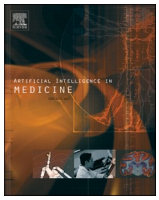
Artificial Intelligence In Medicine 153 (2024) 102887

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed



Data mining and machine learning in HIV infection risk research: An overview and recommendations

Qiwei Ge, Xinyu Lu, Run Jiang, Yuyu Zhang, Xun Zhuang *

Department of Epidemiology and Medical Statistics, School of Public Health, Nantong University, China

ARTICLE INFO

Keywords:
HIV infection risk
Data mining
Machine learning
Infection risk analysis

ABSTRACT

In the contemporary era, the applications of data mining and machine learning have permeated extensively into medical research, significantly contributing to areas such as HIV studies. By reviewing 38 articles published in the past 15 years, the study presents a roadmap based on seven different aspects, utilizing various machine learning techniques for both novice researchers and experienced researchers seeking to comprehend the current state of the art in this area. While traditional regression modeling techniques have been commonly used, researchers are increasingly adopting more advanced fully supervised machine learning and deep learning techniques, which often outperform the traditional methods in predictive performance. Additionally, the study identifies nine new open research issues and outlines possible future research plans to enhance the outcomes of HIV infection risk research. This review is expected to be an insightful guide for researchers, illuminating current practices and suggesting advancements in the field.

1. Introduction

With the advent of digitization, medical research has experienced a surge in data-driven practices, such as data mining, which extracts and identifies valuable patterns from vast data sources. Combining techniques like association correlation analysis, classification and regression, and cluster analysis, data mining allows researchers to uncover diverse patterns, with the choice of method depending on the data type and analysis goals. The impressive performance of data mining has established it as an essential tool for medical practitioners, particularly in complex tasks like mining medical data. In the past several decades, data mining has been widely used in various health management and medicinal applications, including Human Immunodeficiency Virus (HIV) research.

HIV research primarily involves biological and sociological aspects and data mining techniques can establish links between various biological and sociological attributes of individuals and their HIV infection status. It provides a wealth of classification and regression techniques. These techniques can be leveraged to build predictive systems using HIV-related data [1]. A variety of statistical and machine learning techniques have been employed for this task to identify risk prediction indicators for HIV infection. The HIV risk prediction indicators primarily comprise of infected-specific attributes, including demographic and sexual behavior features [2]. Over the past few decades, scientists and clinicians have utilized electronic health records and epidemiological investigation datasets to associate these infected-specific characteristics with HIV infection. Recent technological advancements have facilitated data-driven prediction techniques, aiding in the development of better HIV risk prediction models and addressing the significant challenge posed by high morbidity and mortality rates associated with HIV [3].

Despite the increasing number of studies exploring data mining approaches for predicting HIV infection risk, there exists no singular data mining approach that is universally applicable to all types of datasets [4]. Against the backdrop of the burgeoning trend of data mining and related methods in HIV infection risk prediction studies, this study aims to conduct a comprehensive analysis of the application of data mining and machine learning techniques in predicting HIV infection risk over the past 15 years. It provides a comprehensive description of the processes involved in risk prediction to offer a better theoretical basis for applying data-driven techniques and formulate future research agendas for researchers in the field.

This review is structured as follows: Section 2 outlines the methodology followed in conducting this survey. A detailed and critical analysis of the studies selected for the survey is presented in Section 3. Based on the survey, Section 4 offers various research directions for future researchers seeking to work in this domain. Finally, Section 5 presents the

* Corresponding author at: No 9 Seyuan Road, Nantong, Jiangsu, China.
E-mail address: xzhuang@ntu.edu.cn (X. Zhuang).

https://doi.org/10.1016/j.artmed.2024.102887
Received 22 August 2023; Received in revised form 7 March 2024; Accepted 27 April 2024
Available online 30 April 2024

concluding remarks of the study.

2. Methods

PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology is used for the present literature review. The inception of HIV infection risk model research article dates back to 2009 [5], so we chose the literatures in 2009 as the starting of our study. Various repositories, including PubMed (Medical Subject Headings), IEEE Xplore, Springer and Science Direct, were utilized to retrieve studies published within the past 15 years (2009–2023). The final search was conducted on June 11, 2023, and we discussed the changes of HIV risk prediction methods from the past to the present. Since our focus was primarily on the techniques and dynamics involved, no specific study population type was searched for the review.

Search terms related to disease (HIV OR AHI OR AIDS), task (Infect AND (Predict OR Identify), and techniques (Data mining OR Machine learning OR Classification OR Rule mining OR Sequence mining) were used on the title and abstract for the analysis from each of the databases. A total of 1427 publications were retrieved from the search analysis. After removing 661 duplicated articles and 686 not meeting the eligibility criteria after screening the titles and abstracts, 80 articles for full-text assessment. After full-text evaluation, the remaining 38 papers were finally included in the analysis.

The relevance of the articles during the abstract and full-text screening was manually assessed by two authors {GQW and LYY}, and any disagreement was resolved by discussion with a third author (ZX). The search process is depicted in Fig. 1.The eligibility criteria for the articles considered for the review were as follows:

i. Papers on statistical risk analysis for HIV infection.
ii. Papers using machine learning models for HIV infection risk prediction.
iii. Papers recommending Preventative measures based on the HIV infection risk.
iv. Papers selecting or validating features necessary for predicting HIV infection risk.
v. Original researches or articles.

Reasons for exclusion after reviewing the full-text:

i. Risk factor analysis only.
ii. Focus on other diseases of AIDS patients (e.g., cardiovascular disease).
iii. Unable to determine the specific machine learning method.
iv. Paper involved the theoretical concepts only.
v. Studies that are a review, meta-analysis, report, abstract, or poster.

Each of the 38 articles included was analyzed in detail for the overall methodology utilized in the paper. Based on the methodology, various steps (or sub-processes) were recorded including the dataset used, pre-processing steps, classification methods, platforms and software, along with the validation of techniques. Each article was manually analyzed to determine the dataset used, the type of study population discussed and studied, the various feature selection methods employed, the classification method used to predict infection, and the validation measure used in the study.

Using data mining to predict infection risk involves four primary components, including data acquisition, data preparation, model building and model evaluation. The initial step involves acquiring the appropriate dataset, which is then prepared for model building and validation. Based on these crucial steps, the process of infection risk prediction encompasses seven sub-processes (see Fig. 2). This article's literature review is designed to examine each of these sub-processes, providing an understanding and comparison of different research pieces.

3. Results

3.1. Journal publications and trends

A total of 38 selected articles were published in 20 different journals, including medical informatics journals and clinical journals. The top sources of publication included *Journal of Acquired Immune Deficiency Syndromes* (15.8 %), followed by *AIDS and Behavior* (10.5 %), and *Clinical Infectious Diseases* (10.5 %). There are also publications such as *Lancet HIV*, *BMJ Open*, *Medicine* and so on.

The distribution of 38 articles over the years is depicted in Fig. 3, which reveals that 78.9 % of the studies were published after the year 2017. This phenomenon can be attributed to the fact that machine learning techniques were not widely popular in the early years of the

