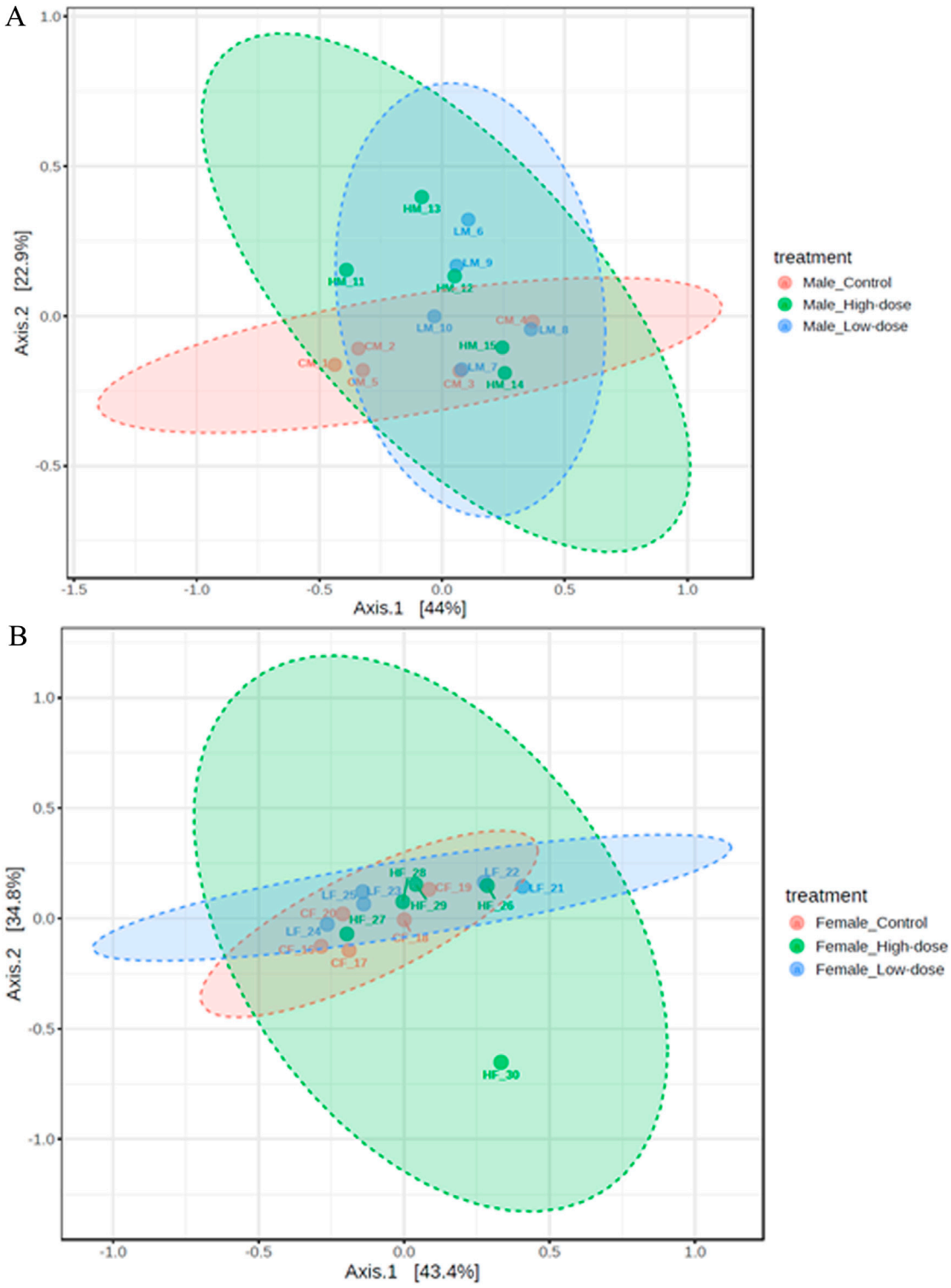


FIGURE 7
Principal Coordinate Analysis (PCoA) plots showing beta diversity in males **(A)** and females **(B)** rats across control (red points), low-dose (blue points), and high-dose groups (green points).

