frontiers | Frontiers in Public Health

TYPE Original Research
PUBLISHED 12 March 2025
DOI 10.3389/fpubh.2025.1544351

Check for updates

OPEN ACCESS

EDITED BY
Surjeet Dalal,
Amity University Gurgaon, India

REVIEWED BY
Yogarajan Gunasekaran,
Mepco Schlenk Engineering College, India
Karan Gupta,
Netaji Subhas University of Technology, India
Muzammal Hussain,
Government College University, Faisalabad, Pakistan

*CORRESPONDENCE
Erxiding Maimaiti
✉ cba624799002@126.com
Maimaitiaili Wubuli
✉ mamatali@126.com

RECEIVED 12 December 2024
ACCEPTED 17 February 2025
PUBLISHED 12 March 2025

CITATION
Chen Y, Pan K, Lu X, Maimaiti E and Wubuli M (2025) Machine learning-based prediction of mortality risk in AIDS patients with comorbid common AIDS-related diseases or symptoms. *Front. Public Health* 13:1544351. doi: 10.3389/fpubh.2025.1544351

Machine learning-based prediction of mortality risk in AIDS patients with comorbid common AIDS-related diseases or symptoms

Yiwei Chen[1], Kejun Pan[2], Xiaobo Lu[2], Erxiding Maimaiti[1]* and Maimaitiaili Wubuli[2]*

[1]Department of Epidemiology and Health Statistics, School of Public Health, Xinjiang Medical University, Urumqi, Xinjiang, China, [2]Department of Infectious Diseases, The First Affiliated Hospital of Xinjiang Medical University, Urumqi, Xinjiang, China

Objective: Early assessment and intervention of Acquired Immune Deficiency Syndrome (AIDS) patients at high risk of mortality is critical. This study aims to develop an optimally performing mortality risk prediction model for AIDS patients with comorbid AIDS-related diseases or symptoms to facilitate early intervention.

Methods: The study included 478 first-time hospital-admitted AIDS patients with related diseases or symptoms. Eight predictors were screened using lasso regression, followed by building eight models and using SHAP values (Shapley's additive explanatory values) to identify key features in the best models. The accuracy and discriminatory power of model predictions were assessed using variable importance plots, receiver operating characteristic curves, calibration curves, and confusion matrices. Clinical benefits were evaluated through decision-curve analyses, and validation was performed with an external set of 48 patients.

Results: Lasso regression identified eight predictors, including hemoglobin, infection pathway, Sulfamethoxazole-Trimethoprim, expectoration, headache, persistent diarrhea, Pneumocystis jirovecii pneumonia, and bacterial pneumonia. The optimal model, XGBoost, yielded an Area Under Curve (AUC) of 0.832, a sensitivity of 0.703, and a specificity of 0.799 in the training set. In the test set, the AUC was 0.729, the sensitivity was 0.717, and the specificity was 0.636. In the external validation set, the AUC was 0.873, the sensitivity was 0.852, and the specificity was 0.762. Furthermore, the calibration curves showed a high degree of fit, and the DCA curves demonstrated the overall high clinical utility of the model.

Conclusion: In this study, an XGBoost-based mortality risk prediction model is proposed, which can effectively predict the mortality risk of patients with co-morbid AIDS-related diseases or symptomatic AIDS, providing a new reference for clinical decision-making.

KEYWORDS

machine learning, XGBoost, AIDS, HIV, prediction model

Introduction

The risk of death from AIDS has been a significant global concern for decades. Among these, AIDS-related diseases and symptoms are important factors that affect the prognosis of patients. The National Institutes of Health (NIH) invested nearly $69 billion in AIDS research between 1982 and 2018 to understand, treat, and prevent HIV infection (1, 2). Studies have demonstrated that HIV infection progressively compromises the patient's immune system, with a gradual depletion of CD4+ T-lymphocytes (3–5). This ultimately leads to various opportunistic infections (OIs) and tumors. Since 1980, OIs have accounted for a large proportion of deaths among HIV-infected patients, particularly in Asia and sub-Saharan Africa and West Africa (6–9).

OIs encompass a variety of bacteria, viruses, fungi, and parasites, some of which are exceedingly rare in immunocompetent populations (10). OIs pose a significant health risk to patients with AIDS, particularly when the CD4+ T-cell count is below 200 cells/μL. The presence of various OIs significantly increases mortality rates in patients.

Tumors in people with AIDS include both AIDS-defining and non-AIDS-defining tumors, with generally low survival rates for these patients. AIDS-defining tumors mainly include Kaposi's sarcoma and non-Hodgkin's lymphoma. Non-AIDS-defining tumors include lung cancer, hepatocellular carcinoma, and perianal tumors, among others. AIDS-defining tumors account for 15–19% of deaths in HIV-infected patients, with most having an early onset and a more aggressive course than non-AIDS-defining tumors (11). Confirmation of AIDS-associated cancers primarily relies on histopathological biopsy (12).

With the widespread use of Combination Antiretroviral Treatment (cART), AIDS is gradually becoming a chronic disease with a limited impact on life expectancy, but this increased survival has also led to a surge in comorbidities (13–16). Early cART effectively prevents OIs and tumors, reducing the risk of developing these conditions. Consequently, the proportion of OIs and tumors in treated patients has greatly reduced, although it remains high (17–20). OIs and tumors remain the leading cause of death among people living with HIV, with significantly higher mortality rates, particularly in low- and middle-income countries (21).

Although some studies have explored factors affecting the prognosis of patients with AIDS, including hemoglobin, viral load, and CD4+ T-cell counts, they have been Cox regression and Logistic regression, which are traditional regression methods that, while providing a basic predictive framework, are usually unable to deal effectively with high-dimensional data or complex variable interactions, and especially exhibit significant limitations when nonlinear effects are involved, may not be able to effectively capture the complex relationships between variables, resulting in inadequate predictive performance (22–27). In addition, machine learning techniques have been gradually introduced into the medical field in recent years, and their advantages in large-scale data analysis and complex model construction have been widely recognized. Recently, machine learning algorithms have become increasingly popular in healthcare, with clinically based machine learning models being used for prognostic predictions in various diseases, such as diabetes mellitus and rectal cancer (28–36). The application of death prediction in infectious diseases is also becoming a growing trend, particularly for predicting patient mortality risk, as evidenced by the COVID-19 outbreak (36, 37). However, predictive studies of the risk of death for first-time HIV admissions are scarce, and no studies have used machine learning methods to predict the risk of death for first-time admissions with co-morbid HIV-related illnesses or symptoms. To fill this research gap, this study develops a machine learning-based optimal mortality risk prediction model based on a comprehensive dataset covering demographic information, clinical manifestations, and laboratory metrics, and combines it with the SHAP tool for model interpretability analysis to help clinicians identify high-risk patients and adjust their treatment plans. In addition, we validate the performance of the model with an external dataset to demonstrate its stability and generalization ability in real-world applications. XGBoost performs better in dealing with complex data structures and nonlinear relationships than other machine learning methods by integrating multiple decision trees, has higher prediction accuracy, more efficient big data processing capability, supports parallelized training, faster training speed, less resource consumption, and has strong generalization ability, therefore, this study considers constructing a prediction model based on XGBoost.

In summary, the main goal of this study is to construct a set of scientifically valid mortality risk prediction models to support clinicians in early diagnosis and individualized interventions for first-time HIV admissions to improve patient prognosis.

Methods

Research design

The data used to construct and test the model in this study were obtained from 478 patients with AIDS who attended the Infectious Diseases-Hepatology Center of the First Affiliated Hospital of Xinjiang Medical University between October 2000 and January 2021, presenting AIDS-related diseases or symptoms at the beginning of their admission to the hospital. We collected demographic data (e.g., sex and age), AIDS-related disease information (e.g., thrush and cryptococcosis), clinical manifestations (e.g., persistent fever and persistent diarrhea), and laboratory test data (e.g., hemoglobin and albumin) from patients at the beginning of the admission period, for a total of 55 variables. Patients were followed up regularly according to their condition after receiving cART until March 7, 2024, with death as the outcome indicator, resulted in a total of 248 survivors and 230 deaths. Inclusion criteria were as follows: (1) patients had a positive HIV antibody confirmatory test; (2) patients' diagnoses of relevant opportunistic infections and tumors were based on clinical manifestations, ancillary investigations, and medical records, confirmed by discharge diagnosis; (3) patients had completed relevant investigations before receiving cART; and (4) patients had good adherence to the study and received timely follow-up visits. Patient treatment adherence was measured by patients' medication use records and regular follow-up data, which were conducted every 3 months. Exclusion criteria were as follows: (1) seriously missing case information; (2) patients with poor adherence.

Statistical methods

Continuous variables in this study were expressed as mean ± SD or median (interquartile range), and categorical variables as frequency (percentage). These variables were then compared between survivors and deceased using Student's t-tests, Mann–Whitney U-tests, chi-square tests, and Fisher's exact tests. All analyses other than comparisons between multiple models were conducted in R 4.3.1, with the CBCgrps package (version 2.8.2) used for the analysis of differences (38, 39). Comparisons between multiple models were performed in Extreme Smart Analysis. All tests were two-sided with a significance level of $\alpha = 0.05$.

Data preprocessing

Excluded variables with more than 20% of the original data missing. The remaining data were filled using the Random Forest method. Then, the training and test sets were split in a ratio of 6:4. The Random Forest approach can effectively handle missing values through multiple interpolation and can maintain the structural integrity of the dataset. In the process of constructing each tree, Random Forest considers different feature subsets to reduce the impact of outliers, while missing values can be handled by the combined results of multiple trees.

Selection of predictors

Using whether death was the dependent variable, first, Receiver Operating Characteristic (ROC) curves and Area Under Curve (AUC) values for all covariates in the complete dataset were generated to gain preliminary insights into the variables. Then, predictors were identified from the variables in the training set using Lasso regression and a min-max normalization was applied to the quantitative data. According to the 10 Events Per Variable (10EPV) rule, the sample size of deceased patients in the training set was ensured to meet the criterion of 10 times the number of predictors.

Modeling

In this study, seven machine learning algorithms (XGBoost, LightGBM, AdaBoost, MLP, SVM, GNB, and KNN) and one traditional regression algorithm (Logistic Regression) were used to initially construct patient mortality risk prediction models. The optimal model was selected through 10-fold Cross-Validation (CV), and then the optimal model was subjected to hyperparameter optimization (including max_depth, eta, gamma, colsample_bytree, min_child_weight, and subsample parameters) to construct the final model. The model was interpreted using the SHAP tool. Finally, a variable importance plot, ranking graph, and variable dependency plot were generated to show the relative importance of each feature in the model.

Evaluation of the model

The AUC is calculated from the ROC curve. The ROC curve is frequently used to assess the discriminative capacity of a predictive model, i.e., its ability to discriminate between different categories.

In this study, calibration curves were used to assess model fit, in which the Brier score was used as an evaluation metric; the lower the Brier score, the better the model fit.

The accuracy and discriminative power of model predictions were evaluated using a confusion matrix. Specific indicators include accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1 score.

The study used Decision Curve Analysis (DCA) to evaluate the clinical utility of the model. The DCA curve plots the threshold probability on the horizontal axis and the net benefit on the vertical axis. The closer the curve is to the upper right corner, the greater the utility of the predictive model.