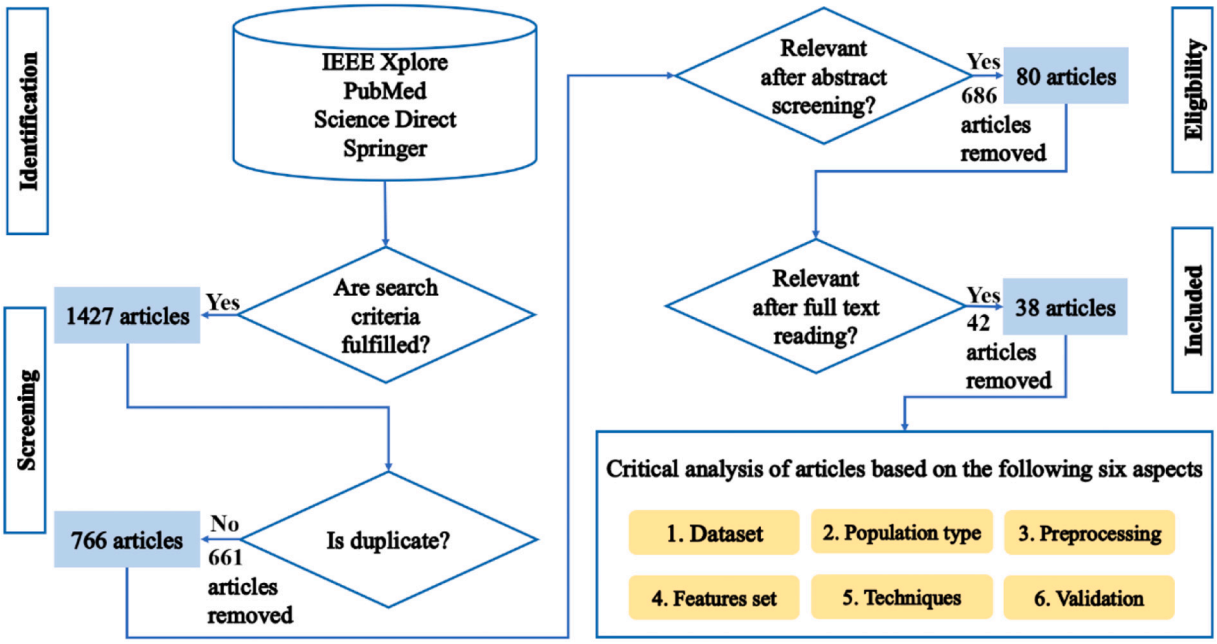


Fig. 1. Search methodology followed in the study.

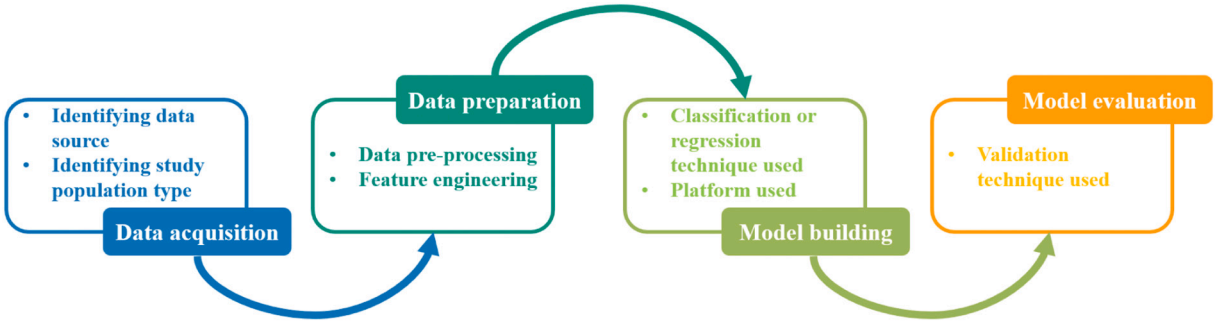


Fig. 2. Data mining steps for infection risk prediction.

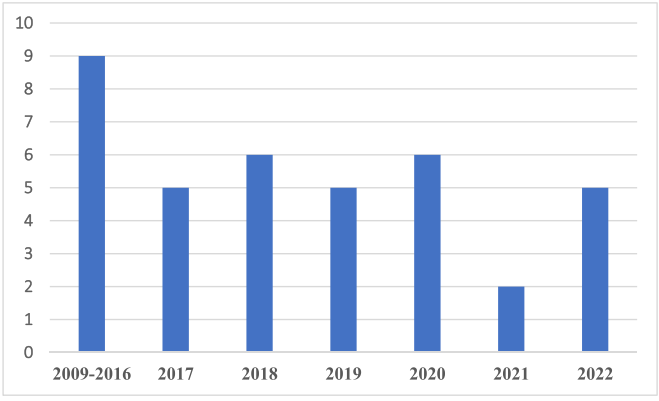


Fig. 3. Yearly distribution of the 38 studies.

21st century. However, after demonstrating promising trends and positive results in the medical field, these techniques have since become a popular topic among researchers.

3.2. HIV infection risk prediction

This section aims to provide a detailed analysis of the 38 existing literatures. Our analysis is oriented around seven aspects including data source, study population type, data pre-processing, feature set, model building techniques, platform or software used and validation, aiming to gain a better understanding of the dynamics involved and to facilitate a more insightful discussion.

3.2.1. Data source

The reliability and usability of results in the prediction process are significantly impacted by both the quality and quantity of the dataset used. The term "quantity" refers to the number of instances in a dataset, while "quality" is determined by the completeness and feature set of the dataset. Computational techniques for data cleaning and pre-processing can handle the handling of missing values or normalization. However, the availability of relevant and exciting feature sets is still scarce.

While many researchers use local hospitals or national repositories for their studies due to their availability and applicability to the local population, some researchers opt to capture data from thematic epidemiological investigations (randomized controlled trials, cohort or cross-sectional studies) that may have interesting data attributes covering more features. Table 1 summarizes the datasets used by each study in this survey. It was assumed that researchers primarily use local hospitals or national repositories due to their easy availability; however, as shown in the table, studies using local hospitals or national repositories are almost equally as common as those using randomized controlled trials, cohort or cross-sectional study datasets in our survey.

Table 1
Data sources used in the 38 studies.

Data source	Number of studies	Studies reference
Local hospitals or national repositories		
Melbourne Sexual Health Centre (MSHC), Australia	3	Bao Y et al. [6], Xu X et al. [7], Xu X et al. [8]
Acute and early HIV Test (AEH), USA	2	Hoenigl M et al. [9], Krakower DS et al. [10]
Public Health—Seattle & King County (PHSKC) STD Clinic electronic records	1	Menza TW et al. [5]
Los Angeles LGBT Center electronic records	1	Beymer MR et al. [11]
Clinical Data Warehouse (CDW) at New York Presbyterian Hospital – Columbia University Medical Center	1	Feller DJ et al. [12]
Kaiser Permanente Northern California electronic health record (EHR)	1	Marcus JL et al. [13]
The Danish National Hospital Registry (DNHR) registries	1	Ahlström MG et al. [14]
A clinical network record in Massachusetts at Atrius Health	1	Gruber S et al. [15]
MSM sentinel surveillance in Zhejiang province, China	1	He J et al. [16]
A metropolitan sexually transmitted disease clinic in Denver, Colorado	1	Haukoos JS et al. [17]
San Francisco City Clinic (SFCC)	1	Wahome E et al. [18]
Randomized controlled trial	5	Smith DK et al. [19], Balkus JE et al. [20], Wand H et al. [21], Balzer LB et al. [22], Peebles K et al. [23]
Cohort study	13	Facente SN et al. [24], Sanders EJ et al. [25], Dijkstra M et al. [26], Lin TC et al. [27], Lin TC et al. [28], Yun K et al. [29], Scott H et al. [30], Jones J et al. [31], Pintye J et al. [32], Lancki N et al. [33], Wahome E et al. [34], Luo Q et al. [35], Tordoff DM et al. [36]
Cross-sectional study	7	Hu P et al. [37], Sanders EJ et al. [25], Zheng M et al. [38], Kabapy AF et al. [39], Liu S et al. [40], Dong Y et al. [41], Yin L et al. [42]

3.2.2. Study population type

Table 2 presents the various types of study populations that were examined by the researchers. The table indicates that most researchers focused their infection risk prediction studies on men who have sex with men (MSM). Among the thirty-eight studies we selected, twenty-three exclusively looked at people with MSM, only five examined women, and the remaining ten studies examined the general population. The

Table 2
Study population type discussed in selected studies.

Study population type	Number of studies	Studies reference
MSM	23	Menza TW et al. [5], Smith DK et al. [19], Hoenigl M et al. [9], Beymer MR et al. [11], Jones J et al. [31], Lancki N et al. [33], Yin L et al. [42], Wahome E et al. [34], Luo Q et al. [35], Tordoff DM et al. [36], Bao Y et al. [6], He J et al. [16], Facente SN et al. [24], Hu P et al. [37], Sanders EJ et al. [25], Dijkstra M et al. [26], Lin TC et al. [28], Lin TC et al. [27], Yun K et al. [29], Scott H et al. [30], Zheng M et al. [38], Liu S et al. [40], Dong Y et al. [41]
Young women	3	Balkus JE et al. [20], Wand H et al. [21], Peebles K et al. [23]
Pregnant women	1	Pintye J et al. [32]
female sex workers	1	Sanders EJ et al. [25]
General population	11	Feller DJ et al. [12], Marcus JL et al. [13], Krakower DS et al. [10], Ahlström MG et al. [14], Balzer LB et al. [22], Gruber S et al. [15], Xu X et al. [7], Xu X et al. [8], Haukoos JS et al. [17], Wahome E et al. [18], Kabapy AF et al. [39]

rationale for this maybe that MSM is a high-risk group for HIV infection, with the proportion of HIV-infected individuals in this group increasing year by year [43]. However, upon reviewing the studies, it is apparent that there are no studies that focus on other populations at high risk of HIV infection, such as elderly male population, drug users, and immigration population [44].