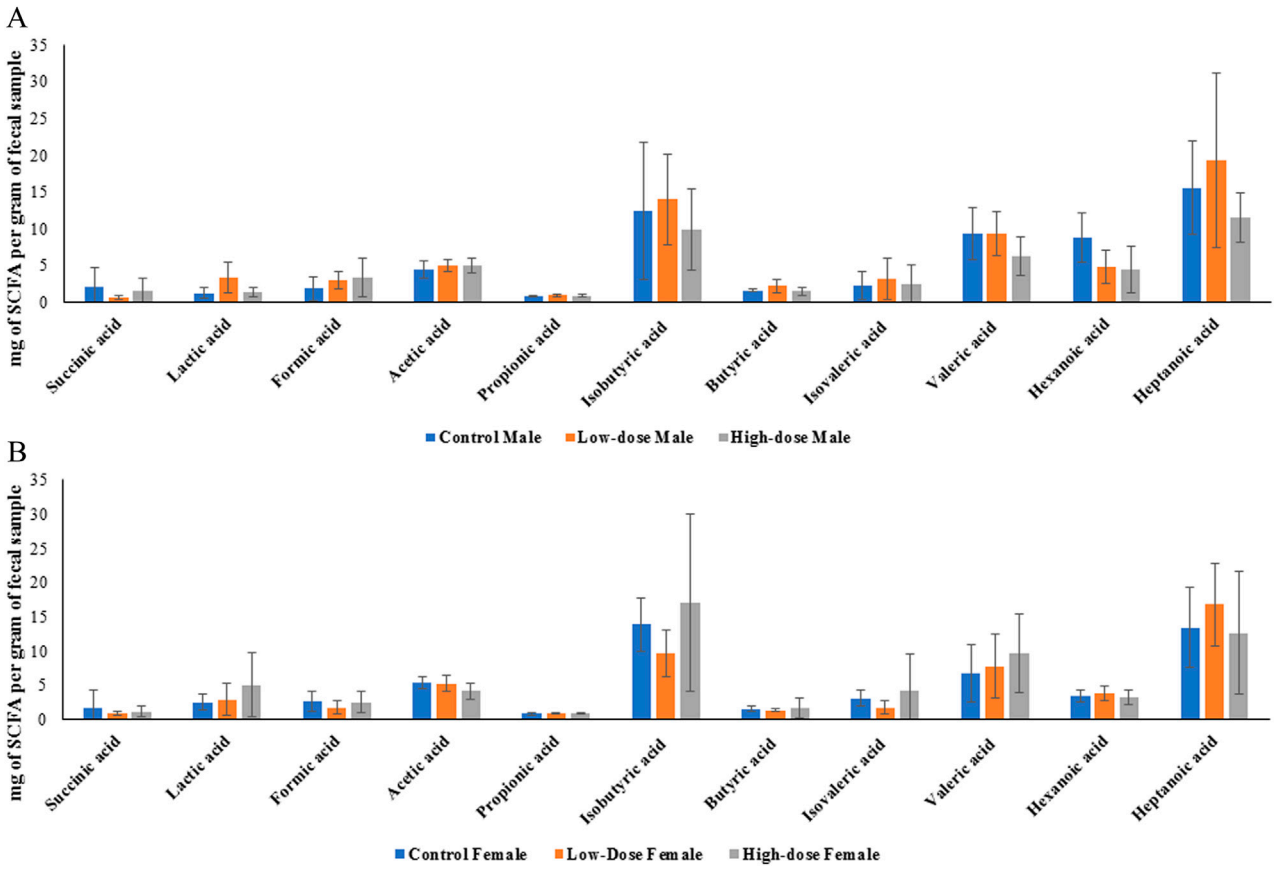


FIGURE 8
Bar graphs showing concentration of individual SCFA in control, low-dose, and high-dose groups for male **(A)** and female **(B)** rats. Each bar represents the mean ± SD of five observations (n = 5).

evaluated concentrations of each SCFA across different doses within the same sex. While we observed different trends in males and females, none of the data displayed statistically significant differences. Nevertheless, it is interesting that in both sexes the treatments tended to alter levels of lactic acid while the trends of other fatty acids were different in males and females (Figure 8).

Sex-dependent differences in SCFA concentrations

Among all the SCFAs analyzed, only hexanoic acid showed a statistically significant difference between sexes, with higher levels observed in control males compared to control females ($p < 0.05$). However, other SCFAs showed varying patterns between male and female rats across different treatment groups (represented by the considerable overlap of error bars in Figure 8).

Given the intricate relationship between gut microbiota, SCFAs, and mucosal immunity, we also examined Immunoglobulin A (IgA) levels in offspring. SCFAs are known to promote intestinal IgA responses, and investigating both parameters provide a more comprehensive view of how gestational ART exposure might influence the developing gut ecosystem.

Effect of perinatal exposure to HIV TC-ART on IgA levels in rat offspring

We quantified both secretory (fecal unbound) and bacterial-bound IgA and compared concentrations between control and treated groups in both males and females. The levels of unbound and bacteria-bound IgA were comparable between treated and control offspring, with no significant differences observed across the treatment groups (Figure 9). Similarly, no significant sex differences were observed between the levels of unbound or bacterial-bound IgA detected in the feces of adult offspring due to the perinatal exposure to HIV TC-ART.

Discussion

It is estimated that 39 million people are infected with HIV and over the past few years [49], life expectancy of individuals

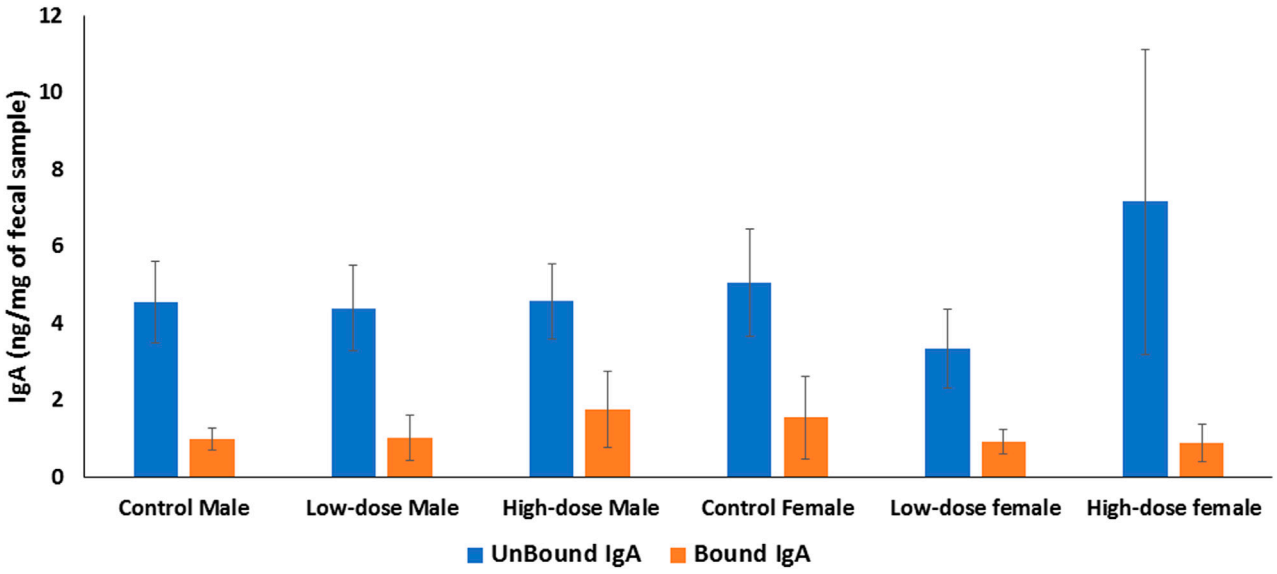


FIGURE 9
Bar graph showing quantification of IgA levels in the offspring feces. The level of IgA in rat feces is expressed as ng per mg of feces. IgA unbound (blue bars); bacteria-bound IgA (orange bars). Error bars represent standard deviation values (n = 5).

living with HIV has improved significantly with the widespread usage of ART [50]. However, as the population of ART-treated individuals continues to grow, there is a pressing need to understand the long-term effects of ART exposure, particularly during important developmental stages such as during gestation [51–53], in the absence of the HIV.