External validation

Forty-eight AIDS patients presenting with AIDS-related diseases or symptoms upon their first admission to the Shayibak District Branch of Urumqi Friendship Hospital (21 survivors and 27 deceased) were included in the external validation cohort. The external validation cohort came from different hospitals in the same area and had similar disease characteristics as the training set. We incorporated the data from the external validation set into the model constructed from the training set, and assessed the performance, goodness of fit, and clinical benefit of the model by plotting ROC curves, calibration curves, and DCA curves; a higher AUC value and a high degree of fit, and a wide interval of the DCA curves illustrated a high degree of generalizability of the model.

Ethics statement

The study protocol was approved by the Ethical Review Committee of the First Affiliated Hospital of Xinjiang Medical University (Ethical approval number: K202409-31). All experiments were conducted in accordance with relevant designated guidelines and regulations. Due to the retrospective nature of the study, the ethical review committee of the First Affiliated Hospital of Xinjiang Medical University waived the need of obtaining informed consent.

Results

Research flowchart

The flowchart of the study is presented in Figure 1. The flowchart is divided into four sections: research steps, methods, research content, and research design.

Patient characteristics

A total of 478 AIDS patients (248 survivors and 230 deceased) were included in this study up to 7 March 2024, who presented with AIDS-related diseases or symptoms at the beginning of their admission to the hospital between October 2000 and January 2021 at the Infectious Diseases-Hepatology Center of the First Affiliated Hospital of Xinjiang Medical University. The results of the analysis of

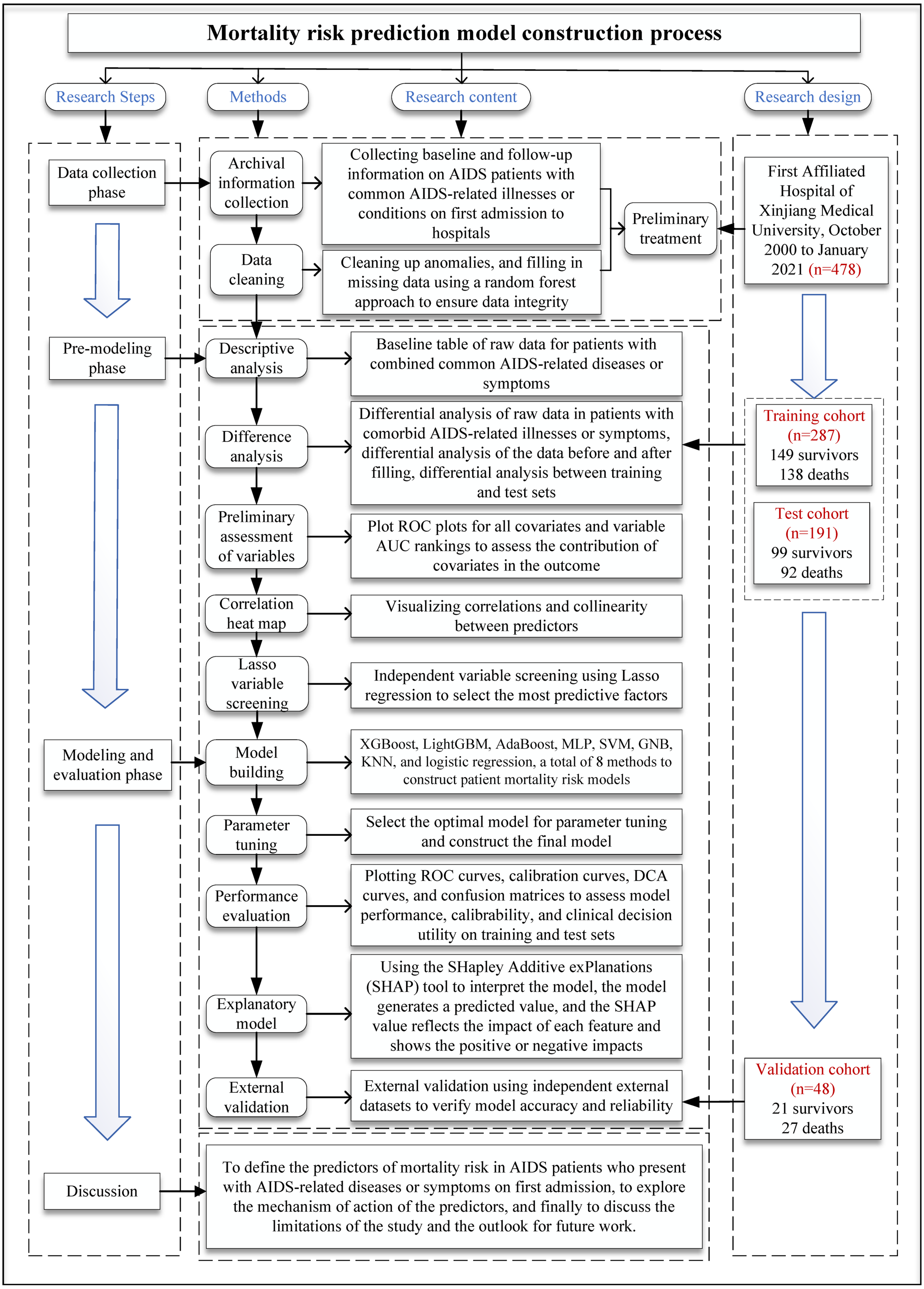

Mortality risk prediction model construction process
Research Steps
Methods
Research content
Research design
Data collection phase
Archival information collection
Collecting baseline and follow-up information on AIDS patients with common AIDS-related illnesses or conditions on first admission to hospitals
Data cleaning
Cleaning up anomalies, and filling in missing data using a random forest approach to ensure data integrity
Preliminary treatment
First Affiliated Hospital of Xinjiang Medical University, October 2000 to January 2021 (n=478)
Pre-modeling phase
Descriptive analysis
Baseline table of raw data for patients with combined common AIDS-related diseases or symptoms
Difference analysis
Differential analysis of raw data in patients with comorbid AIDS-related illnesses or symptoms, differential analysis of the data before and after filling, differential analysis between training and test sets
Training cohort (n=287) 149 survivors 138 deaths
Test cohort (n=191) 99 survivors 92 deaths
Preliminary assessment of variables
Plot ROC plots for all covariates and variable AUC rankings to assess the contribution of covariates in the outcome
Correlation heat map
Visualizing correlations and collinearity between predictors
Lasso variable screening
Independent variable screening using Lasso regression to select the most predictive factors
Modeling and evaluation phase
Model building
XGBoost, LightGBM, AdaBoost, MLP, SVM, GNB, KNN, and logistic regression, a total of 8 methods to construct patient mortality risk models
Parameter tuning
Select the optimal model for parameter tuning and construct the final model
Performance evaluation
Plotting ROC curves, calibration curves, DCA curves, and confusion matrices to assess model performance, calibrability, and clinical decision utility on training and test sets
Explanatory model
Using the SHapley Additive exPlanations (SHAP) tool to interpret the model, the model generates a predicted value, and the SHAP value reflects the impact of each feature and shows the positive or negative impacts
External validation
External validation using independent external datasets to verify model accuracy and reliability
Validation cohort (n=48) 21 survivors 27 deaths
Discussion
To define the predictors of mortality risk in AIDS patients who present with AIDS-related diseases or symptoms on first admission, to explore the mechanism of action of the predictors, and finally to discuss the limitations of the study and the outlook for future work.

FIGURE 1
Flowchart of the mortality risk prediction model development.

variance between the two groups of raw data (Table 1) indicate that the variables of patients' marital status, infection pathway, treatment time group, OHL, esophageal candidiasis, PJP, CMV, bacterial pneumonia, persistent diarrhea, nausea, headache, WHO stage, SMZ-TMP, and treatment plan exhibited a differential distribution between the two groups. Furthermore, deceased patients exhibited

TABLE 1 Baseline characterization of raw data and analysis of differences.

Variables	Total (*n* = 478)	Survival (*n* = 248)	Deceased (*n* = 230)	*p*
Marital status, n (%)				0.021
Single	58 (12.2)	35 (14.1)	23 (10.1)	
Married or cohabiting	338 (71.0)	182 (73.4)	156 (68.4)	
Divorced or widowed	80 (16.8)	31 (12.5)	49 (21.5)	
Treatment time group, n (%)				0.032
0–30 days	238 (49.8)	115 (46.4)	123 (53.5)	
31–90 days	108 (22.6)	68 (27.4)	40 (17.4)	
91–365 days	69 (14.4)	30 (12.1)	39 (17.0)	
>365 days	63 (13.2)	35 (14.1)	28 (12.2)	
Infection pathway, n (%)				<0.001
Blood-borne (transfusion + apheresis)	22 (4.6)	8 (3.2)	14 (6.1)	
Intravenous drug addiction	149 (31.2)	51 (20.6)	98 (42.6)	
homosexual transmission	11 (2.3)	8 (3.2)	3 (1.3)	
Heterosexual transmission	264 (55.2)	173 (69.8)	91 (39.6)	
Other (mother-to-child transmission + unknown route)	32 (6.7)	8 (3.2)	24 (10.4)	
OHL, n (%)				0.035
No	465 (97.3)	237 (95.6)	228 (99.1)	
Yes	13 (2.7)	11 (4.4)	2 (0.9)	
PJP, n (%)				<0.001
No	430 (90.0)	206 (83.1)	224 (97.4)	
Yes	48 (10.0)	42 (16.9)	6 (2.6)	
CMV, n (%)				0.031
No	472 (98.7)	242 (97.6)	230 (100.0)	
Yes	6 (1.3)	6 (2.4)	0 (0)	
Bacterial pneumonia, n (%)				<0.001
No	440 (92.1)	244 (98.4)	196 (85.2)	
Yes	38 (7.9)	4 (1.6)	34 (14.8)	
Persistent diarrhea, n (%)				0.017
No	418 (87.4)	226 (91.1)	192 (83.5)	
Yes	60 (12.6)	22 (8.9)	38 (16.5)	
Nausea, n (%)				0.048
No	432 (90.4)	231 (93.1)	201 (87.4)	
Yes	46 (9.6)	17 (6.9)	29 (12.6)	
Projectile vomiting, n (%)				0.053
No	474 (99.2)	248 (100)	226 (98.3)	
Yes	4 (0.8)	0 (0)	4 (1.7)	
Headache, n (%)				0.035
No	440 (92.1)	235 (94.8)	205 (89.1)	
Yes	38 (7.9)	13 (5.2)	25 (10.9)	

(Continued)

TABLE 1 (Continued)

Variables	Total (*n* = 478)	Survival (*n* = 248)	Deceased (*n* = 230)	*p*
WHO, n (%)				<0.001
Stage 1	22 (4.6)	9 (3.6)	13 (5.7)	
Stage 2	399 (83.8)	222 (89.5)	177 (77.6)	
Stage 3	31 (6.5)	5 (2.0)	26 (11.4)	
Stage 4	24 (5.0)	12 (4.8)	12 (5.3)	
SMZ-TMP, n (%)				<0.001
No	170 (35.8)	60 (24.2)	110 (48.5)	
Yes	305 (64.2)	188 (75.8)	117 (51.5)	
Plan, n (%)				0.006
AZT + 3TC + DDI	1 (0.2)	0 (0)	1 (0.4)	
AZT + 3TC + EFV	126 (26.4)	77 (31.0)	49 (21.3)	
AZT + 3TC + LVP	3 (0.6)	0 (0)	3 (1.3)	
AZT + 3TC + NVP	168 (35.1)	73 (29.4)	95 (41.3)	
D4T + 3TC + EFV	41 (8.6)	22 (8.9)	19 (8.3)	
D4T + 3TC + NVP	30 (6.3)	14 (5.6)	16 (7.0)	
TDF + 3TC + EFV	89 (18.6)	55 (22.2)	34 (14.8)	
TDF + 3TC + LVP	12 (2.5)	5 (2.0)	7 (3.0)	
TDF + 3TC + NVP	3 (0.6)	1 (0.4)	2 (0.9)	
3TC + DTG	1 (0.2)	0 (0)	1 (0.4)	
BIC/FTC/TAF	1 (0.2)	1 (0.4)	0 (0)	
EVG/c/FTC/TAF	3 (0.6)	0 (0)	3 (1.3)	
HDL, mmol/L	0.8 ± 0.3	0.9 ± 0.3	0.7 ± 0.3	0.032
AST, U/L	30.2 (22.0, 46.8)	29.0 (21.0, 40.0)	33.6 (23.0, 52.0)	0.007
ALT, U/L	27.0 (19.0, 44.1)	24.1 (17.4, 40.8)	29.0 (21.6, 49.4)	0.002
GGT, U/L	45.3 (26.0, 87.0)	41.0 (24.0, 81.0)	58.7 (41.8, 98.4)	0.011

Treatment time group, Time from discovery of HIV positivity to initiation of treatment; OHL, Oral Hairy Leukoplakia; PJP, Pneumocystis Jirovecii Pneumonia; CMV, Cytomegalovirus; SMZ-TMP, Sulfamethoxazole-Trimethoprim; HDL, High Density Lipoprotein; AST, Aspartate Aminotransferase; ALT, Alanine Aminotransferase; GGT, γ-Glutamyl transpeptidase.