3.2.3. Data pre-processing

This section examines the data pre-processing techniques used by researchers for infection risk prediction. Typically, data pre-processing for this application involves handling missing data, selecting appropriate features, and balancing the dataset.

A very few authors have used appropriate imputation techniques to handle missing data issues, while others have simply deleted instances with missing data. For instance, Bao Y et al. [6] used the mean value to replace missing values and demonstrated that the performance with imputation methods is much better than without imputation. Xu X et al. [7] solved the missing data problem using the random forest method, which can capture the correlations between different variables to estimate missing values more accurately without introducing artificial biases. In the same year, Xu X et al. [8] used another method that did not impute missing data but created a binary feature vector indicating missing values. The results showed that this method has lower error rates and higher efficiency than the imputation method.

As classification techniques are prone to overfitting when dealing with imbalanced datasets, several authors have attempted to address this issue through sampling techniques [45]. Over-sampling techniques aim to balance datasets by replicating existing minority class samples, while under-sampling technique reduces the majority of class samples [46]. Synthetic Minority Over-sampling Technique (SMOTE) have been the most commonly used balancing technique [47]. SMOTE balancing technique, reduces the risk of overfitting by generating new minority class samples through interpolation in the feature space, as opposed to under-sampling techniques. When dealing with significant class imbalances, SMOTE performs better in increasing the quantity of minority class samples because it generates new samples instead of simply replicating the existing ones. In a study conducted by He et al. [16], the use of SMOTE to balance the dataset resulted in an improved area under the receiver operating characteristic (ROC) curve (AUC) when compared to the original dataset across various machine learning methods for predicting infection risk. Other techniques for addressing imbalanced datasets include adaptive synthetic Sampling, sample weighting, threshold shifting, and other methods, but these were not utilized in the collected studies [48].

Except for the three studies that conducted external validation, all other articles utilized feature selection techniques. Table 3 presents the main feature selection techniques employed in the selected articles. Simple approaches include feature selection based on epidemiological evidence, expert opinions, and previous evaluations. Various regression methods, such as bivariate regression, stepwise regression, elastic net regression, and least absolute shrinkage and selection operator (lasso) regression, were widely used. Notably, both studies [10,13] published in Lancet HIV utilized lasso regression for feature selection and modeling, effectively reducing the number of features and achieving the highest AUC. Lasso regression combines linear regression with L1 regularization, which introduces an L1 regularization term to constrain the model's complexity and drives some feature coefficients to zero through coefficient shrinkage, thereby achieving the effect of feature selection [49]. In addition, Liu S et al. adopted three variable selection methods, including Boruta, Stepwise selection and Univariate selection. Boruta is based on the same idea as forming a random forest classifier, that is, by adding randomness to the system and collecting results from random sample sets, the misleading influence of random fluctuation and correlation can be reduced.

3.2.4. Feature set

In prediction analysis, the selection of features plays a crucial role. Understanding the various predictors available requires a comprehensive summary and classification of the features utilized in each study. Sociodemographic characteristics emerged as the prevailing feature category, encompassing fundamental information about the subjects. These characteristics provided essential insights into the social and demographic aspects of the individuals under study. These may include:

- Age
- Gender
- Marital status
- Race

Table 3
Some key feature selection techniques used in the 38 studies.

Feature selection technique	Number of studies	Studies reference
Simplicity	3	Menza TW et al. [5], Hoenigl M et al. [9], Wand H et al. [21]
Previous evaluation	3	Pintye J et al. [32], Xu X et al. [7], Xu X et al. [8]
Mutual information criteria	1	Feller DJ et al. [12]
Literature reference	1	Tordoff DM et al. [36]
Clinical expertise	1	Gruber S et al. [15]
Boruta	1	Liu S et al. [40]
Stepwise regression	4	Smith DK et al. [19], Balkus JE et al. [20], Yin L et al. [42], Liu S et al. [40]
Bivariate Cox regression	3	Beymer MR et al. [11], Yun K et al. [29], Peebles K et al. [23]
Bivariate Poisson regression	1	Wahome E et al. [34]
Univariate logistic regression	16	Lancki N et al. [33], Balzer LB et al. [22], Bao Y et al. [6], He J et al. [16], Facente SN et al. [24], Hu P et al. [37], Haukoos JS et al. [17], Wahome E et al. [18], Sanders EJ et al. [25], Dijkstra M et al. [26], Lin TC et al. [28], Lin TC et al. [27], Scott H et al. [30], Zheng M et al. [38], Kabapy AF et al. [39], Liu S et al. [40]
LASSO	3	Marcus JL et al. [13], Krakower DS et al. [10], Dong Y et al. [41]
Elastic network regression	1	Ahlström MG et al. [14]

Sexual behavior: All the characteristics pertaining to sexual activity within the past 3 to 6 months were considered. Examples of these characteristics include:

- Condom use
- Multiple sexual partners
- Group sex
- Commercial sex

Sexually transmitted infections (STIs): Study subjects who were diagnosed with STIs or had a history of STIs in their medical records were included. Examples of specific STDs that may be considered in these studies include:

- Syphilis
- Gonorrhea
- Genital Herpes
- Genital Warts

Intervention measures: The interventions related to HIV/AIDS that subjects received in the past were taken into account. These interventions encompassed a range of measures and tests, including:

- Publicity seminar
- Peer education
- HIV/AIDS counseling and testing
- Condom distribution

Symptoms: Symptoms of the body at time of testing or during 14 days prior to testing, including:

- Headache
- Pharyngitis
- Rash
- Myalgia

Health scale score: Scores on a series of scientific scales about physical health and mental health, such as self-esteem, loneliness and depression, including:

- Rosenberg self-esteem scale (RSES)
- Patient health questionnaire-9 (PHQ-9)
- Defeat scale (DS)
- Interpersonal needs questionnaire (INQ-15)

Medical records: Documents created and maintained by doctors, nurses, and other medical professionals during the diagnosis, treatment, and monitoring of patient health. Some records associated with HIV infection include:

- Diagnosis
- Laboratory tests
- Imaging examination
- Medication history

…

Table 4 summarizes the categories of features used in survey-based studies. It demonstrates that sociodemographic characteristics are consistently selected across local hospitals or national repositories as well as thematic epidemiological investigations, as they encompass basic information about all study subjects. However, there are differences in the types of information included in national and hospital databases versus thematic epidemiological investigations. Local hospitals or national repositories tend to include more information related to disease diagnoses, laboratory examinations and outcomes. On the other hand, thematic epidemiological investigations incorporate more features related to sexual behavior and interventions. In the prediction of acute HIV infection, the symptoms displayed by the subjects are the most important feature set [27,28].

Table 4
Categories of features used in the 38 studies.