Given the critical role of the gut microbiome in shaping immune function, metabolism, and neurodevelopment, investigation of the potential impact of ART on the gut microbiome and its consequences for the health of the offspring is warranted. The gut microbiome's influence extends beyond the intestinal environment, playing roles in gut-brain communication, liver function, and cardiovascular health through complex bidirectional interactions known as the gut-brain axis, gut-liver axis, and gut-heart axis, respectively. In this study, we investigated the effects of perinatal ART exposure on the composition of the gut microbiome and their metabolites (SCFA's) in aged rat offspring.

Our findings suggest that perinatal exposure to ART is associated with alterations in the gut microbiome composition in aged rat offspring at multiple taxonomic levels, with notable sex-specific differences. However, despite these changes in the gut microbiome composition, we did not observe statistically significant differences in SCFA levels across treatment groups or between sexes, which may be due to high individual variability.

Interestingly, our study did not find significant differences in the alpha diversity indices (Chao1 or Shannon) across treatment groups in either male or female rats. This suggests that the developmental ART exposure has not significantly affected the overall species richness or diversity of the fecal microbiome in the aged offspring at 12 months. The lack of significant differences in our study may be attributed to several factors, such as the subtle effects of gestational ART exposure on the fecal microbiome of the offspring, the long-term nature of the study allowing for microbial community recovery, and the high individual variability masking potential treatment effects [54]. However, it is important to consider that alpha diversity measures provide a broad overview of the microbial community structure and may not capture subtle changes in specific bacterial taxa [55, 56].

To gain a deeper understanding of the effects of developmental ART exposure on the fecal microbiome, we performed an integrated analysis of taxonomic levels, examining the changes at the phylum, genus, and species levels and their interconnectedness. Our results revealed dose-dependent and sex-specific alterations in the relative abundances of various bacterial taxa.

At the phylum level, we observed distinct differences between control and TC-ART treated groups in both males and females. In males, *Firmicutes* and *Bacteroidetes* exhibited a dose-dependent decrease, while *Actinobacteria* showed a dose-dependent increase. In females, the response was more complex, with *Firmicutes* decreasing in the low-dose group but increasing in the high-dose group, and *Actinobacteria* showing a dose-dependent increase. These findings suggest that TC-ART treatment may modulate the fecal microbiome composition in a dose- and sex-specific manner. Similar effects of antibiotics on the gut microbiome have been reported in previous studies [57–61].

The reduction in *Lactobacillus* species may increase offspring susceptibility to gastrointestinal disturbances and infections [62]. Conversely, the increase in *Bifidobacterium* species could offer some protective effects, given their association with improved immune function and metabolic health [63]. These alterations share similarities with findings from studies on early-life antibiotic exposure [64], suggesting that various early-life factors can induce long-lasting changes in gut microbiota composition. The sex-specific differences observed echo earlier findings [65] on sex-specific microbial patterns. This highlights the complex interplay between early-life exposures, sex hormones, and gut microbiome development. The species-level analysis, including the heatmap and clustering patterns, further confirmed the dose-dependent and sex-specific effects of TC-ART treatment on the gut microbiome composition. Specific bacterial species showed dose-dependent increases or decreases, while others exhibited sex-specific patterns of change. These species-level alterations drove the changes observed at the genus and phylum levels, highlighting the interconnectedness of taxonomic levels in the microbiome. Sex hormones such as estrogen and testosterone are known to influence gut microbiota composition and immune responses, potentially leading to distinct microbial community structures between males and females [66].

The observed changes in the gut microbiome composition may have important implications for extraintestinal organ functions, as discussed earlier regarding the gut microbiome's role in immune function, metabolism, and gut-extraintestinal organ axes. *Lactobacillus* and *Bifidobacterium* species, which were affected by TC-ART treatment, are known for their probiotic properties and have been associated with various benefits, such as improved immune function, reduced inflammation, and protection against pathogens [67–70]. Conversely, a decrease in these beneficial bacteria has been linked to an increased risk of metabolic disorders, inflammatory bowel disease, and infections [71–73].

Moreover, the sex-specific alterations in key genera, such as *Akkermansia*, may have differential effects on health outcomes. *Akkermansia muciniphila*, which was more abundant in females, has been inversely associated with obesity, diabetes, and inflammation [74, 75]. The higher prevalence of this species in females may confer some protection against metabolic disorders, while its reduction in males may increase their susceptibility to these conditions [65, 76–78].

The dose-dependent effects of TC-ART treatment on specific bacterial species also warrant attention. For instance, the increase in *B. pseudolongum* in a dose-dependent manner may have positive implications for gastrointestinal tract, as this species has been shown to exert anti-inflammatory effects and improve gut barrier function [79–81]. However, the decrease in *Lactobacillus* species with increasing doses of TC-ART treatment may compromise the beneficial effects of these bacteria on the host.

The beta diversity analysis revealed sex-specific responses to developmental TC-ART exposure in the gut microbiome composition of rat offspring. Males exhibited distinct clustering patterns associated with treatment groups, indicating that gestational TC-ART exposure alters the gut microbiome composition in male offspring. The observed sex-specific alterations underscore the complex interplay between host factors, such as sex hormones, and the gut microbiome in response to early-life exposures [82]. In contrast to the male microbiome, the PCoA plot for females revealed a high degree of overlap among the control, low-dose, and high-dose groups, indicating that gestational ART exposure did not significantly alter the overall microbial community structure in female offspring. Colonization of bacterial community during early development may play a more dominant role in shaping the microbial community structure [65, 83, 84].