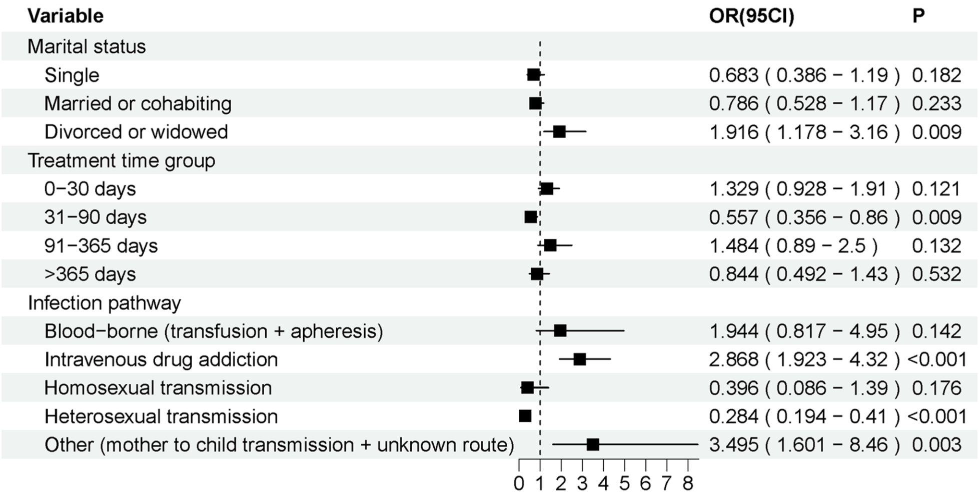
lower levels of HDL and higher levels of AST, ALT, and GGT compared to surviving patients. Subgroup analyses (Figure 2) were performed to explore variations in mortality risk across different patient characteristics. Multivariate logistic regression models were applied to estimate odds ratios (ORs) and 95% confidence intervals (CIs) for each subgroup. Variables included marital status (single, married or cohabiting, divorced or widowed), Treatment time group (0–30 days, 31–90 days, 91–365 days, >365 days), and infection pathways (blood-borne, intravenous drug addiction, homosexual transmission, heterosexual transmission, other).

The study excluded covariates with more than 20% missing data and filled in data with less than 20% missing using the random forest approach. The analysis of differences in the data before and after filling (Supplementary Material Table 1) showed no statistically significant differences in any variable. Subsequently, the filled patient data were randomly divided into a training set and a test set in a ratio of 6:4. An analysis of differences was performed between the two datasets (Supplementary Material Table 2). The differences in each variable were not statistically significant, and the data were balanced and comparable.

Predictor selection

First, an exploratory analysis of the data was conducted to plot ROC curves for all independent variables in the filled dataset (Figure 3a) to initially determine the relationships between all independent variables and the outcome variable. Then, the contributions of all independent variables were ranked (Figure 3b).

Finally, lasso regression was performed on the training set to obtain the Lambda chart (Figure 4a) and the cross-validation diagram (Figure 4b). Predictors were screened from 59 independent variables, and non-zero coefficient positive and negative bar plots are shown in Figure 4c. Under the λ-1se dashed line, the model fit was good, and the number of predictors was appropriate. Ultimately, eight predictors were identified, including one continuous variable (HB) and seven categorical variables (bacterial pneumonia, persistent diarrhea, headache, expectoration, infection pathway, SMZ-TMP, PJP), and a min-max standardized transformation was performed for the continuous variable HB. Four variables were positively related to mortality (high-risk variables: bacterial pneumonia, persistent diarrhea, headache, expectoration), and four variables were negatively



Variable	OR(95CI)	P
Marital status		
Single	0.683 (0.386 − 1.19)	0.182
Married or cohabiting	0.786 (0.528 − 1.17)	0.233
Divorced or widowed	1.916 (1.178 − 3.16)	0.009
Treatment time group		
0−30 days	1.329 (0.928 − 1.91)	0.121
31−90 days	0.557 (0.356 − 0.86)	0.009
91−365 days	1.484 (0.89 − 2.5)	0.132
>365 days	0.844 (0.492 − 1.43)	0.532
Infection pathway		
Blood−borne (transfusion + apheresis)	1.944 (0.817 − 4.95)	0.142
Intravenous drug addiction	2.868 (1.923 − 4.32)	<0.001
Homosexual transmission	0.396 (0.086 − 1.39)	0.176
Heterosexual transmission	0.284 (0.194 − 0.41)	<0.001
Other (mother to child transmission + unknown route)	3.495 (1.601 − 8.46)	0.003

FIGURE 2
Subgroup analysis forest map. Subgroup analyses showed a significantly higher risk of death in patients who were divorced or widowed ($p < 0.05$); in patients with a time from detection of HIV positivity to initiation of treatment of 31–90 days ($p < 0.05$); and in patients whose infection pathway was intravenous drug addiction or heterosexual transmission ($p < 0.05$).

related to mortality (low-risk variables: infection pathway, SMZ-TMP, HB, PJP).

Model building and evaluation

This study selected seven machine learning algorithms (XGBoost, LightGBM, AdaBoost, MLP, SVM, GNB, KNN) and one traditional regression method (Logistic Regression) to build a patient mortality risk prediction model. The ROC curve, calibration curve, and DCA curve of the training set and ten-fold cross-validation were drawn (Figures 5a–d). According to Figure 5 and the cross-validation model parameter results (Table 2), it was found that among the eight models, the XGBoost model had the highest AUC, the lowest Brier score, the best-combined model parameter results, and the highest clinical benefit, so XGBoost was chosen for modeling. In order to optimize the performance of the XGBoost model, this study uses a grid search method for hyper-parameter tuning, with a nrounds of 200, max_depth of 3, eta of 0.01, gamma of 0.1, colsample_bytree of 0.7, min_child_weight of 3, and subsample of 0.7.

Subsequently, the ROC curves and AUC values (Figures 6a,b), calibration curves and Brier scores (Figures 6c,d), and DCA curves (Figures 6e,f) of the training and test sets were plotted. Figures 6a,b shows that the training set AUC = 0.832, the test set AUC = 0.729, and the performance of the model is good; Figures 6c,d shows that the training set Brier = 0.187 and the test set Brier = 0.214 have low values and good fit; In Figures 6e,f, the blue lines indicate the clinical intervention benefits, and most of the blue lines are above the two thresholds, showing that the clinical benefits are relatively high. The accuracy and discriminant power of the model predictions were visualized using the confusion matrix plots (Figures 7a,b), and the performance of the prediction model was calculated (Table 3), including accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1 score.

The model was interpreted using the SHAP tool. For each sample, the model generated a predicted value. The SHAP value reflects the impact of each feature, indicating positive or negative effects. We plotted a SHAP beeswarm plot (Figure 8a) to visualize variable importance. The horizontal axis is the SHAP value, which indicates the magnitude and direction of each feature's contribution to the prediction results; the color indicates the magnitude of the feature's value (red for high values and blue for low values), with red dots concentrating in the positive direction and blue dots concentrating in the negative direction, which indicates that the higher the value of the feature, the higher the positive contribution it will make to the deaths. The SHAP summary plot (Figure 8b) was plotted to rank the importance of the variables, with importance decreasing from top to bottom, and the longer the horizontal axis, the longer the feature, indicating that it has a greater overall impact on the model's predictions. A SHAP dependence plot (Figure 9a) was drawn to show the interaction between the variables, where the horizontal axis is the original value of a feature, the vertical axis is the corresponding SHAP value, and the color of the dots indicates the magnitude of the value of the other interacting feature. It can be seen from Figure 8b that infection pathway, HB, SMZ-TMP, and PJP are the top four important characteristics in terms of contribution, and they are negatively correlated with death. A higher value is associated with a lower risk of mortality; conversely, characteristics such as bacterial pneumonia, persistent diarrhea, headache, and expectoration were positively related to death, with higher values indicating a higher risk of death. Finally, we selected representative patients #10 and #45 to draw SHAP plots (Figure 9b). The prediction of death risk for patient #10 can be described by these characteristics: no PJP (SHAP value +0.0844), indicating a positive impact on the prediction of death. Expectoration symptoms (SHAP value +0.204) increased the likelihood of death. Headache symptoms (SHAP value +0.265) also increased the possibility of death. An HB level of 72 (SHAP value +0.386) was the most influential feature, significantly increasing