Feature category	Studies reference
Sociodemographic	Menza TW et al. [5], Smith DK et al. [19], Hoenigl M et al. [9], Beymer MR et al. [11], Lancki N et al. [33], Yin L et al. [42], Wahome E et al. [34], Luo Q et al. [35], Tordoff DM et al. [36], Bao Y et al. [6], He J et al. [16], Balkus JE et al. [20], Wand H et al. [21], Pintye J et al. [32], Feller DJ et al. [12], Marcus JL et al. [13], Krakower DS et al. [10], Ahlström MG et al. [14], Balzer LB et al. [22], Gruber S et al. [15], Xu X et al. [7], Xu X et al. [8], Facente SN et al. [24], Hu P et al. [37], Haukoos JS et al. [17], Wahome E et al. [18], Sanders EJ et al. [25], Lin TC et al. [28], Lin TC et al. [27], Yun K et al. [29], Scott H et al. [30], Zheng M et al. [38], Peebles K et al. [23], Kabapy AF et al. [39], Liu S et al. [40], Dong Y et al. [41]
Sexual behavior	Menza TW et al. [5], Smith DK et al. [19], Hoenigl M et al. [9], Beymer MR et al. [11], Lancki N et al. [33], Yin L et al. [42], Wahome E et al. [34], Luo Q et al. [35], Tordoff DM et al. [36], Bao Y et al. [6], He J et al. [16], Balkus JE et al. [20], Wand H et al. [21], Pintye J et al. [32], Feller DJ et al. [12], Xu X et al. [7], Xu X et al. [8], Hu P et al. [37], Haukoos JS et al. [17], Wahome E et al. [18], Dijkstra M et al. [26], Yun K et al. [29], Scott H et al. [30], Zheng M et al. [38], Peebles K et al. [23], Kabapy AF et al. [39], Liu S et al. [40], Dong Y et al. [41]
STIs	Menza TW et al. [5], Hoenigl M et al. [9], Beymer MR et al. [11], Yin L et al. [42], Wahome E et al. [34], Luo Q et al. [35], Tordoff DM et al. [36], Bao Y et al. [6], He J et al. [16], Balkus JE et al. [20], Wand H et al. [21], Pintye J et al. [32], Feller DJ et al. [12], Marcus JL et al. [13], Krakower DS et al. [10], Ahlström MG et al. [14], Balzer LB et al. [22], Gruber S et al. [15], Xu X et al. [7], Xu X et al. [8], Hu P et al. [37], Yun K et al. [29], Scott H et al. [30], Zheng M et al. [38], Peebles K et al. [23], Kabapy AF et al. [39], Liu S et al. [40]
Interventions	He J et al. [16], Lancki N et al. [33]
Symptoms	Sanders EJ et al. [25], Dijkstra M et al. [26], Lin TC et al. [28], Lin TC et al. [27]
Health scale score	Liu S et al. [40], Dong Y et al. [41]
Medical records	Bao Y et al. [6], Feller DJ et al. [12], Marcus JL et al. [13], Krakower DS et al. [10], Ahlström MG et al. [14], Gruber S et al. [15], Xu X et al. [7], Xu X et al. [7], Facente SN et al. [24]
Smoking/alcohol intake	Scott H et al. [30], Kabapy AF et al. [39], Liu S et al. [40], Dong Y et al. [41]
Drug use	Haukoos JS et al. [17], Wahome E et al. [18], Beymer MR et al. [11], Lancki N et al. [33], Feller DJ et al. [12], Wahome E et al. [34], Luo Q et al. [35], Yun K et al. [29], Zheng M et al. [38], Kabapy AF et al. [39], Bao Y et al. [6], Xu X et al. [7], Xu X et al. [8]

In recent years, there has been an increasing recognition of the importance of incorporating mental health as a predictive feature in research studies [40,41]. Additionally, the inclusion of smoking and drinking habits has also been observed [39–41]. This could be attributed to the fact that individuals in disadvantaged areas are more prone to engaging in unprotected sexual activities following smoking/alcohol intake [46].

3.2.5. Model building techniques

Table 5 presents the machine learning modeling techniques used in the 38 articles. It is undeniable that traditional regression techniques are still the most widely used predictive model construction techniques. In 2009, the Menza team established the first HIV infection risk prediction model for men who have sex with men using Cox regression techniques. However, due to the complexity of medical data, traditional regression approaches fall short in accurately modeling its complexity [50]. In response, researchers have extensively utilized fully supervised machine learning methods such as random forest, K-Nearest Neighbors (KNN),

Table 5
Categories of techniques used for prediction in the 38 studies.

Techniques	Studies reference
Cox regression	Menza TW et al. [5], Balkus JE et al. [20], Beymer MR et al. [11], Pintye J et al. [32], Wand H et al. [21], Tordoff DM et al. [36], Yun K et al. [29], Peebles K et al. [23]
Poisson regression	Lancki N et al. [33], Wahome E et al. [34]
Generalized Estimating Equations (GEE)	Smith DK et al. [19], Facente SN et al. [24], Sanders EJ et al. [25], Dijkstra M et al. [26]
Logistic regression	Hoenigl M et al. [9], Yin L et al. [42], Krakower DS et al. [10], Ahlström MG et al. [14], Gruber S et al. [15], Bao Y et al. [6], He J et al. [16], Xu X et al. [7], Xu X et al. [8], Hu P et al. [37], Haukoos JS et al. [17], Wahome E et al. [18], Lin TC et al. [28], Lin TC et al. [27], Scott H et al. [30], Zheng M et al. [38], Liu S et al. [40], Dong Y et al. [41]
Random forest	Feller DJ et al. [12], Marcus JL et al. [13], Krakower DS et al. [10], Ahlström MG et al. [14], Gruber S et al. [15], He J et al. [16], Xu X et al. [16], Xu X et al. [8]
LASSO	Marcus JL et al. [13], Krakower DS et al. [10], Ahlström MG et al. [14], Gruber S et al. [15], Xu X et al. [8]
Ridge regression	Krakower DS et al. [10], Ahlström MG et al. [14], Gruber S et al. [15], Xu X et al. [8]
Elastic network regression	Krakower DS et al. [10], Ahlström MG et al. [14], Gruber S et al. [15], Xu X et al. [8]
Naïve Bayes	Xu X et al. [8]
KNN	Xu X et al. [7]
SVM	Krakower DS et al. [10], Gruber S et al. [15], He J et al. [16], Xu X et al. [7]
Decision Tree	He J et al. [16]
Gradient Boosting Machine	Bao Y et al. [6]
Neural Networks	Xu X et al. [8]
Deep learning	Bao Y et al. [6]
XG Boosting	Bao Y et al. [6]
Stacking ensemble learning	Balzer LB et al. [22], Xu X et al. [7], Xu X et al. [8]

and Support Vector Machine (SVM). In recent years, medical models based on deep learning techniques exhibit the ability to capture intricate details and patterns within the data [51], researchers can harness the power of unlabeled data and delve deeper into the complex interplay of factors influencing HIV infection risk.

There are a wide variety of machine learning modeling techniques, and choosing the best modeling technique has become a new problem. The first type of researchers choose to use multiple modeling techniques at the same time and select the model with the best predictive effect. For instance, Bao Y et al. and others used logistic regression, random forests, Gradient Boosting Machine (GBM), and Extreme Gradient Boosting (XGBoost) at the same time, and ultimately GBM showed the best predictive effect. The second type of researchers opts for stacking ensemble learning (an ensemble learning method that trains a new model based on the combined predictions of 2 or more previous machine learning models) has also been explored in the latest studies to generate better performs than individual machine learning techniques. Fig. 4 depicts the twenty-six ensemble learning models developed by Xu X et al. These models are derived from various combinations of the five base learner models. Specifically, there are ten models that result from combinations of two distinct base learner models, ten additional models that incorporate combinations of three distinct base learner models, five models that are formed by combining four distinct base learner models, and one model that integrates all five base learner models. What's more, some scholars have established a tool called Super Learner which is an R-based tool that applies ensemble learning by integrating multiple weighted classifiers, leveraging cross-validation techniques. The tool supports many classification and regression algorithms, such as random forest, LASSO, and support vector machines, and the preferred algorithms can be selected with very simple operations.

3.2.6. *Platform or software used*

Table 6 has been compiled with the intention of providing a comprehensive overview of the software employed in constructing the

Table 6
Platforms or Software used in studies.