Several factors may contribute to the observed sex differences in response to gestational ART exposure, including hormonal influences [85, 86], gender-specific immune responses, genetic and epigenetic variations [87, 88]. In addition to the gut microbiome composition, we further investigated the impact of developmental ART exposure on the concentrations of SCFAs in the offspring. SCFAs are important microbial metabolites that play a role in maintaining gut homeostasis, regulating immune function, and influencing metabolic processes [89–91]. However, our analysis of SCFA and IgA concentrations revealed high variability within treatment groups, complicating the interpretation of the results. The high variability observed could be attributed to individual differences in gut microbiome composition and the complex nature of short-chain fatty acid production and metabolism. In addition, the ongoing studies using metatranscriptomics analysis would help to understand if the observed microbiome changes result in different metabolic and immune responses between males and females. Moreover, integration of multi omics approaches such as metatranscriptomics, and metabolomics, would be valuable to better understand the impact of developmental exposure of TC-ART on the gut microbiome function and metabolic outputs.

Conclusion

Our study showed that gestational and lactational exposure to TC-ART was associated with alterations in the fecal microbiome composition of aged rat offspring, with notable sex-specific differences. These changes were observed at various taxonomic levels and were characterized by dose-dependent and sex-specific patterns. Despite these changes in the fecal microbiome, we did not observe significant differences in IgA levels and SCFA concentrations across treatment groups or sexes. However, given the complex interplay between the

microbiome and host factors, other functional consequences may exist that were not captured by these specific markers. Further investigations, including meta transcriptomics, could help to determine the impact of TC-ART on the gut microbiome and metabolic function in both males and females to account for sex differences.

Author contributions

CM: Formal analysis, Investigation, Methodology, Visualization, Writing – original draft; YY: Analysis, Investigation, Methodology, Writing – review and editing; KG: Conceptualization, Methodology, Investigation, Visualization, Writing – review and editing; KK: Methodology; VS: Conceptualization, Project administration, Writing – review and editing; HC: Methodology, Writing – review and editing; VS: Conceptualization, Project administration, Writing – review and editing; JS: Methodology, Project administration, Writing – review and editing; SK: Conceptualization, Methodology, Investigation, Formal analysis, Supervision, Project administration, Writing – review and editing. All authors contributed to the article and approved the submitted version.

Author disclaimer

The findings and conclusions presented in this manuscript are those of the authors and do not necessarily represent the views of the U.S. Food and Drug Administration or the National Institutes of Health. Any mention of a commercial product is for clarification only, it is not an endorsement for the use of it.

Data availability

The datasets presented in this article will be available as per the guidelines of the U.S. Food and Drug Administration data sharing policy. Requests to access the datasets should be directed to Sangeeta.khare@fda.hhs.gov.

Ethics statement

The animal study was approved by Institutional Animal Care and Use Committee of AmplyfyBio. The study was conducted in accordance with the local legislation and institutional requirements.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This study was supported by the NIEHS under an Interagency Agreement between FDA and NIEHS (FDA IAG#224-17-0502 and NIH IAG#AES17011001-1-0-5). It was also supported by HHSN273201400015C. VS, JS, and HC were supported by the Intramural Research Program of the National Institutes of Health ZIAES103369-03. CM, YY, and KK were supported by an appointment to the Postgraduate Research Program at the NCTR administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement with the U.S. Department of Energy and the U.S. Food and Drug Administration.

Acknowledgments

The authors would like to thank Dr. Radwa Hanafy (NCTR), Dr. Kanungo Jyotshnabala (NCTR), and Dr. Kelly Shipkowski (National Institutes of Health) for reviewing the manuscript and providing valuable comments and suggestions.

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Generative AI statement

The authors declare that no Generative AI was used in the creation of this manuscript.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.ebm-journal.org/articles/10.3389/ebm.2025.10468/full#supplementary-material

References

1. Jandhyala SM, Talukdar R, Subramanyam C, Vuyyuru H, Sasikala M, Nageshwar Reddy D. Role of the normal gut microbiota. *World J Gastroenterol* (2015) **21**(29):8787–803. doi:10.3748/wjg.v21.i29.8787

2. Thursby E, Juge N. Introduction to the human gut microbiota. *Biochem J* (2017) **474**(11):1823–36. doi:10.1042/bcj20160510

3. Hasan N, Yang H. Factors affecting the composition of the gut microbiota, and its modulation. *PeerJ* (2019) **7**:e7502. doi:10.7717/peerj.7502

4. DeGruttola AK, Low D, Mizoguchi A, Mizoguchi E. Current understanding of dysbiosis in disease in human and animal models. *Inflamm Bowel Dis* (2016) **22**(5): 1137–50. doi:10.1097/mib.0000000000000750

5. Ferretti P, Pasolli E, Tett A, Asnicar F, Gorfer V, Fedi S, et al. Mother-to-Infant microbial transmission from different body sites shapes the developing infant gut microbiome. *Cell Host and Microbe* (2018) **24**(1):133–45.e5. doi:10.1016/j.chom.2018.06.005

6. Mitchell CM, Mazzoni C, Hogstrom L, Bryant A, Bergerat A, Cher A, et al. Delivery mode affects stability of early infant gut microbiota. *Cell Rep Med* (2020) **1**(9):100156. doi:10.1016/j.xcrm.2020.100156

7. Rutayisire E, Huang K, Liu Y, Tao F. The mode of delivery affects the diversity and colonization pattern of the gut microbiota during the first year of infants' life: a systematic review. *BMC Gastroenterol* (2016) **16**(1):86. doi:10.1186/s12876-016-0498-0