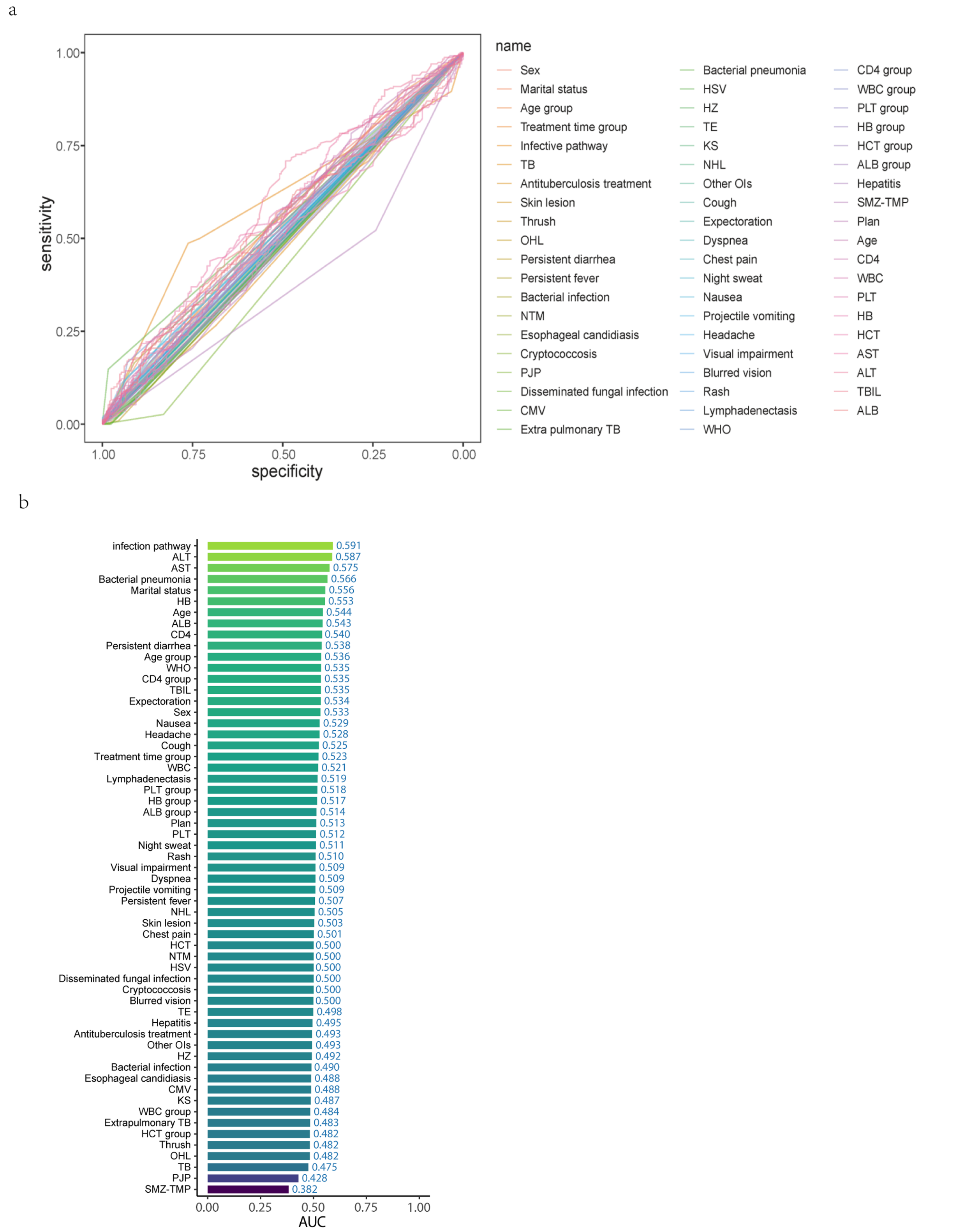


FIGURE 3
Initially determine the relationship between all independent variables and the ending variable and rank the contribution of the independent variables. **(a)** ROC curve for all independent variables. **(b)** AUC ranking of all independent variables. TB, Tuberculosis; OHL, Oral Hairy Leukoplakia; NTM, Nontuberculous Mycobacteria; PJP, Pneumocystis Jirovecii Pneumonia; CMV, Cytomegalovirus; HSV, Herpes Simplex Virus; HZ, Herpes Zoster; TE, Toxoplasmic Encephalitis; KS, Kaposi's Sarcoma; NHL, Non-Hodgkin lymphoma; OIs, Opportunistic Infections; SMZ-TMP, Sulfamethoxazole-Trimethoprim; WBC, White Blood Cell; PLT, Platelet; HB, Hemoglobin; HCT, Hematocrit; AST, Aspartate Aminotransferase; ALT, Alanine Aminotransferase; TBIL, Total Bilirubin; ALB, Albumin.

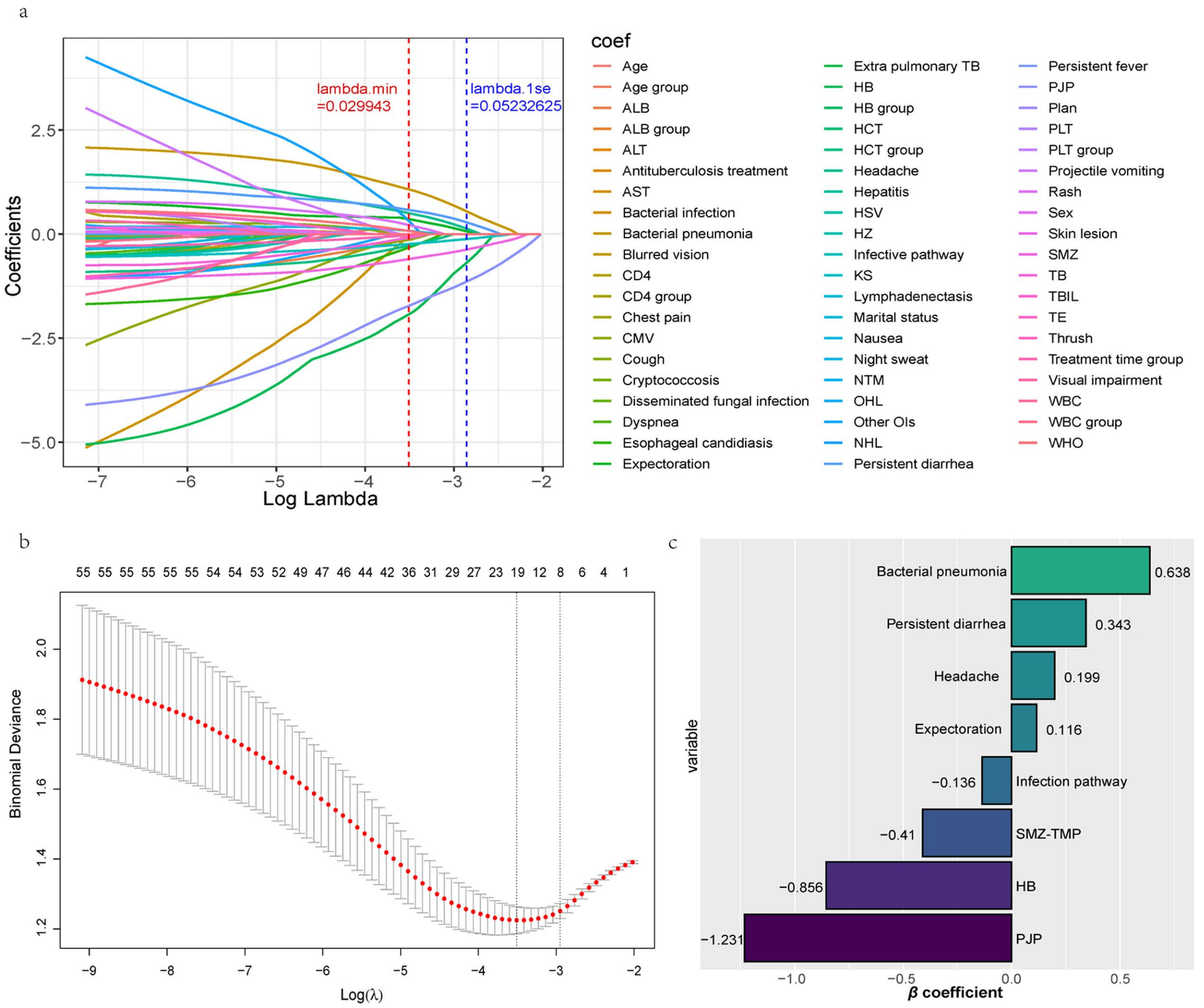


FIGURE 4
Screening of independent variables to obtain predictors. **(a)** Lambda plot. **(b)** Cross-validation plot. **(c)** Bar plots of positive and negative non-zero coefficients. ALB, Albumin; ALT, Alanine Aminotransferase; AST, Aspartate Aminotransferase; CMV, Cytomegalovirus; HB, Hemoglobin; HCT, Hematocrit; HSV, Herpes Simplex Virus; HZ, Herpes Zoster; KS, Kaposi's Sarcoma; NTM, Nontuberculous Mycobacteria; OHL, Oral Hairy Leukoplakia; OIs, Opportunistic Infections; NHL, Non-Hodgkin lymphoma; PJP, Pneumocystis Jirovecii Pneumonia; PLT, Platelet; SMZ-TMP, Sulfamethoxazole-Trimethoprim; TB, Tuberculosis; TBIL, Total Bilirubin; TE, Toxoplasmic Encephalitis; WBC, White Blood Cell. Variable selection is based on the LASSO regression method using 10-fold cross-validation to determine the optimal penalty parameter λ. Variables screened are those with non-zero coefficients in the model. The horizontal axis is the penalty parameter λ, the vertical axis is the coefficients of the variables, and the vertical dashed line indicates the optimal value of λ selected through cross-validation.

the possibility of death. Heterosexual transmission (SHAP value −0.19) as the infection pathway reduced the possibility of death. Using SMZ-TMP (SHAP value −0.129) also reduced the likelihood of death. The absence of persistent diarrhea symptoms and bacterial pneumonia reduced the probability of death, but due to their small contribution, specific SHAP values were not shown in the figure. The final prediction score f(x) was 0.494, while the model's baseline prediction or expectation E[f(x)] was −0.0751. This indicates that the combination of these characteristics resulted in a prediction leaning toward death compared to the baseline prediction. Red bars indicated features that increased the probability of predicted death, while blue bars indicated features that decreased it. The predictive description of the risk of death for patient #45 was similar to that described above, and the predictive results tended to be survival.

External validation

In this study, we developed and internally validated a mortality risk prediction model based on the XGBoost algorithm. To further verify the generalization capability and practical application value of the model, we used the dataset of an external hospital for external validation. Forty-eight AIDS patients presenting with AIDS-related diseases or symptoms upon their first admission to the Shayibak District Branch of Urumqi Friendship Hospital (21