Platform/software	Number of studies
R	20
STATA	7
SAS	5
SPSS	5
Python	2
MATLAB	1

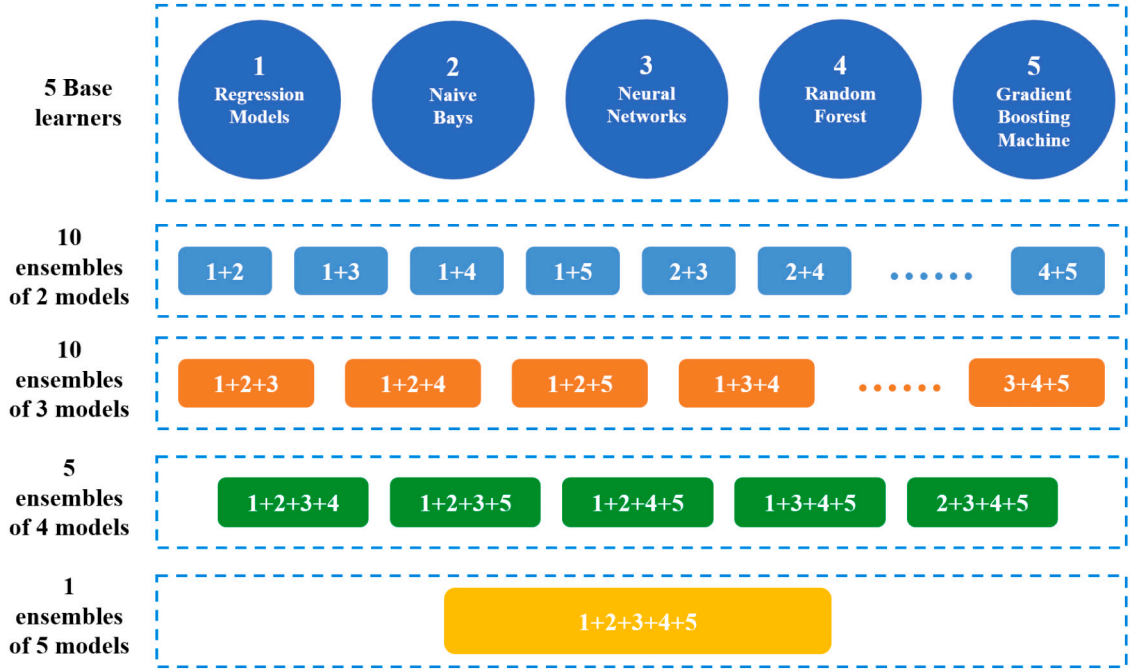


Fig. 4. Twenty-six ensemble learning models combined five base models.

prediction models, with the aim of offering clarity to researchers. While some researchers did not disclose or specify the software utilized for building their prediction models, the majority of researchers opted for the R package. The primary reason for this preference may be its open-access availability. it is evident that approximately 70 % of the studies utilized open-access tools for conducting their experiments and evaluations. It is worth noting that software such as STATA and SAS require a subscription or purchase as they are licensed. Several authors employed multiple tools for their analysis. For example, SPSS was utilized for statistical analysis, while the R package was employed for creating graphs or models [27]. Xu X et al. utilized the R package for model building, while MATLAB was employed for computing information gain [7]. Furthermore, it is apparent that software such as Python and MATLAB are not commonly utilized among machine learning researchers.

3.2.7. Validation

During the infection risk prediction study, the final step involves validating the proposed approach using various measures. While accuracy, sensitivity, and Area Under Curve (AUC) were commonly employed in the selected articles, the specific validation measures varied among researchers. Fig. 5(a) depicts the different validation methods utilized by the researchers in their studies.

Approximately 33 % of the studies utilized the K-fold cross-validation technique. Consequently, Fig. 5(b) illustrates a distinct classification of different cross-validation techniques. Researchers commonly opted for K-fold cross-validation to minimize bias in their results. However, it does not guarantee consistent outcomes. Following K-fold cross-validation, the holdout method emerged as the preferred choice among researchers. Furthermore, 8 out of the total 38 studies incorporated the nested cross-validation technique to ensure fair performance comparisons.

Two studies published in *Lancet HIV* employed ten folds for prospective cross-validation, with each fold further divided into ten folds for internal cross-validation [10,13]. In contrast, Xu X et al. utilized five folds for outer cross-validation and ten folds for inner cross-validation [7,8]. Fig. 6 visually presents the structure of the nested validation scheme, consisting of ten outer folds and ten inner folds. One inner set is designated as the test set (TS), one as the validation set (VL), and the remaining eight inner sets for training purposes (TR).

4. Open issues and possible future aspects

In 2013, UNAIDS reported a total of 3.7 million individuals infected with HIV. This number has significantly increased to approximately 38 million by 2022 [52]. As mentioned earlier, researchers have made efforts to develop various models for assessing infection risk. While some authors have pointed out the limitations of existing models and attempted to address these issues, there are still numerous drawbacks and challenges that need to be addressed to mitigate the growing burden of HIV on a global scale. This section aims to highlight these issues and discuss potential future agendas for tackling them. Furthermore, it recognizes some approaches taken by researchers in addressing these problems to provide better understanding and relevant examples.

4.1. Dataset interoperability

There remains a dearth of studies that effectively leverage and integrate heterogeneous data in the field. To optimize the utilization of such diverse data, the incorporation of advanced computational models and improved data integration frameworks becomes indispensable. In America, health care data standards have been widely embraced, offering defined protocols and methodologies for recording, storing, and sharing data within medical institutions [53]. Nevertheless, the current state of standardization still falls short in guaranteeing seamless interoperability (i.e., the ability to access, exchange, integrate, and collaboratively utilize data in a coordinated manner) and shareability of public health data. This issue is particularly relevant in the realm of HIV research. By employing semantic technologies, such as ontologies, during the creation and representation of intricate information and the associations between concepts, one can effectively interpret information by identifying the pertinent context and aligning data with established terminologies [54]. It is imperative to urgently develop additional standardized ontologies, terminologies, and common data elements specific to HIV research. This concerted effort will ensure enhanced data interoperability and facilitate the seamless integration and sharing of valuable research findings.

4.2. Establishing a comprehensive HIV/AIDS registration system

While researchers have access to certain large datasets, it is important to note that these datasets often capture information from a restricted population subset. Both EHR datasets and thematic epidemiological survey datasets typically encompass only a specific country or region's population. Notably, the impact of the same infection risk rating scale can vary across different countries and races, as confirmed by recent empirical studies [31]. Consequently, there arises a need for a more comprehensive HIV infection registry that encompasses patients from diverse countries, races, and religions [31]. Such a registry would effectively accommodate the heterogeneous and homogeneous characteristics of patients, enabling a more comprehensive understanding of the disease.

4.3. Limited types of subjects of study

The existing body of literature in HIV research has primarily

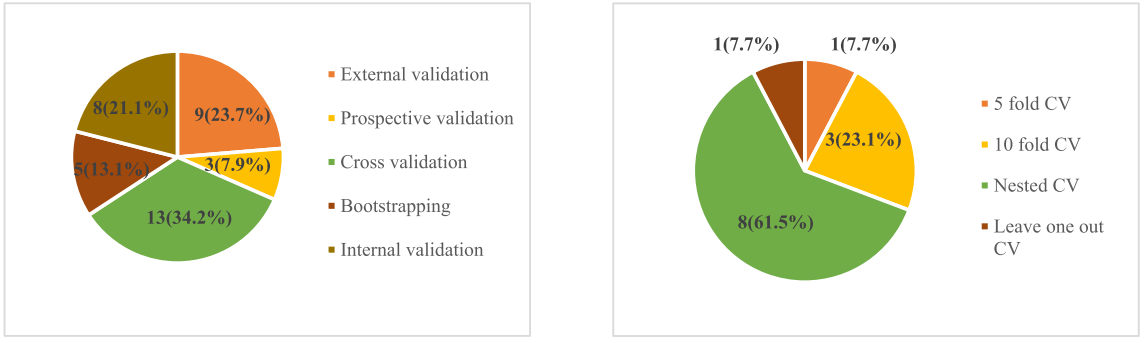


Fig. 5. Validation techniques used by the 38 studies.

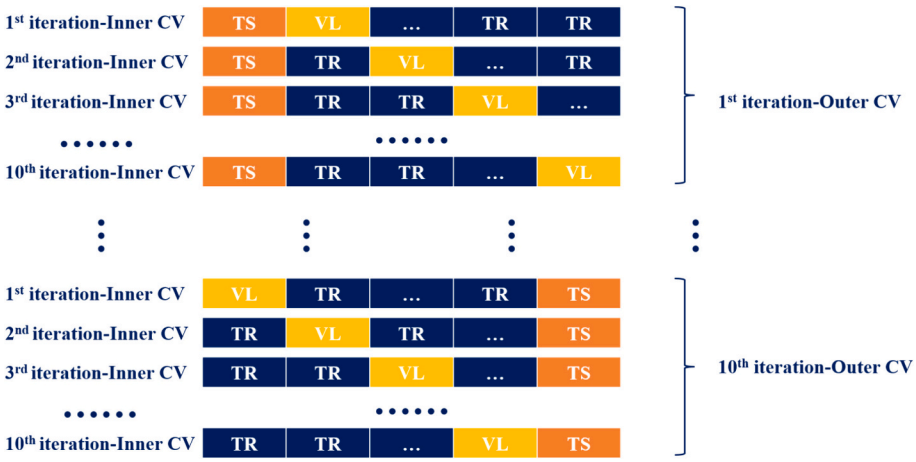


Fig. 6. Nested cross-validation example.

concentrated on a limited range of study population types. Although it is well established that MSM bear the highest HIV infection rates worldwide, it is vital to recognize that effective prevention strategies should extend beyond this demographic. Numerous high-risk groups, including drug addicts, young students, elderly men (whose HIV incidence has exhibited a concerning rise in recent years), as well as individuals returning from international travel following the COVID-19 pandemic, require careful attention [55].