8. Ma J, Li Z, Zhang W, Zhang C, Zhang Y, Mei H, et al. Comparison of gut microbiota in exclusively breast-fed and formula-fed babies: a study of 91 term infants. *Scientific Rep* (2020) **10**(1):15792. doi:10.1038/s41598-020-72635-x

9. Catassi G, Aloi M, Giorgio V, Gasbarrini A, Cammarota G, Ianiro G. The role of diet and nutritional interventions for the infant gut microbiome. *Nutrients* (2024) **16**(3):400. doi:10.3390/nu16030400

10. Clarke G, Stilling RM, Kennedy PJ, Stanton C, Cryan JF, Dinan TG. Minireview: gut microbiota: the neglected endocrine organ. *Mol Endocrinol* (2014) **28**(8):1221–38. doi:10.1210/me.2014-1108

11. Sherman MP, Zaghouani H, Niklas V. Gut microbiota, the immune system, and diet influence the neonatal gut-brain axis. *Pediatr Res* (2015) **77**(1-2):127–35. doi:10.1038/pr.2014.161

12. Nash MJ, Frank DN, Friedman JE. Early microbes modify immune system development and metabolic homeostasis-the "restaurant" hypothesis revisited. *Front Endocrinol (Lausanne)* (2017) **8**:349. doi:10.3389/fendo.2017.00349

13. Laforest-Lapointe I, Arrieta MC. Patterns of early-life gut microbial colonization during human immune development: an ecological perspective. *Front Immunol* (2017) **8**:788. doi:10.3389/fimmu.2017.00788

14. Patel PH, Zulfiqar H. Reverse transcriptase inhibitors in: *StatPearls. Treasure island (FL)*. StatPearls Publishing. (2023).

15. Pau AK, George JM. Antiretroviral therapy: current drugs. *Infect Dis Clin North America* (2014) **28**(3):371–402. doi:10.1016/j.idc.2014.06.001

16. Braitstein P, Brinkhof MWG, Dabis F, Schechter M, Boulle A, Miotti P, et al. Mortality of HIV-1-infected patients in the first year of antiretroviral therapy: comparison between low-income and high-income countries. *Lancet* (2006) **367**(9513):817–24. doi:10.1016/S0140-6736(06)68337-2

17. Shetty AK, Maldonado Y. Antiretroviral drugs to prevent mother-to-child transmission of HIV during breastfeeding. *Curr HIV Res* (2013) **11**(2):102–25. doi:10.2174/1570162x11311020004

18. Kesho Bora Study Group, de Vincenzi I. Triple antiretroviral compared with zidovudine and single-dose nevirapine prophylaxis during pregnancy and breastfeeding for prevention of mother-to-child transmission of HIV-1 (Kesho Bora study): a randomised controlled trial. *Lancet Infect Dis* (2011) **11**(3):171–80. doi:10.1016/S1473-3099(10)70288-7

19. Chagomerana MB, Miller WC, Tang JH, Hoffman IF, Mthiko BC, Phulusa J, et al. Optimizing prevention of HIV mother to child transmission: duration of antiretroviral therapy and viral suppression at delivery among pregnant Malawian women. *PLoS One* (2018) **13**(4): e0195033. doi:10.1371/journal.pone.0195033

20. Gupta A, Verma A, Kashyap M, Gautam P. ART in prevention of mother-to-child transmission of HIV. *The J Obstet Gynecol India* (2020) **70**(1):18–22. doi:10.1007/s13224-019-01263-x

21. clinicalinfo.hiv.gov. *Guidelines for the Use of Antiretroviral Agents in Adults and Adolescents With HIV*. (2023). Available online at: https://clinicalinfo.hiv.gov/sites/default/files/guidelines/documents/adult-adolescent-arv/guidelines-adult-adolescent-arv.pdf. (Accessed May 5, 2025).

22. Imahashi M, Ode H, Kobayashi A, Nemoto M, Matsuda M, Hashiba C, et al. Impact of long-term antiretroviral therapy on gut and oral microbiotas in HIV-1-infected patients. *Sci Rep* (2021) **11**(1):960. doi:10.1038/s41598-020-80247-8

23. Williams B, Landay A, Presti RM. Microbiome alterations in HIV infection a review. *Cell Microbiol* (2016) **18**(5):645–51. doi:10.1111/cmi.12588

24. Bender JM, Li F, Martelly S, Byrt E, Rouzier V, Leo M, et al. Maternal HIV infection influences the microbiome of HIV-uninfected infants. *Sci Transl Med* (2016) **8**(349):349ra100. doi:10.1126/scitranslmed.aaf5103

25. Zilberman-Schapira G, Zmora N, Itav S, Bashiardes S, Elinav H, Elinav E. The gut microbiome in human immunodeficiency virus infection. *BMC Med* (2016) **14**(1):83. doi:10.1186/s12916-016-0625-3

26. Li S, Armstrong A, Neff C, Shaffer M, Lozupone C, Palmer B. Complexities of gut microbiome dysbiosis in the context of HIV infection and antiretroviral therapy. *Clin Pharmacol and Ther* (2016) **99**(6):600–11. doi:10.1002/cpt.363

27. Portincasa P, Bonfrate L, Vacca M, De Angelis M, Farella I, Lanza E, et al. Gut microbiota and short chain fatty acids: implications in glucose homeostasis. *Int J Mol Sci* (2022) **23**(3):1105. doi:10.3390/ijms23031105

28. Liu P, Wang Y, Yang G, Zhang Q, Meng L, Xin Y, et al. The role of short-chain fatty acids in intestinal barrier function, inflammation, oxidative stress, and colonic carcinogenesis. *Pharmacol Res* (2021) **165**:105420. doi:10.1016/j.phrs.2021.105420

29. Corrêa-Oliveira R, Fachi JL, Vieira A, Sato FT, Vinolo MAR. Regulation of immune cell function by short-chain fatty acids. *Clin and Translational Immunol* (2016) **5**(4):e73. doi:10.1038/cti.2016.17