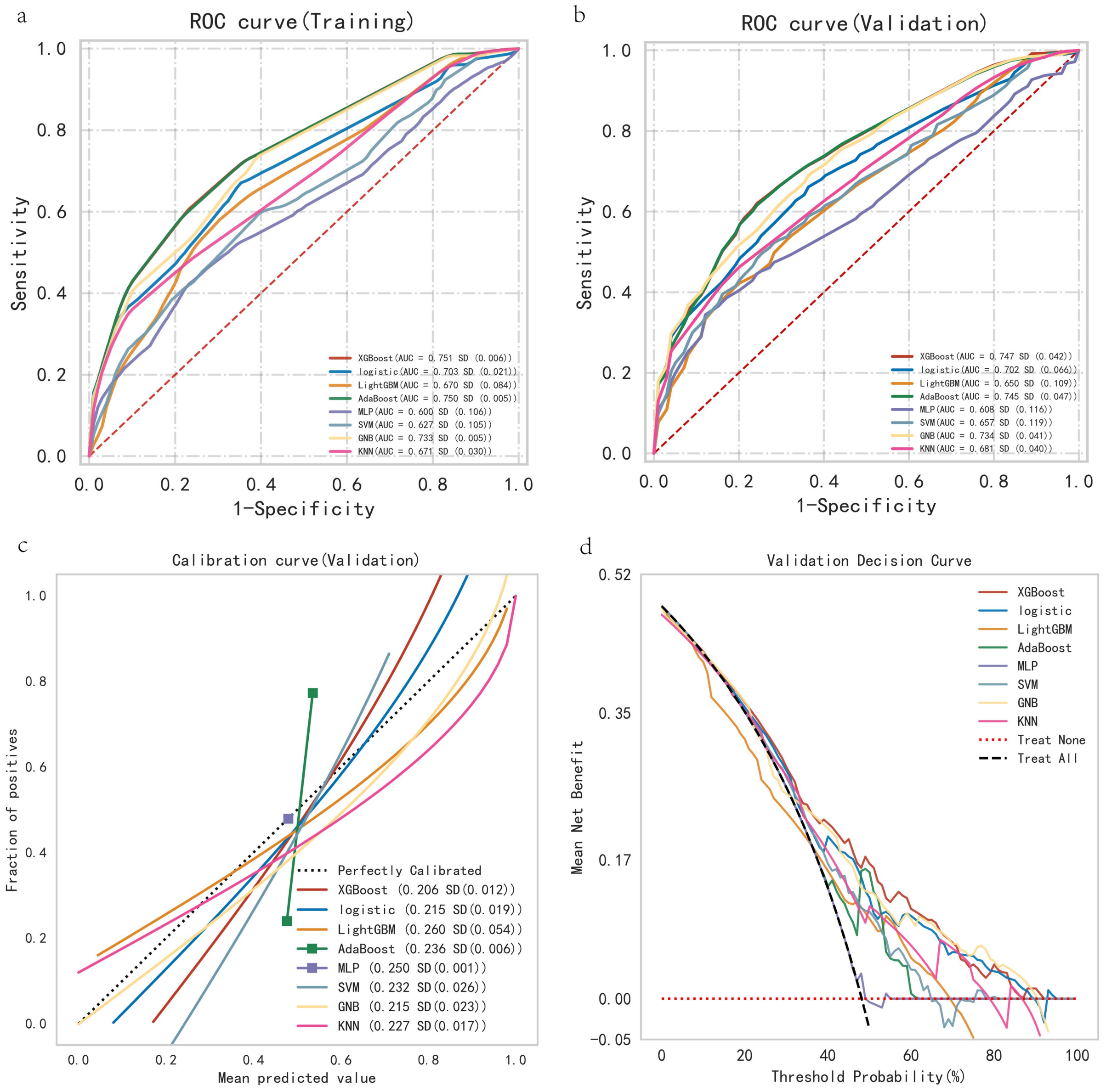


FIGURE 5
Performance evaluation of all models. **(a)** Training set ROC curve. **(b)** Ten-fold cross-validation ROC curve. **(c)** Ten-fold cross-validation calibration curve. **(d)** Ten-fold cross-validation DCA curve. Model XGBoost has the highest area under the ROC curve AUC and the most superior performance of the model. XGBoost has the lowest Brier score, which indicates that the closer the prediction probability is to the true outcome, the more accurate the model is. Model XGBoost has the highest clinical benefit for the DCA curve, with an interval of 20–90%.

TABLE 2 Ten-fold cross-validation model performance parameters.

Model	AUC (SD)	Accuracy (SD)	Sensitivity (SD)	Specificity (SD)	F1 Score (SD)
XGBoost	0.747 (0.042)	0.680 (0.049)	0.674 (0.102)	0.734 (0.091)	0.685 (0.067)
Logistic	0.702 (0.066)	0.636 (0.071)	0.574 (0.152)	0.786 (0.130)	0.599 (0.119)
LightGBM	0.650 (0.109)	0.610 (0.106)	0.648 (0.201)	0.656 (0.286)	0.621 (0.080)
AdaBoost	0.745 (0.047)	0.665 (0.056)	0.670 (0.107)	0.734 (0.091)	0.704 (0.071)
MLP	0.608 (0.116)	0.594 (0.069)	0.470 (0.269)	0.807 (0.233)	0.494 (0.200)
SVM	0.657 (0.119)	0.636 (0.081)	0.735 (0.141)	0.604 (0.242)	0.659 (0.063)
GNB	0.734 (0.041)	0.663 (0.047)	0.635 (0.144)	0.770 (0.129)	0.639 (0.098)
KNN	0.681 (0.040)	0.590 (0.072)	0.535 (0.218)	0.768 (0.229)	NaN (NaN)

The XGBoost model has the best overall performance.

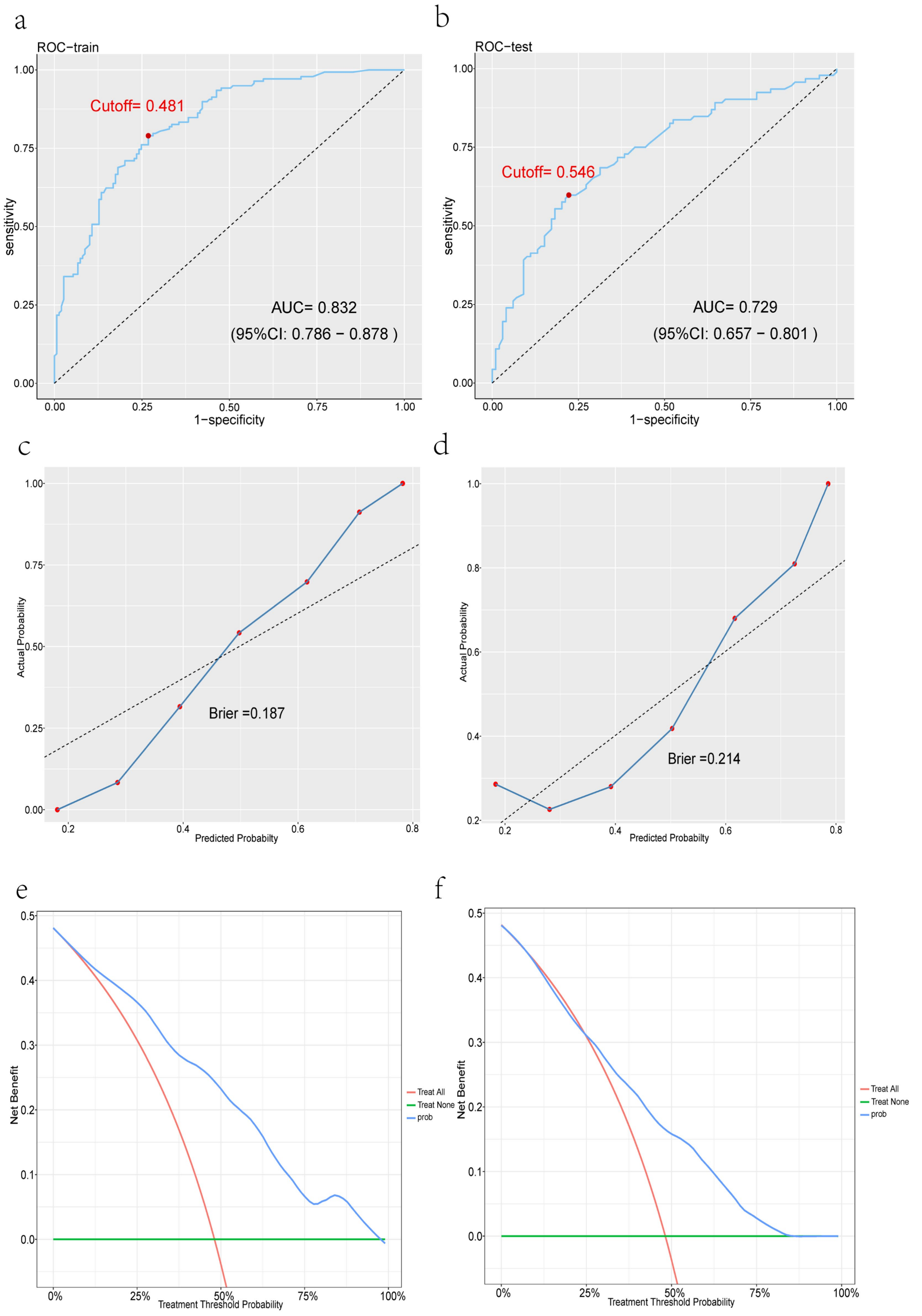


FIGURE 6
Performance evaluation of the XGBoost model. **(a)** XGBoost training set ROC curve. **(b)** XGBoost test set ROC curve. **(c)** XGBoost training set calibration curve. **(d)** XGBoost test set calibration curve. **(e)** XGBoost training set DCA curve. **(f)** XGBoost test set DCA curve.

survivors and 27 deceased) were included in the external validation cohort. In external validation, the model also demonstrated superior predictive performance compared to traditional prediction methods. Specifically, it included a confusion matrix plot (Figure 10a), an ROC curve (Figure 10b), a calibration curve (Figure 10c), and a DCA curve (Figure 10d). Among them, important indicators such as accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1 score were used to evaluate the performance of the model (Table 4).

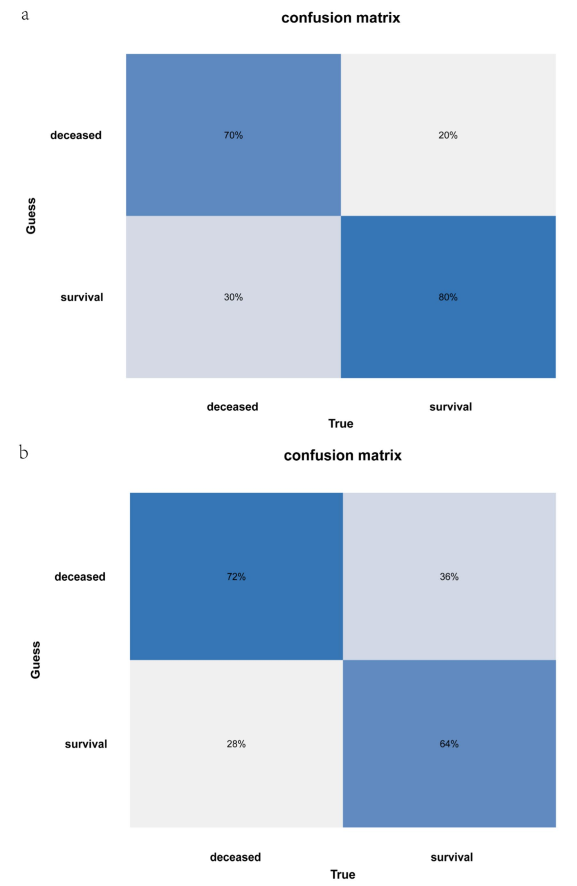


FIGURE 7
Accuracy and discriminative power of XGBoost model predictions. **(a)** Training set confusion matrix. **(b)** Test set confusion matrix. Confusion matrix plots show 70% and 72% sensitivity, reflecting the proportion of all individuals who were actually dead that the model model correctly predicted as dead. 80% and 64% specificity, referring to the proportion of all individuals who were actually alive that the model correctly predicted as alive.

Discussion

In this study, the XGBoost algorithm demonstrated better discrimination (AUC = 0.751) compared to the other seven models (Logistic Regression, LightGBM, AdaBoost, etc.). After adjusting the parameters, the overall efficacy of the XGBoost model in predicting mortality risk was relatively high, reflected in the model's ability to identify high-risk individuals. Specific indicators include prediction accuracy (Accuracy = 0.753), sensitivity (Sensitivity = 0.703), specificity (Specificity = 0.799), positive predictive value (Pos Pred Value = 0.744), negative predictive value (Neg Pred Value = 0.764), and F1 Score (F1 Score = 0.690). The AUC metric is used to measure the overall performance of the classification model, and the closer the value is to 1, the better the discriminative ability of the model. Meanwhile, since the goal of the study is to minimize the leakage of high-risk patients, Sensitivity and Negative Predictive Value are key indicators, and higher values indicate higher predictive reliability of the model. In addition, F1 Score serves as a balance between Positive Predictive Value/Precision and Sensitivity/Recall, with higher values representing a better measure of the model's classification ability.

In subsequent external validation, we found that the performance on the external dataset was also relatively good, indicating that the model has good generalizability. The results of this study, although based on a region-specific sample, are considered to have some generalizability, especially among other groups of AIDS patients in similar regions. Additionally, the DCA curve of the model indicates that, compared to the intervention of all patients and the non-intervention of patients, the predictive model of the internally verified test set has higher clinical intervention benefits in predicting the risk of death for patients within the range of 25–85%. The predictive model of the external validation set has higher clinical intervention benefits in predicting the risk of death for patients within the range of 15–100%.