To mitigate the HIV epidemic, it is imperative to explore different study population types comprehensively. By encompassing a broader range of demographics, researchers can gain valuable insights and develop prevention strategies that cater to the unique needs and challenges faced by these populations. Comprehensive efforts are needed to ensure that prevention measures are universally applicable, leaving no population group behind.

4.4. Missing value processing

It is widely acknowledged that medical data are highly prone to missing and inconsistent data. The process of classifying datasets into labeled classes represents only one stage in the data mining process, also known as knowledge discovery from data. However, without a high-quality dataset, machine learning techniques may not yield accurate results. Therefore, it is crucial for researchers to preprocess the dataset before applying any classification or regression techniques.

In the previous section, it was apparent that many researchers do not effectively address missing data. Fortunately, there exist several successful techniques for handling missing data that can be employed during the preprocessing stage. However, it is important to note that a dataset with a substantial amount of missing data can produce unreliable results, potentially leading to an overfitted model. As a result, a more effective approach might involve removing instances with more than a predefined threshold value (e.g., 50 %) of missing data and applying imputation techniques to the remaining attributes [56]. A similar approach can be applied when deciding whether to retain or discard specific attributes during the preprocessing phase. By adopting such strategies, researchers can mitigate the impact of missing data and enhance the overall quality of the dataset, thus improving the reliability of subsequent analyses and models.

4.5. Feature sets

In most studies, sociodemographic feature sets are utilized to predict infection risk. However, it is important to note that the sociodemographic characteristics collected by different research groups may vary. Additionally, sexual behavior characteristics play a crucial role in predicting the risk of HIV infection, but such information is often lacking in EHR.

For future work, it is valuable to explore the design of infection risk prediction models that encompass multiple categories of characteristics. By incorporating a broader range of characteristic attributes, researchers can potentially improve the accuracy and comprehensiveness of infection risk prediction models. Furthermore, it is important to acknowledge that characteristic attributes may vary depending on the specific study population. However, there is still a need to identify certain characteristic attributes that can be universally applied across different populations for infection risk prediction [4]. By determining these universal attributes, researchers can establish a foundation for consistent and reliable risk prediction models that can be widely applied.

4.6. Use of unsupervised and semi-supervised approaches for infection risk analysis

Among the studies we reviewed, it was evident that the majority employed full-supervised learning methods for infection risk prediction models. These approaches have demonstrated superior performance when ample labeled datasets are available for training. However, the collection and management of medical datasets pose significant challenges. As a result, many studies opted to remove instances without class labels. While this is a more reliable approach, it is important to recognize that unlabeled instances can contain valuable information [57]. Surprisingly, only a few of the selected studies explored the use of unsupervised or semi-supervised classification techniques for infection risk prediction.

Unsupervised and semi-supervised classification techniques can be particularly advantageous when dealing with medical datasets, especially in cases where the infection risk outcomes of certain patients cannot be obtained. These techniques offer a promising avenue for extracting meaningful patterns and insights from such datasets. Therefore, it is imperative to further explore and utilize unsupervised and semi-supervised classification approaches to develop prediction models in the medical domain [58]. By doing so, researchers can tap into the potential of unlabeled instances and leverage the full range of available data to improve the accuracy and robustness of infection risk prediction models.

4.7. Use of other bio-inspired computing approaches

While researchers are actively seeking ways to improve infection risk prediction results, there remains a lack of methodological exploration in the current studies. Meta-heuristic techniques have gained prominence in various classification fields as they offer the potential to enhance

results. However, their application in HIV infection risk prediction studies is still relatively unexplored.

In a notable study by Wang et al., Particle Swarm Optimization was employed for feature selection in conjunction with traditional machine learning techniques, resulting in improved performance [59]. Nevertheless, the use of diverse meta-heuristic algorithms, and potentially hyper-heuristic algorithms, remains largely untapped in this context. Exploring the application of these algorithms in HIV infection risk prediction studies presents an opportunity for novel advancements.

As a potential future research agenda, leveraging algorithms such as Cuckoo search, Flower Pollination algorithm, and others in conjunction with suitable machine learning approaches holds promise for achieving better prediction outcomes [60]. By combining these meta-heuristic algorithms with appropriate machine learning techniques, researchers can potentially unlock new insights and achieve enhanced results in the field of HIV infection risk prediction.

4.8. Communication barrier between AI and medical workers

In recent years, there has been a growing utilization of neural networks and deep learning techniques among researchers in the field of infection risk prediction. It is worth noting that these models often require a substantial amount of time for training due to their complexity. The reason behind this choice is the high-performance results achieved by neural networks. However, it is important to acknowledge that most machine learning approaches, including neural networks, are often perceived as black boxes by medical workers [61].

While these models can provide estimations of infection risk, such as high or low, medical workers often struggle to grasp the intricate details or reasoning behind these results. It is crucial for medical workers to be able to explain to the subjects the possible outcomes and the appropriate underlying reasons. In this context, the concept of explainable AI (XAI) becomes relevant. By incorporating XAI techniques, medical workers can better comprehend and trust the results generated by prediction models.

The inclusion of an explanation interface in prediction models can lead to the development of a responsible AI system. This interface would enable medical workers to gain insights into the decision-making process of the models and understand the factors influencing the predictions. By embracing XAI, medical workers can effectively communicate the results to subjects, providing them with a comprehensive understanding of the predicted infection risk and the rationale behind it. This promotes transparency, accountability, and fosters trust in the AI system.

4.9. Validation techniques

Among the selected research studies, k-fold cross-validation emerges as the most commonly employed validation method. While many studies have undergone external verification, it is important to note that some studies still lack external validation of their results, meaning that the results are verified using the same dataset on which they were trained. This approach can present challenges, as the error incurred from evaluating on the same training data tends to be relatively small.

As part of future plans, incorporating multiple datasets with similar attributes can be beneficial for external verification purposes. By utilizing different datasets for validation, researchers can obtain a more comprehensive and reliable assessment of the performance and generalizability of their models. This approach enhances the credibility and robustness of the research findings, facilitating a deeper understanding of the strengths and limitations of existing technologies. Consequently, establishing a standardized experimental setup and promoting the use of diverse datasets for external validation will contribute to the advancement and reproducibility of research in the field.

5. Conclusion

Predicting the risk of HIV infection within a population is of utmost importance in the timely detection of infected individuals. In order to present a thorough and insightful overview in the past years, we carefully selected 38 articles employing data mining techniques to assess HIV infection risk. A majority of these articles were published post-2017. Our study further encompassed a detailed analysis of various aspects within this domain. Specifically, we examined the utilization of different public domains as data sources, revealing that local hospitals or national repositories, as well as thematic epidemiological investigations, were the most commonly employed sources.

Among the targeted populations, researchers primarily focused on men who have sex with men (MSM) due to their globally highest infection rates. However, it is imperative to conduct detailed analyses of other study populations as well. Evaluating the preprocessing techniques employed in these studies, we observed that while many researchers utilized feature selection techniques to identify critical dataset features, the majority did not employ imputation techniques to handle missing data.

Notably, researchers are increasingly turning to the latest machine learning techniques, such as neural networks, ensemble approaches, and deep learning techniques, for classification purposes. These techniques are validated through various validation methods, including k-fold cross-validation and the holdout method. Researchers commonly rely on open-source tools, including R packages and Python, for their data analysis.

Conclusively, we have outlined nine challenges prevalent in the existing literature and provided future recommendations for researchers in this field. This comprehensive review serves as a roadmap for both novice researchers seeking to comprehend the current state-of-the-art in this area and experienced researchers aiming to identify key issues and potential areas for improvement in terms of performance and reliability.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to improve readability and language. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no conflict of interest.