30. Ghosh S, Whitley CS, Haribabu B, Jala VR. Regulation of intestinal barrier function by microbial metabolites. *Cell Mol Gastroenterol Hepatol* (2021) **11**(5): 1463–82. doi:10.1016/j.jcmgh.2021.02.007

31. Scheithauer TPM, Rampanelli E, Nieuwdorp M, Vallance BA, Verchere CB, van Raalte DH, et al. Gut microbiota as a trigger for metabolic inflammation in obesity and type 2 diabetes. *Front Immunol* (2020) **11**:571731. doi:10.3389/fimmu.2020.571731

32. Moreno-Indias I, Cardona F, Tinahones FJ, Queipo-Ortuño MI. Impact of the gut microbiota on the development of obesity and type 2 diabetes mellitus. *Front Microbiol* (2014) **5**:190. doi:10.3389/fmicb.2014.00190

33. Hartstra AV, Bouter KE, Bäckhed F, Nieuwdorp M. Insights into the role of the microbiome in obesity and type 2 diabetes. *Diabetes Care* (2014) **38**(1):159–65. doi:10.2337/dc14-0769

34. Yukino-Iwashita M, Nagatomo Y, Kawai A, Taruoka A, Yumita Y, Kagami K, et al. Short-chain fatty acids in gut–heart Axis: their role in the pathology of heart failure. *J Personalized Med* (2022) **12**(11):1805. doi:10.3390/jpm12111805

35. Modrego J, Ortega-Hernández A, Goirigolzarri J, Restrepo-Córdoba MA, Bäuerl C, Cortés-Macías E, et al. Gut microbiota and derived short-chain fatty acids are linked to evolution of heart failure patients. *Int J Mol Sci* (2023) **24**(18):13892. doi:10.3390/ijms241813892

36. Hu T, Wu Q, Yao Q, Jiang K, Yu J, Tang Q. Short-chain fatty acid metabolism and multiple effects on cardiovascular diseases. *Ageing Res Rev* (2022) **81**:101706. doi:10.1016/j.arr.2022.101706

37. Tang Q, Jin G, Wang G, Liu T, Liu X, Wang B, et al. Current sampling methods for gut microbiota: a call for more precise devices. *Front Cell Infect Microbiol* (2020) **10**:151. doi:10.3389/fcimb.2020.00151

38. Murali A, Giri V, Cameron HJ, Behr C, Sperber S, Kamp H, et al. Elucidating the relations between gut bacterial composition and the plasma and fecal metabolomes of antibiotic treated wistar rats. *Microbiol Res* (2021) **12**(1): 82–122. doi:10.3390/microbiolres12010008

39. Wensel CR, Pluznick JL, Salzberg SL, Sears CL. Next-generation sequencing: insights to advance clinical investigations of the microbiome. *The J Clin Invest* (2022) **132**(7):e154944. doi:10.1172/jci154944

40. Usyk M, Peters BA, Karthikeyan S, McDonald D, Sollecito CC, Vazquez-Baeza Y, et al. Comprehensive evaluation of shotgun metagenomics, amplicon sequencing, and harmonization of these platforms for epidemiological studies. *Cell Rep Methods* (2023) **3**(1):100391. doi:10.1016/j.crmeth.2022.100391

41. Chetty A, Blekhman R. Multi-omic approaches for host-microbiome data integration. *Gut Microbes* (2024) **16**(1):2297860. doi:10.1080/19490976.2023.2297860

42. Qiao C, He M, Wang S, Jiang X, Wang F, Li X, et al. Multi-omics analysis reveals substantial linkages between the oral-gut microbiomes and inflamm-aging molecules in elderly pigs. *Front Microbiol* (2023) **14**:1250891. doi:10.3389/fmicb.2023.1250891

43. Kulkarni V, Ruprecht RM. Mucosal IgA responses: damaged in established HIV infection—yet, effective weapon against HIV transmission. *Front Immunol* (2017) **8**:1581. doi:10.3389/fimmu.2017.01581

44. Klein MB, Young J, Ortiz-Paredes D, Wang S, Walmsley S, Wong A, et al. Virological outcomes after switching to abacavir/lamivudine/dolutegravir combined with adherence support in people living with HIV with poor adherence: a phase IV, multicentre randomized prospective open label study (TriiADD-CTN 286). *Patient Preference and Adherence* (2022) **16**:3267–81. doi:10.2147/ppa.s379065

45. (MFMER), M.F.f.M.E.a.R. Abacavir, dolutegravir, and lamivudine (oral route) (2025). Available online at: https://www.mayoclinic.org/drugs-supplements/abacavir-dolutegravir-and-lamivudine-oral-route/proper-use/drg-20122502.

46. MicrobiomeAnalyst 2.0: comprehensive statistical, functional and integrative analysis of microbiome data. (2025). Available online at: https://www.microbiomeanalyst.ca/MicrobiomeAnalyst/ModuleView.xhtml. (Accessed May 5, 2025).

47. Lu Y, Zhou G, Ewald J, Pang Z, Shiri T, Xia J. MicrobiomeAnalyst 2.0: comprehensive statistical, functional and integrative analysis of microbiome data. *Nucleic Acids Res* (2023) **51**(W1):W310–W318. doi:10.1093/nar/gkad407 (Accessed May 5, 2025).

48. Lahiani M, Gokulan K, Sutherland V, Cunny HC, Cerniglia CE, Khare S. Early developmental exposure to triclosan impacts fecal microbial populations, IgA and functional activities of the rat microbiome. *J Xenobiotics* (2024) **14**(1):193–213. doi:10.3390/jox14010012

49. Organisation WH. Global situation and trends (2025). Available online at: https://www.who.int/data/gho/data/themes/hiv-aids#:~:text=Globally%2C%2039.0%20million%20%5B33.1%E2%80%93,considerably%20between%20countries%20and%20regions. (Accessed May 5, 2025).