During the modeling process, it was observed that a reduction in HB increases the risk of death in patients. The potential reasons for this are as follows: HIV itself has myelosuppressive manifestations that lead to a decrease in HB, so a decrease in HB is one of the common clinical manifestations of HIV infection (40). In HIV patients, anemia is a factor in accelerated disease progression and reduced quality of life, and prolonged anemia also increases the risk of death (22, 41–43). Some studies have indicated that intravenous drug users are at a higher risk of death due to the route of infection, and bacterial pneumonia is the third most common cause of AIDS-related death, which is consistent with our findings (19, 44, 45). The present study found that expectoration, headache, and persistent diarrhea were associated with an increased risk of death. It is postulated that this may be due to the presence of expectoration symptoms suggestive of lung diseases, such as pneumonia and PJP; headache symptoms indicative of central nervous system disease; and persistent diarrhea symptoms commonly observed in patients with advanced AIDS. A study found that a decrease in HB may lead to a worse prognosis in patients with comorbid PJP, and anemia should be managed aggressively in AIDS patients with comorbid PJP if their HB is less than 90 g/L (46). The present study found that the prophylactic use of SMZ-TMP reduces the risk of death in patients with AIDS. This is because SMZ-TMP is effective in reducing the incidence of PJP, which is an important mortality factor for AIDS patients (47–49). Additionally, PJP was found to be a unique predictor showing a negative correlation in mortality risk prediction. The presumed reason is that PJP is a serious opportunistic infection and patients usually receive standardized treatment immediately upon diagnosis. This early diagnosis and intervention may significantly improve patient prognosis (50).

Some limitations of our study need to be acknowledged. First, the sample size of the study is relatively small, particularly for external validation. Although the patient data originate from two large hospitals, the sample size and number of outcome events may still limit the accuracy of extrapolating the results to other regions and may not be fully representative of all patient groups (e.g., different ages, genders, regions, etc.). Moreover, the validation cohort is derived from data from hospitals in a specific region, which may be subject to regional bias, disease distribution, and treatment differences. Patient characteristics (e.g., disease spectrum, treatments, lifestyle habits, etc.) in different regions may affect the predictive effectiveness of the model. Therefore, more external test sets from different hospitals or

TABLE 3 Confusion matrix values.

	Accuracy	Sensitivity	Specificity	Pos pred value	Neg pred value	F1 score
Training set	0.753	0.703	0.799	0.744	0.764	0.732
Test set	0.675	0.717	0.636	0.708	0.647	0.680

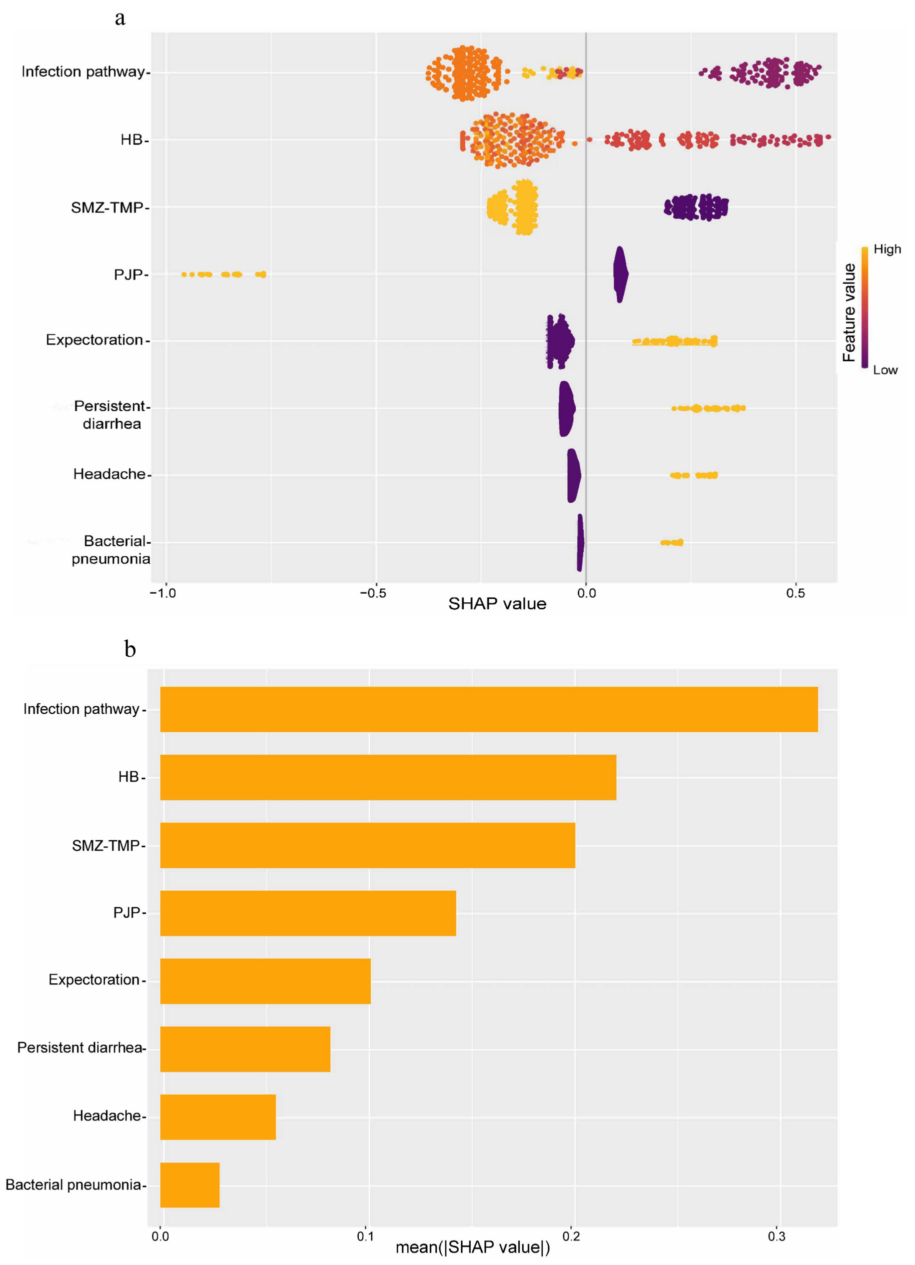


FIGURE 8
Importance and degree of contribution of model features. **(a)** SHAP beeswarm plot. **(b)** SHAP summary plot. HB, Hemoglobin; SMZ-TMP, Sulfamethoxazole-Trimethoprim; PJP, Pneumocystis Jirovecii Pneumonia. The horizontal axis of the SHAP beeswarm plot is the SHAP value, indicating the size and direction of the contribution of each feature to the prediction results, the color indicates the size of the feature value, the red dot feature value is large, the direction of its concentration, indicating the direction of the contribution to the prediction of the deaths. SHAP summary plot is to rank the importance of the variables, the importance of which decreases from the top to the bottom, and the longer the horizontal axis is, the longer the features are, indicating that they have a greater impact on the overall prediction of the model. The longer the horizontal axis and the longer the feature, indicating the greater its impact on the model's overall prediction.

regions are needed to enhance the model's robustness. As this study utilized retrospective data from hospital records, selection bias may have been introduced. Patients included in this study might represent those with more severe conditions or better treatment adherence, potentially limiting the generalizability of the findings to a broader population of AIDS patients. Second, the duration of follow-up from

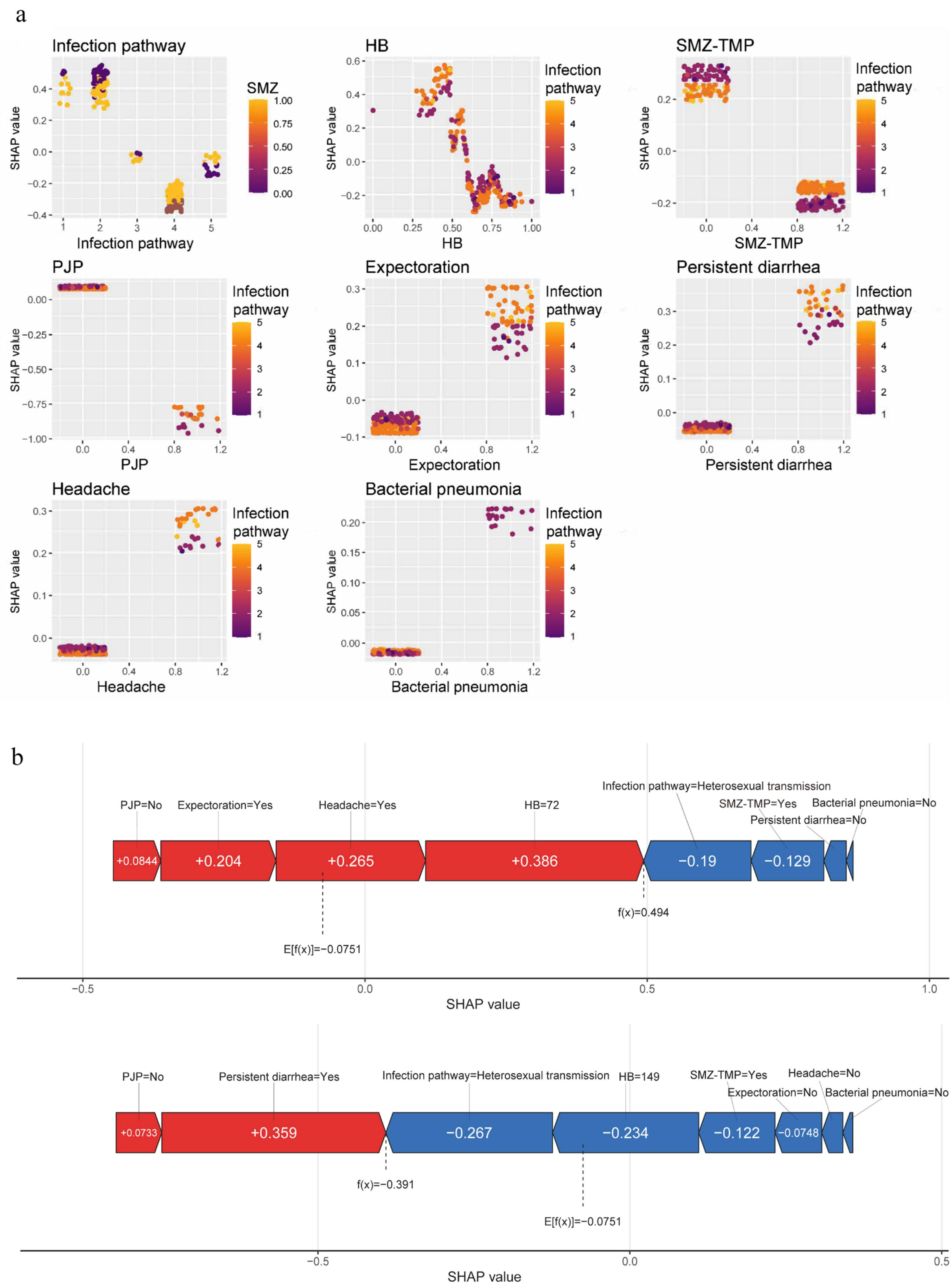


FIGURE 9
Interaction effects between features and model interpretation. **(a)** SHAP dependence plots. **(b)** SHAP force plot for Patient #10 & Patient #45. HB, Hemoglobin; SMZ-TMP, Sulfamethoxazole-Trimethoprim; PJP, Pneumocystis Jirovecii Pneumonia. The SHAP dependence plots show the interactions between variables, where the horizontal axis is the original value of a feature, the vertical axis is the corresponding SHAP value, and the color of the dot indicates the magnitude of the value of the other feature that is interacting. Red bars of the SHAP force plot indicate features that increase the

(Continued)

FIGURE 9 (Continued)
predicted probability of death, and blue bars indicate features that decrease the predicted probability of death. The final SHAP values for each predictor are summed to score f(x), which is compared to the model's baseline prediction or expectation, E[f(x)]. f(x) greater than E[f(x)] indicates that the combination of these predictors results in a prediction that favors death when compared to the baseline prediction, and conversely the prediction favors survival.