Acknowledgements

This work was supported by the National Science and Technology Major Project on A Multimodality model guided tailored screening, precise diagnosis, and prevention strategy (MODERN) for HIV Prevention and Control (2022YFC2304901). The project was also supported by the Graduate Research & Practice Innovation Program of Jiangsu Province, China (KYCX23_3435) and Preventive Medicine Research Project of Jiangsu Provincial Health Commission (Ym2023079).

We would like to express our deepest gratitude to Chen Cui and Wenjie Jiang for their invaluable assistance during the revision process of this manuscript. Their expert insights in addressing the reviewers' comments and their meticulous care in fine-tuning the language significantly contributed to the refinement of this work.

References

[1] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25(1):44–56. https://doi.org/10.1038/s41591-018-0300-7.

[2] Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018;1:18. https://doi.org/10.1038/s41746-018-0029-1.
[3] Bisaso KR, Anguzu GT, Karungi SA, Kiragga A, Castelnuovo B. A survey of machine learning applications in HIV clinical research and care. Comput Biol Med. 2017;91:366–371. doi:10.10 16/j.compbiomed.2017.11.001.
[4] Xiang Y, Du J, Fujimoto K, Li F, Schneider J, Tao C. Application of artificial intelligence and machine learning for HIV prevention interventions. Lancet HIV. 2022;9(1):e54-e62. doi:10. 1016/S2352–3018(21)00247–2.
[5] Menza TW, Hughes JP, Celum CL, Golden MR. Prediction of HIV acquisition among men who have sex with men. Sex Transm Dis 2009;36(9):547–55. https://doi.org/10.1097/OLQ.0b013e3181a9cc41.
[6] Bao Y, Medland NA, Fairley CK, et al. Predicting the diagnosis of HIV and sexually transmitted infections among men who have sex with men using machine learning approaches. J Infect 2021;82(1):48–59. https://doi.org/10.1016/j.jinf.2020.11.007.
[7] Xu X, Ge Z, Chow EPF, et al. A machine-learning-based risk-prediction tool for HIV and sexually transmitted infections acquisition over the next 12 months. JCM 2022;11(7):1818. https://doi.org/10.3390/jcm11071818.
[8] Xu X, Yu Z, Ge Z, et al. Web-based risk prediction tool for an individual's risk of HIV and sexually transmitted infections using machine learning algorithms: development and external validation study. J Med Internet Res 2022;24(8):e37850. https://doi.org/10.2196/37850.
[9] Hoenigl M, Weibel N, Mehta SR, et al. Development and validation of the San Diego Early Test Score to predict acute and early HIV infection risk in men who have sex with men. Clin Infect Dis 2015;61(3):468–75. https://doi.org/10.1093/cid/civ335.
[10] Krakower DS, Gruber S, Hsu K, et al. Development and validation of an automated HIV prediction algorithm to identify candidates for pre-exposure prophylaxis: a modelling study. The Lancet HIV 2019;6(10):e696–704. https://doi.org/10.1016/S2352-3018(19)30139-0.
[11] Beymer MR, Weiss RE, Sugar CA, et al. Are Centers for Disease Control and Prevention guidelines for preexposure prophylaxis specific enough? Formulation of a personalized HIV risk score for pre-exposure prophylaxis initiation. Sexual Trans Dis 2017;44(1):49–57. https://doi.org/10.1097/OLQ.0000000000000535.
[12] Feller DJ, Zucker J, Yin MT, Gordon P, Elhadad N. Using clinical notes and natural language processing for automated HIV risk assessment. JAIDS Journal of Acquired Immune Deficiency Syndromes 2018;77(2):160–6. https://doi.org/10.1097/QAI.0000000000001580.
[13] Marcus JL, Hurley LB, Krakower DS, Alexeeff S, Silverberg MJ, Volk JE. Use of electronic health record data and machine learning to identify candidates for HIV pre-exposure prophylaxis: a modelling study. The Lancet HIV 2019;6(10):e688–95. https://doi.org/10.1016/S2352-3018(19)30137-7.
[14] Ahlström MG, Ronit A, Omland LH, Vedel S, Obel N. Algorithmic prediction of HIV status using nation-wide electronic registry data. EClinicalMedicine 2019;17:100203. https://doi.org/10.1016/j.eclinm.2019.10.016.
[15] Gruber S, Krakower D, Menchaca JT, et al. Using electronic health records to identify candidates for human immunodeficiency virus pre-exposure prophylaxis: an application of super learning to risk prediction when the outcome is rare. Stat Med 2020;39(23):3059–73. https://doi.org/10.1002/sim.8591.
[16] He J, Li J, Jiang S, et al. Application of machine learning algorithms in predicting HIV infection among men who have sex with men: model development and validation. Front Public Health 2022;10:967681. https://doi.org/10.3389/fpubh.2022.967681.
[17] Haukoos JS, Lyons MS, Lindsell CJ, et al. Derivation and validation of the Denver human immunodeficiency virus (HIV) risk score for targeted HIV screening. Am J Epidemiol 2012;175(8):838–46. https://doi.org/10.1093/aje/kwr389.
[18] Wahome E, Fegan G, Okuku HS, et al. Evaluation of an empiric risk screening score to identify acute and early HIV-1 infection among MSM in Coastal Kenya. AIDS 2013;27(13):2163–6. https://doi.org/10.1097/QAD.0b013e3283629095.
[19] Smith DK, Pals SL, Herbst JH, Shinde S, Carey JW. Development of a clinical screening index predictive of incident HIV infection among men who have sex with men in the United States. JAIDS Journal of Acquired Immune Deficiency Syndromes 2012;60(4):421–7. https://doi.org/10.1097/QAI.0b013e318256b2f6.
[20] Balkus JE, Brown E, Palanee T, et al. An empiric HIV risk scoring tool to predict HIV-1 acquisition in African women. JAIDS Journal of Acquired Immune Deficiency Syndromes 2016;72(3):333–43. https://doi.org/10.1097/QAI.0000000000000974.
[21] Wand H, Reddy T, Naidoo S, et al. A simple risk prediction algorithm for HIV transmission: results from HIV prevention trials in KwaZulu Natal, South Africa (2002–2012). AIDS Behav 2018;22(1):325–36. https://doi.org/10.1007/s10461-017-1785-7.
[22] Balzer LB, Havlir DV, Kamya MR, et al. Machine learning to identify persons at high-risk of human immunodeficiency virus acquisition in rural Kenya and Uganda. Clin Infect Dis 2020;71(9):2326–33. https://doi.org/10.1093/cid/ciz1096.
[23] Peebles K, Palanee-Phillips T, Balkus JE, et al. Age-specific risk scores do not improve HIV-1 prediction among women in South Africa. J Acquir Immune Defic Syndr 2020;85(2):156–64. https://doi.org/10.1097/QAI.0000000000002436.
[24] Facente SN, Pilcher CD, Hartogensis WE, et al. Performance of risk-based criteria for targeting acute HIV screening in San Francisco. PloS One 2011;6(7):e21813. https://doi.org/10.1371/journal.pone.0021813.
[25] Sanders EJ, Wahome E, Powers KA, et al. Targeted screening of at-risk adults for acute HIV-1 infection in sub-Saharan Africa. AIDS. 2015;29 Suppl 3(03):S221–230. doi:https://doi.org/10.1097/QAD. 0000000000000924.
[26] Dijkstra M, de Bree GJ, Stolte IG, et al. Development and validation of a risk score to assist screening for acute HIV-1 infection among men who have sex with men. BMC Infect Dis 2017;17(1):425. https://doi.org/10.1186/s12879-017-2508-4.
[27] Lin TC, Gianella S, Tenenbaum T, Little SJ, Hoenigl M. A simple symptom score for acute human immunodeficiency virus infection in a San Diego Community-Based Screening Program. Clin Infect Dis 2018;67(1):105–11. https://doi.org/10.1093/cid/cix1130.
[28] Lin TC, Dijkstra M, De Bree GJ. Schim van der Loeff MF, Hoenigl M. Brief Report: the Amsterdam symptom and risk-based score predicts for acute HIV infection in men who have sex with men in San Diego. J Acquir Immune Defic Syndr 2018;79(2):e52–5. https://doi.org/10.1097/QAI.0000000000001800.
[29] Yun K, Xu J, Leuba S, et al. Development and validation of a personalized social media platform-based HIV incidence risk assessment tool for men who have sex with men in China. J Med Internet Res 2019;21(6):e13475. https://doi.org/10.2196/13475.
[30] Scott H, Vittinghoff E, Irvin R, et al. Development and validation of the personalized sexual health promotion (SexPro) HIV risk prediction model for men who have sex with men in the United States. AIDS Behav 2020;24(1):274–83. https://doi.org/10.1007/s10461-019-02616-3.
[31] Jones J, Hoenigl M, Siegler AJ, Sullivan PS, Little S, Rosenberg E. Assessing the performance of 3 human immunodeficiency virus incidence risk scores in a cohort of Black and White Men who have sex with men in the South. Sexual Trans Dis. 2017;44(5):297–302. doi:10.10 97/OLQ.0000000000000596.
[32] Pintye J, Drake AL, Kinuthia J, et al. A risk assessment tool for identifying pregnant and postpartum women who may benefit from pre-exposure prophylaxis (PrEP). CLINID. Published online December 28, 2016:ciw850. doi:https://doi.org/10.1093/cid/ciw850.
[33] Lancki N, Almirol E, Alon L, McNulty M, Schneider JA. Preexposure prophylaxis guidelines have low sensitivity for identifying seroconverters in a sample of young Black MSM in Chicago. AIDS 2018;32(3):383–92. https://doi.org/10.1097/QAD.0000000000001710.
[34] Wahome E, Thiong'o AN, Mwashigadi G, et al. An empiric risk score to guide PrEP targeting among MSM in coastal Kenya. AIDS Behav 2018;22(S1):35–44. https://doi.org/10.1007/s10461-018-2141-2.
[35] Luo Q, Huang X, Li L, et al. External validation of a prediction tool to estimate the risk of human immunodeficiency virus infection amongst men who have sex with men. Medicine 2019;98(29):e16375. https://doi.org/10.1097/MD.0000000000016375.
[36] Tordoff DM, Barbee LA, Khosropour CM, Hughes JP, Golden MR. Derivation and validation of an HIV risk prediction score among gay, bisexual, and other men who have sex with men to inform PrEP initiation in an STD clinic setting. JAIDS Journal of Acquired Immune Deficiency Syndromes 2020;85(3):263–71. https://doi.org/10.1097/QAI.0000000000002438.
[37] Hu P, Zhong F, Cheng WB, Xu HF, Ling L. Study on the infectious risk model of AIDS among men who have sex with men in Guangzhou. Zhonghua Liu Xing Bing Xue Za Zhi 2012;33(7):667–71.
[38] Zheng M, He J, Yuan Z, et al. Risk assessment and identification of HIV infection among men who have sex with men: a cross-sectional study in Southwest China. BMJ Open. 2020;10(11):e 039557. doi:https://doi.org/10.1136/bmjopen-2020-039557.
[39] Kabapy AF, Shatat HZ, Abd El-Wahab EW. Identifying factors increasing the risk of acquiring HIV among Egyptians to construct a consensus web-based tool for HIV risk assessment. Curr Med Res Opin 2021;37(6):973–84. https://doi.org/10.1080/03007995.2021.1901678.
[40] Liu S, Xia D, Wang Y, et al. Predicting the risk of HIV infection among internal migrant MSM in China: an optimal model based on three variable selection methods. Front Public Health 2022;10:1015699. https://doi.org/10.3389/fpubh.2022.1015699.
[41] Dong Y, Liu S, Xia D, et al. Prediction model for the risk of HIV infection among MSM in China: validation and stability. Int J Environ Res Public Health. 2022;19(2):1010. doi:10. 3390/ijerph19021010.
[42] Yin L, Zhao Y, Peratikos MB, et al. Risk prediction score for HIV infection: development and internal validation with cross-sectional data from men who have sex with men in China. AIDS Behav 2018;22(7):2267–76. https://doi.org/10.1007/s10461-018-2129-y.
[43] Nevendorff L, Schroeder SE, Pedrana A, Bourne A, Stoové M. Prevalence of sexualized drug use and risk of HIV among sexually active MSM in East and South Asian countries: systematic review and meta-analysis. J Int AIDS Soc 2023;26(1):e26054. https://doi.org/10.1002/jia2.26054.
[44] Collins PY, Velloza J, Concepcion T, et al. Intervening for HIV prevention and mental health: a review of global literature. J Int AIDS Soc. 2021;24 Suppl 2(Suppl 2):e25710. doi:https://doi.org/10.1002/jia2. 25710.
[45] Ma S, Huang J. Penalized feature selection and classification in bioinformatics. Brief Bioinform 2008;9(5):392–403. https://doi.org/10.1093/bib/bbn027.
[46] Liu X, Li N, Liu S, et al. Normalization methods for the analysis of unbalanced transcriptome data: a review. Front Bioeng Biotechnol 2019;7:358. https://doi.org/10.3389/fbioe.2019.00358.
[47] Albaradei S, Thafar M, Alsaedi A, et al. Machine learning and deep learning methods that use omics data for metastasis prediction. Comput Struct Biotechnol J. 2021;19:5008–5018. doi:10. 1016/j.csbj.2021.09.001.
[48] Cnaan A, Laird NM, Slasor P. Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. Stat Med. 1997;16(20):2349–2380. doi:10.1002/(sici)1097-0258(19971030)16:20<2349::aid-sim667>3.0.co;2-e.
[49] Bose G, Healy BC, Lokhande HA, et al. Early predictors of clinical and MRI outcomes using least absolute shrinkage and selection operator (LASSO) in