50. Wandeler G, Johnson LF, Egger M. Trends in life expectancy of HIV-positive adults on antiretroviral therapy across the globe: comparisons with general population. *Curr Opin HIV AIDS* (2016) **11**(5):492–500. doi:10.1097/coh.0000000000000298

51. Dzanibe S, Jaspan HB, Zulu MZ, Kiravu A, Gray CM. Impact of maternal HIV exposure, feeding status, and microbiome on infant cellular immunity. *J Leukoc Biol* (2019) **105**(2):281–9. doi:10.1002/jlb.mr0318-120r

52. Mtintsilana A, Norris SA, Dlamini SN, Nyati LH, Aronoff DM, Koethe JR, et al. The impact of HIV and ART exposure during pregnancy on fetal growth: a prospective study in a South African cohort. *BMC Pregnancy Childbirth* (2023) **23**(1):415. doi:10.1186/s12884-023-05743-x

53. Furman D, Campisi J, Verdin E, Carrera-Bastos P, Targ S, Franceschi C, et al. Chronic inflammation in the etiology of disease across the life span. *Nat Med* (2019) **25**(12):1822–32. doi:10.1038/s41591-019-0675-0

54. Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R. Diversity, stability and resilience of the human gut microbiota. *Nature* (2012) **489**(7415):220–30. doi:10.1038/nature11550

55. Willis AD. Rarefaction, alpha diversity, and statistics. *Front Microbiol* (2019) **10**:2407. doi:10.3389/fmicb.2019.02407

56. Worsley SF, Davies CS, Mannarelli ME, Hutchings MI, Komdeur J, Burke T, et al. Gut microbiome composition, not alpha diversity, is associated with survival in a natural vertebrate population. *Anim Microbiome* (2021) **3**(1):84. doi:10.1186/s42523-021-00149-6

57. Patangia DV, Anthony Ryan C, Dempsey E, Paul Ross R, Stanton C. Impact of antibiotics on the human microbiome and consequences for host health. *Microbiologyopen* (2022) **11**(1):e1260. doi:10.1002/mbo3.1260

58. Buffie CG, Jarchum I, Equinda M, Lipuma L, Gobourne A, Viale A, et al. Profound alterations of intestinal microbiota following a single dose of clindamycin results in sustained susceptibility to Clostridium difficile-induced colitis. *Infect Immun* (2012) **80**(1):62–73. doi:10.1128/iai.05496-11

59. Lawley TD, Clare S, Walker AW, Goulding D, Stabler RA, Croucher N, et al. Antibiotic treatment of *Clostridium difficile* carrier mice triggers a supershedder state, spore-mediated transmission, and severe disease in immunocompromised hosts. *Infect Immun* (2009) **77**(9):3661–9. doi:10.1128/iai.00558-09

60. Miyoshi J, Bobe AM, Miyoshi S, Huang Y, Hubert N, Delmont TO, et al. Peripartum antibiotics promote gut dysbiosis, loss of immune tolerance, and inflammatory bowel disease in genetically prone offspring. *Cell Rep* (2017) **20**(2):491–504. doi:10.1016/j.celrep.2017.06.060

61. Stokholm J, Sevelsted A, Bønnelykke K, Bisgaard H. Maternal propensity for infections and risk of childhood asthma: a registry-based cohort study. *The Lancet Respir Med* (2014) **2**(8):631–7. doi:10.1016/s2213-2600(14)70152-3

62. Aggarwal N, Kitano S, Puah GRY, Kittelmann S, Hwang IY, Chang MW. Microbiome and human health: current understanding, engineering, and enabling technologies. *Chem Rev* (2023) **123**(1):31–72. doi:10.1021/acs.chemrev.2c00431

63. Hidalgo-Cantabrana C, Delgado S, Ruiz L, Ruas-Madiedo P, Sánchez B, Margolles A. Bifidobacteria and their health-promoting effects. *Microbiol Spectr* (2017) **5**(3). doi:10.1128/microbiolspec.bad-0010-2016

64. Pantazi AC, Balasa AL, Mihai CM, Chisnoiu T, Lupu VV, Kassim MAK, et al. Development of gut microbiota in the first 1000 Days after birth and potential interventions. *Nutrients* (2023) **15**(16):3647. doi:10.3390/nu15163647

65. Kim YS, Unno T, Kim BY, Park MS. Sex differences in gut microbiota. *World J Mens Health* (2020) **38**(1):48–60. doi:10.5534/wjmh.190009

66. d'Afflitto M, Upadhyaya A, Green A, Peiris M. Association between sex hormone levels and gut microbiota composition and diversity-A systematic review. *J Clin Gastroenterol* (2022) **56**(5):384–92. doi:10.1097/mcg.0000000000001676

67. Dempsey E, Corr SC. Lactobacillus spp. for gastrointestinal health: current and future perspectives. *Front Immunol* (2022) **13**:840245. doi:10.3389/fimmu.2022.840245

68. Gavzy SJ, Kensiski A, Lee ZL, Mongodin EF, Ma B, Bromberg JS. Bifidobacterium mechanisms of immune modulation and tolerance. *Gut Microbes* (2023) **15**(2):2291164. doi:10.1080/19490976.2023.2291164

69. Vlasova AN, Kandasamy S, Chattha KS, Rajashekara G, Saif LJ. Comparison of probiotic lactobacilli and bifidobacteria effects, immune responses and rotavirus vaccines and infection in different host species. *Vet Immunol Immunopathology* (2016) **172**:72–84. doi:10.1016/j.vetimm.2016.01.003

70. Mazziotta C, Tognon M, Martini F, Torreggiani E, Rotondo JC. Probiotics mechanism of action on immune cells and beneficial effects on human health. *Cells* (2023) **12**(1):184. doi:10.3390/cells12010184

71. Khan I, Ullah N, Zha L, Bai Y, Khan A, Zhao T, et al. Alteration of gut microbiota in inflammatory bowel disease (IBD): cause or consequence? IBD treatment targeting the gut microbiome. *Pathogens* (2019) **8**(3):126. doi:10.3390/pathogens8030126