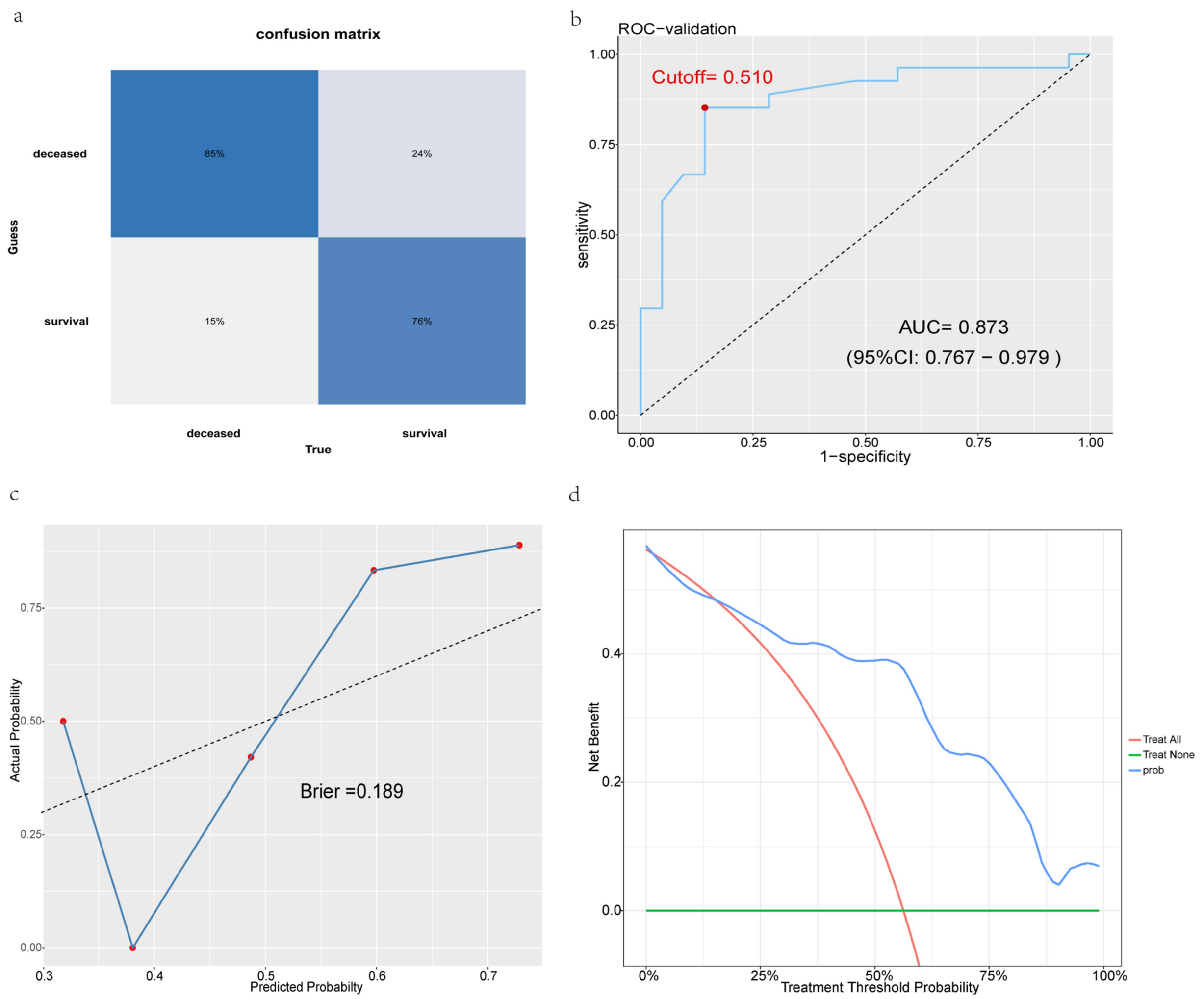


FIGURE 10
External validation shows excellent performance in models. **(a)** Validation set confusion matrix. **(b)** XGBoost validation set ROC curve. **(c)** XGBoost validation set calibration curve. **(d)** XGBoost validation set DCA curve.

TABLE 4 Validation confusion matrix values.

	Accuracy	Sensitivity	Specificity	Pos pred value	Neg pred value	F1 score
Validation set	0.813	0.852	0.762	0.800	0.821	0.836

admission to the study cut-off point varied for each patient, which cannot be completely avoided in clinical practice, and which may have influenced the study results, particularly among patients with shorter follow-up times who may not have reached the study endpoints. Fortunately, external validation demonstrated excellent performance, suggesting the model is suitable for predicting the risk of death in the study's patients, and related studies have confirmed that the performance of machine learning prediction models is not affected by the duration of follow-up (51). However, the predictive performance of the model still does not fully meet the expected level, which may be due to the limited sample size. Finally, different patients may require different treatment regimens, which may be adjusted during follow-up depending on the patient's specific situation. Due to ethical considerations and the observational nature of the study, we could not ensure the impact of patient treatment regimens on outcomes.

The availability of high-quality, consistent patient data remains a major barrier to implementing predictive models in resource-limited settings. Missing data and reliance on historical datasets may result in information bias, selection bias, and temporal bias. Missing data may result in incomplete or inaccurate information about the variables on which the model relies, thus affecting the veracity and reliability of the analyzed results. Removal of missing values may tend to retain patient populations with complete data, leading to an underestimation or overestimation of the broad applicability of study results. The time span of historical data may introduce changes in medical technology, standards of care, and patient characteristics, affecting the comparability of studies and the applicability of results. The lack of computational infrastructure may also hinder the integration of machine learning models into routine clinical practice. Most importantly, ethical issues, such as the potential for predictive models to exacerbate health inequalities, need to be carefully considered. The application of predictive modeling in clinical decision-making may involve issues of patient privacy and autonomy and needs to be implemented with due consideration of ethical implications.

In future research, we will first continue to increase the sample size based on the existing results, conduct joint studies with several hospitals to enhance the robustness and accuracy of the model, and continue to optimize the model. Second, the follow-up duration will be controlled to ensure that patients are followed for a consistent length of time. Third, we will integrate more data types, such as patient imaging data and genetic data, to improve model performance. Additionally, we will explore advanced techniques such as feature engineering and ensemble methods to further improve the prediction performance. Feature engineering includes constructing interactive features to combine variables to improve data representation, e.g., BMI = weight/height^2. Ensemble methods include combining XGBoost, LGBM, or neural networks to enhance model generalization. Fourth, we will explore the potential of applying the model to other related fields, such as predicting the risk of death in patients with other chronic diseases or using it in different epidemiological studies.

Conclusion

In conclusion, the following variables were identified as important predictors of the risk of death in patients: infection pathway, HB, SMZ-TMP, PJP, expectoration, persistent diarrhea, headache, and bacterial pneumonia. The findings assist clinicians in assessing disease severity in various ways. This study may serve as a reference for future clinical studies and potential applications.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: due to relevant national laws and the specificity of the population involved in the study, the datasets used to generate the graphs and analyses in this study are not publicly available to protect the privacy of people living with HIV, but can be obtained from the corresponding authors upon reasonable request. Requests to access these datasets should be directed to Yiwei Chen, chen1973492893@126.com.

Ethics statement

The studies involving humans were approved by Ethics Review Committee of the First Affiliated Hospital of Xinjiang Medical University (Ethical approval number: K202409-31). The studies were conducted in accordance with the local legislation and institutional requirements. The ethics committee/institutional review board waived the requirement of written informed consent for participation from the participants or the participants' legal guardians/next of kin because this study used historical data and did not intervene with patients.

Author contributions

YC: Conceptualization, Formal analysis, Methodology, Validation, Writing – original draft. KP: Data curation, Writing – review & editing. XL: Funding acquisition, Supervision, Writing – review & editing. EM: Supervision, Writing – review & editing. MW: Supervision, Project administration, Resources, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was supported by the Government of Xinjiang Uygur Autonomous Region, China, under the AIDS Clinical Key Specialty Construction Project. The funder played no role in study design, data collection, data analysis, interpretation and report writing.

Acknowledgments

We thank the Xinjiang Uygur Autonomous Region Government for the financial support of the AIDS Clinical Key Specialty Construction Project and the members of the group for their contributions.

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.

Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2025.1544351/full#supplementary-material

References

1. Schwetz TA, Fauci AS. The extended impact of human immunodeficiency virus/AIDS research. *J Infect Dis*. (2019) 219:6–9. doi: 10.1093/infdis/jiy441

2. Fauci AS, Lane HC. Four decades of HIV/AIDS — much accomplished, much to do. *N Engl J Med*. (2020) 383:1–4. doi: 10.1056/NEJMp1916753

3. Cohn LB, Chomont N, Deeks SG. The biology of the HIV-1 latent reservoir and implications for cure strategies. *Cell Host Microbe*. (2020) 27:519–30. doi: 10.1016/j.chom.2020.03.014

4. Van Welzen BJ, Oomen PGA, Hoepelman AIM. Dual antiretroviral therapy—all quiet beneath the surface? *Front Immunol*. (2021) 12:637910. doi: 10.3389/fimmu.2021.637910

5. Meng S, Tang Q, Xie Z, Wu N, Qin Y, Chen R, et al. Spectrum and mortality of opportunistic infections among HIV/AIDS patients in southwestern China. *Eur J Clin Microbiol Infect Dis*. (2023) 42:113–20. doi: 10.1007/s10096-022-04528-y

6. Puplampu P, Asafu-Adjaye O, Harrison M, Tetteh J, Ganu VJ. Opportunistic infections among newly diagnosed HIV patients in the largest tertiary facility in Ghana. *Ann Glob Health*. (2024) 90:13. doi: 10.5334/aogh.4149

7. Wembulua BS, Cisse VMP, Ka D, Ngom NF, Mboup A, Diao I, et al. Changes in early HIV/AIDS mortality rates in people initiating antiretroviral treatment between 2013 and 2023: a 10-year multicenter survival study in Senegal. *Infect Dis Now*. (2024) 54:104990. doi: 10.1016/j.idnow.2024.104990

8. Jassat W, Mudara C, Ozougwu L, Welch R, Arendse T, Masha M, et al. Trends in COVID-19 admissions and deaths among people living with HIV in South Africa: analysis of national surveillance data. *Lancet HIV*. (2024) 11:e96–e105. doi: 10.1016/S2352-3018(23)00266-7

9. Kidie AA, Masresha SA, Bizuneh FK. Statistical analysis on the incidence and predictors of death among second-line ART patients in public hospitals of north wollo and waghemira zones, Ethiopia, 2021. *Sci Rep*. (2024) 14:10893. doi: 10.1038/s41598-024-60119-1

10. Patel K, Zhang A, Zhang MH, Bunachita S, Baccouche BM, Hundal H, et al. Forty years since the epidemic: modern paradigms in HIV diagnosis and treatment. *Cureus*. (2021) 13:e14805. doi: 10.7759/cureus.14805