multiple sclerosis. Ann Neurol 2022;92(1):87–96. https://doi.org/10.1002/ana.26370.
[50] Schober P, Vetter TR. Logistic regression in medical research. Anesth Analg 2021;132(2):365–6. https://doi.org/10.1213/ANE.0000000000005247.
[51] Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to machine learning, neural networks, and deep learning. Transl Vis Sci Technol. 2020;9(2):14. doi:10. 1167/tvst.9.2.14.
[52] Fogarty C, Peter T, Karatzas N, Dave S, Belinsky N, Pant Pai N. Global health facility-based interventions to achieve UNAIDS 90-90-90: a systematic review and narrative analysis. AIDS Behav 2022;26(5):1489–503. https://doi.org/10.1007/s10461-021-03503-6.
[53] Weber S, Heitmann KU. Interoperability in healthcare: also prescribed for digital health applications (DiGA). Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2021;64(10):1262–8. https://doi.org/10.1007/s00103-021-03414-w.
[54] Saripalle R, Runyan C, Russell M. Using HL7 FHIR to achieve interoperability in patient health record. J Biomed Inform 2019;94:103188. https://doi.org/10.1016/j.jbi.2019.103188.
[55] Du M, Yuan J, Jing W, Liu M, Liu J. The effect of international travel arrivals on the new HIV infections in 15–49 years aged group among 109 countries or territories from 2000 to 2018. Front Public Health. 2022;10:833551. doi:https://doi.org/10.3389/fpubh.2022.833551.
[56] Yang F, Wang K, Sun L, Zhai M, Song J, Wang H. A hybrid sampling algorithm combining synthetic minority over-sampling technique and edited nearest neighbor for missed abortion diagnosis. BMC Med Inform Decis Mak 2022;22(1):344. https://doi.org/10.1186/s12911-022-02075-2.
[57] Liu Z, Alavi A, Li M, Zhang X. Self-supervised contrastive learning for medical time series: a systematic review. Sensors (Basel) 2023;23(9):4221. https://doi.org/10.3390/s23094221.
[58] Shi W, Huang G, Song S, Wang Z, Lin T, Wu C. Self-supervised discovering of interpretable features for reinforcement learning. IEEE Trans Pattern Anal Mach Intell 2022;44(5):2712–24. https://doi.org/10.1109/TPAMI.2020.3037898.
[59] Wang KJ, Makond B, Chen KH, Wang KM. A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients. Appl Soft Comput 2014;20:15–24. https://doi.org/10.1016/j.asoc.2013.09.014.
[60] Abdalkareem ZA, Al-Betar MA, Amir A, Ehkan P, Hammouri AI, Salman OH. Discrete flower pollination algorithm for patient admission scheduling problem. Comput Biol Med 2022;141:105007. https://doi.org/10.1016/j.compbiomed.2021.105007.
[61] Stahlschmidt SR, Ulfenborg B, Synnergren J. Multimodal deep learning for biomedical data fusion: a review. Brief Bioinform. 2022;23(2):bbab569. doi: https://doi.org/10.1093/bib/bbab569.