72. Quaglio AEV, Grillo TG, Oliveira ECSD, Stasi LCD, Sassaki LY. Gut microbiota, inflammatory bowel disease and colorectal cancer. *World J Gastroenterol* (2022) **28**(30):4053–60. doi:10.3748/wjg.v28.i30.4053

73. Boulangé CL, Neves AL, Chilloux J, Nicholson JK, Dumas ME. Impact of the gut microbiota on inflammation, obesity, and metabolic disease. *Genome Med* (2016) **8**(1):42. doi:10.1186/s13073-016-0303-2

74. Schneeberger M, Everard A, Gómez-Valadés AG, Matamoros S, Ramírez S, Delzenne NM, et al. Akkermansia muciniphila inversely correlates with the onset of inflammation, altered adipose tissue metabolism and metabolic disorders during obesity in mice. *Sci Rep* (2015) **5**:16643. doi:10.1038/srep16643

75. Rodrigues VF, Elias-Oliveira J, Pereira ÍS, Pereira JA, Barbosa SC, Machado MSG, et al. Akkermansia muciniphila and gut immune system: a Good friendship that attenuates inflammatory bowel disease, obesity, and diabetes. *Front Immunol* (2022) **13**:934695. doi:10.3389/fimmu.2022.934695

76. Gao A, Su J, Liu R, Zhao S, Li W, Xu X, et al. Sexual dimorphism in glucose metabolism is shaped by androgen-driven gut microbiome. *Nat Commun* (2021) **12**(1):7080. doi:10.1038/s41467-021-27187-7

77. Shobeiri P, Kalantari A, Teixeira AL, Rezaei N. Shedding light on biological sex differences and microbiota–gut–brain axis: a comprehensive review of its roles in neuropsychiatric disorders. *Biol Sex Differences* (2022) **13**(1):12. doi:10.1186/s13293-022-00422-6

78. Zapała B, Pustelnik J, Dudek A, Milewicz T. Differences in the composition of Akkermansia species and families of christensenellaceae and ruminococcaceae bacteria in the gut microbiota of healthy polish women following a typical western diet. *Diversity* (2023) **15**(10):1103. doi:10.3390/d15101103

79. Guo W, Mao B, Cui S, Tang X, Zhang Q, Zhao J, et al. Protective effects of a novel probiotic Bifidobacterium pseudolongum on the intestinal barrier of colitis mice via modulating the ppary/STAT3 pathway and intestinal microbiota. *Foods* (2022) **11**(11):1551. doi:10.3390/foods11111551

80. Rivière A, Selak M, Lantin D, Leroy F, De Vuyst L. Bifidobacteria and butyrate-producing colon bacteria: importance and strategies for their stimulation in the human gut. *Front Microbiol* (2016) **7**:979. doi:10.3389/fmicb.2016.00979

81. Li J, Wang J, Wang M, Zheng L, Cen Q, Wang F, et al. Bifidobacterium: a probiotic for the prevention and treatment of depression. *Front Microbiol* (2023) **14**:1174800. doi:10.3389/fmicb.2023.1174800

82. Sarkar A, Yoo JY, Valeria Ozorio Dutra S, Morgan KH, Groer M. The association between early-life gut microbiota and long-term health and diseases. *J Clin Med* (2021) **10**(3):459. doi:10.3390/jcm10030459

83. Dias SP, Brouwer MC, van de Beek D. Sex and gender differences in bacterial infections. *Infect Immun* (2022) **90**(10):e0028322. doi:10.1128/iai.00283-22

84. Center CSM. Antibiotics affect male and female gut microbiomes differently (2025). Available online at: https://www.sciencedaily.com/releases/2022/07/220720150612.htm (Accessed July 20, 2022).

85. He S, Li H, Yu Z, Zhang F, Liang S, Liu H, et al. The gut microbiome and sex hormone-related diseases. *Front Microbiol* (2021) **12**:711137. doi:10.3389/fmicb.2021.711137

86. Klein SL, Flanagan KL. Sex differences in immune responses. *Nat Rev Immunol* (2016) **16**(10):626–38. doi:10.1038/nri.2016.90

87. Jašarević E, Morrison KE, Bale TL. Sex differences in the gut microbiome–brain axis across the lifespan. *Philosophical Trans R Soc B: Biol Sci* (2016) **371**(1688):20150122. doi:10.1098/rstb.2015.0122

88. Stilling RM, Dinan TG, Cryan JF. Microbial genes, brain and behaviour – epigenetic regulation of the gut–brain axis. *Genes, Brain Behav* (2014) **13**(1):69–86. doi:10.1111/gbb.12109

89. Fusco W, Lorenzo MB, Cintoni M, Porcari S, Rinninella E, Kaitsas F, et al. Short-chain fatty-acid-producing bacteria: key components of the human gut microbiota. *Nutrients* (2023) **15**(9):2211. doi:10.3390/nu15092211

90. Kim CH. Complex regulatory effects of gut microbial short-chain fatty acids on immune tolerance and autoimmunity. *Cell and Mol Immunol* (2023) **20**(4):341–50. doi:10.1038/s41423-023-00987-1

91. Liu XF, Shao J, Liao YT, Wang LN, Jia Y, Dong P, et al. Regulation of short-chain fatty acids in the immune system. *Front Immunol* (2023) **14**:1186892. doi:10.3389/fimmu.2023.1186892

EBM Experimental Biology and Medicine

EBM is the official journal of the Society for Experimental Biology and Medicine

Experimental Biology and Medicine (EBM) is a global, peer-reviewed journal dedicated to the publication of multidisciplinary and interdisciplinary research in the biomedical sciences.

Discover more of our Special Issues

See more →

Contact

development@ebm-journal.org

See more

ebm-journal.org

publishingpartnerships.frontiersin.org/our-partners