11. Javadi S, Menias CO, Karbasian N, Shaaban A, Shah K, Osman A, et al. HIV-related malignancies and mimics: imaging findings and management. *Radiogr Rev Publ Radiol Soc N Am Inc*. (2018) 38:2051–68. doi: 10.1148/rg.2018180149

12. AIDS and Hepatitis C Professional Group, Society of Infectious Diseases, Chinese Medical Association, Chinese Center for Disease Control and Prevention. Chinese guidelines for diagnosis and treatment of HIV/AIDS (2021 edition). *Zhonghua Nei Ke Za Zhi*. (2021) 60:1106–28. doi: 10.3760/cma.j.cn112138-20211006-00676

13. Barbier F, Mer M, Szychowiak P, Miller RF, Mariotte É, Galicier L, et al. Management of HIV-infected patients in the intensive care unit. *Intensive Care Med*. (2020) 46:329–42. doi: 10.1007/s00134-020-05945-3

14. Ron R, Martínez-Sanz J, Herrera S, Ramos-Ruperto L, Díez A, Sainz T, et al. CD4/CD8 ratio and CD8+ T-cell count as prognostic markers for non-AIDS mortality in people living with HIV. A systematic review and meta-analysis. *Front Immunol*. (2024) 15:1343124. doi: 10.3389/fimmu.2024.1343124

15. Su J, Liu J, Qin F, Chen R, Qin T, Tao X, et al. Effect of antiretroviral therapy on the mortality of HIV-1 infection long-term non-progressors: a cohort study. *BMC Infect Dis*. (2025) 25:72. doi: 10.1186/s12879-025-10448-x

16. Li Y, Ni Y, He Q, Hu X, Zhang Y, He X, et al. Survival analysis and immune differences of HIV long-term non-progressors in Xinjiang China: a 12-year prospective cohort observation. *AIDS Behav*. (2024) 28:3151–60. doi: 10.1007/s10461-024-04396-x

17. Yang X, Su B, Zhang X, Liu Y, Wu H, Zhang T. Incomplete immune reconstitution in HIV/AIDS patients on antiretroviral therapy: challenges of immunological non-responders. *J Leukoc Biol*. (2020) 107:597–612. doi: 10.1002/JLB.4MR1019-189R

18. Dagnaw Tegegne K, Cherie N, Tadese F, Tilahun L, Kassaw MW, Biset G. Incidence and predictors of opportunistic infections among adult HIV infected patients on anti-retroviral therapy at Dessie comprehensive specialized hospital, Ethiopia: a retrospective follow-up study. *HIVAIDS Res Palliat Care*. (2022) 14:195–206. doi: 10.2147/HIV.S346182

19. Zhao Y, Wei L, Dou Z, Zhao D, Gan X, Wu Y, et al. Changing mortality and patterns of death causes in HIV infected patients - China, 2013-2022. *China CDC Wkly*. (2023) 5:1073–8. doi: 10.46234/ccdcw2023.201

20. Masur H, Brooks JT, Benson CA, Holmes KK, Pau AK, Kaplan JE. Prevention and treatment of opportunistic infections in HIV-infected adults and adolescents: updated guidelines from the centers for disease control and prevention, national institutes of health, and HIV medicine association of the infectious diseases society of America. *Clin Infect Dis*. (2014) 58:1308–11. doi: 10.1093/cid/ciu094

21. Girma D, Dejene H, Adugna Geleta L, Tesema M, Bati F. Time to occurrence, predictors, and patterns of opportunistic infections incidence among HIV-positive patients attending antiretroviral therapy Clinic of Salale University Comprehensive Specialized Hospital: a retrospective cohort study. *Medicine*. (2022) 101:e29905. doi: 10.1097/MD.0000000000029905

22. Hou X, Wang D, Zuo J, Li J, Wang T, Guo C, et al. Development and validation of a prognostic nomogram for HIV/AIDS patients who underwent antiretroviral therapy: data from a China population-based cohort. *EBioMedicine*. (2019) 48:414–24. doi: 10.1016/j.ebiom.2019.09.031

23. Lang L, Wang T, Xie L, Yang C, Skudder-Hill L, Jiang J, et al. An independently validated nomogram for individualised estimation of short-term mortality risk among patients with severe traumatic brain injury: a modelling analysis of the CENTER-TBI China registry study. *EClinicalMedicine*. (2023) 59:101975. doi: 10.1016/j.eclinm.2023.101975

24. Yu A, Li Y, Zhang H, Hu G, Zhao Y, Guo J, et al. Development and validation of a preoperative nomogram for predicting the surgical difficulty of laparoscopic colectomy for right colon cancer: a retrospective analysis. *Int J Surg*. (2023) 109:870–8. doi: 10.1097/JS9.0000000000000352

25. Cachay ER, Gilbert T, Deiss R, Mathews WC. Shared decision-making concerning anal cancer screening in persons with human immunodeficiency virus. *Clin Infect Dis*. (2023) 76:582–91. doi: 10.1093/cid/ciac491

26. Chen J, Li L, Chen T, Yang X, Ru H, Li X, et al. Predicting the risk of active pulmonary tuberculosis in people living with HIV: development and validation of a nomogram. *BMC Infect Dis*. (2022) 22:388. doi: 10.1186/s12879-022-07368-5

27. Dong Y, Liu S, Xia D, Xu C, Yu X, Chen H, et al. Prediction model for the risk of HIV infection among MSM in China: validation and stability. *Int J Environ Res Public Health*. (2022) 19:1010. doi: 10.3390/ijerph19021010

28. Han C, Kim HI, Soh S, Choi JW, Song JW, Yoon D. Machine learning with clinical and intraoperative biosignal data for predicting postoperative delirium after cardiac surgery. *iScience*. (2024) 27:109932. doi: 10.1016/j.isci.2024.109932

29. Xie H, Deng Y-M, Li J-Y, Xie K-H, Tao T, Zhang J-F. Predicting the risk of primary sjögren's syndrome with key N7-methylguanosine-related genes: a novel XGBoost model. *Heliyon*. (2024) 10:e31307. doi: 10.1016/j.heliyon.2024.e31307

30. Wei Q, Mease PJ, Chiorean M, Iles-Shih L, Matos WF, Baumgartner A, et al. Machine learning to understand risks for severe COVID-19 outcomes: a retrospective cohort study of immune-mediated inflammatory diseases, immunomodulatory medications, and comorbidities in a large US health-care system. *Lancet Digit Health*. (2024) 6:e309–22. doi: 10.1016/S2589-7500(24)00021-9

31. Peng Z, Li X-J, Wang Y, Li Z-Y, Wang J, Chen C-L, et al. Gender potentially affects early postoperative hyponatremia in pituitary adenoma: XGBoost-based predictive modeling. *Heliyon*. (2024) 10:e28958. doi: 10.1016/j.heliyon.2024.e28958

32. Xie W, Li Y, Meng X, Zhao M. Machine learning prediction models and nomogram to predict the risk of in-hospital death for severe DKA: a clinical study based on MIMIC-IV, eICU databases, and a college hospital ICU. *Int J Med Inform*. (2023) 174:105049. doi: 10.1016/j.ijmedinf.2023.105049

33. Premeaux TA, Bowler S, Friday CM, Moser CB, Hoenigl M, Lederman MM, et al. Machine learning models based on fluid immunoproteins that predict non-AIDS adverse events in people with HIV. *iScience*. (2024) 27:109945. doi: 10.1016/j.isci.2024.109945

34. Nisa SU, Mahmood A, Ujager FS, Malik M. HIV/AIDS predictive model using random forest based on socio-demographical, biological and behavioral data. *Egypt Inform J*. (2023) 24:107–15. doi: 10.1016/j.eij.2022.12.005

35. Burns CM, Pung L, Witt D, Gao M, Sendak M, Balu S, et al. Development of a human immunodeficiency virus risk prediction model using electronic health record data from an academic health system in the southern United States. *Clin Infect Dis*. (2023) 76:299–306. doi: 10.1093/cid/ciac775

36. Amiri P, Montazeri M, Ghasemian F, Asadi F, Niksaz S, Sarafzadeh F, et al. Prediction of mortality risk and duration of hospitalization of COVID-19 patients with chronic comorbidities based on machine learning algorithms. *Digit Health*. (2023) 9:205520762311704. doi: 10.1177/20552076231170493

37. Gao Y, Cai G-Y, Fang W, Li H-Y, Wang S-Y, Chen L, et al. Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. *Nat Commun*. (2020) 11:5033. doi: 10.1038/s41467-020-18684-2

38. Zhang Z. Univariate description and bivariate statistical inference: the first step delving into data. *Ann Transl Med*. (2016) 4:91–1. doi: 10.21037/atm.2016.02.11

39. Zhang Z, Gayle AA, Wang J, Zhang H, Cardinal-Fernández P. Comparing baseline characteristics between groups: an introduction to the CBCgrps package. *Ann Transl Med*. (2017) 5:484–4. doi: 10.21037/atm.2017.09.39

40. Huang L, Xie B, Zhang K, Xu Y, Su L, Lv Y, et al. Prediction of the risk of cytopenia in hospitalized HIV/AIDS patients using machine learning methods based on electronic medical records. *Front Public Health*. (2023) 11:1184831. doi: 10.3389/fpubh.2023.1184831

41. Leal JA, Fausto MA, Carneiro M, Tubinambás U. Prevalence of hypoalbuminemia in outpatients with HIV/AIDS. *Rev Soc Bras Med Trop*. (2018) 51:203–6. doi: 10.1590/0037-8682-0093-2017

42. Mandikiyana Chirimuta LA, Shamu T, Chimbetete C, Part C. Incidence and risk factors of anaemia among people on antiretroviral therapy in Harare. *South Afr J HIV Med*. (2024) 25:1605. doi: 10.4102/sajhivmed.v25i1.1605

43. Lang R, Coburn SB, Gill MJ, Grossman J, Gebo KA, Horberg MA, et al. The association of anemia with survival among people with HIV following antiretroviral initiation in the NA-ACCORD 2007-2016. *J Acquir Immune Defic Syndr*. (2024) 97:334–43. doi: 10.1097/QAI.0000000000003502

44. Metcalfe R, Fraser R, Trayner KMA, Glancy M, Yeung A, Sills L, et al. Rising mortality among people who inject drugs living with HIV in Scotland, UK: a 20-year retrospective cohort study. *HIV Med*. (2024) 26:265–74. doi: 10.1111/hiv.13733

45. Wang S, Tang H, Zhao D, Cai C, Jin Y, Qin Q, et al. Survival of people living with HIV/AIDS from pre-ART era to treat-all era - China, 1985-2022. *China CDC Wkly*. (2024) 6:1264–70. doi: 10.46234/ccdcw2024.253

46. Feng Q, Hao J, Li A, Tong Z. Nomograms for death from pneumocystis jirovecii pneumonia in HIV-uninfected and HIV-infected patients. *Int J Gen Med*. (2022) 15:3055–67. doi: 10.2147/IJGM.S349786

47. Wolinsky E. Mycobacterial diseases other than tuberculosis. *Clin Infect Dis Off Publ Infect Dis Soc Am*. (1992) 15:1–12. doi: 10.1093/clinids/15.1.1

48. Wu L, Zhang Z, Wang Y, Hao Y, Wang F, Gao G, et al. A model to predict in-hospital mortality in HIV/AIDS patients with pneumocystis pneumonia in China: the clinical practice in real world. *Biomed Res Int*. (2019) 2019:1–11. doi: 10.1155/2019/6057028

49. Gri J, Jain V. Pneumocystis jirovecii pneumonia: A case report. *J Med Case Rep*. (2024) 18:52. doi: 10.1186/s13256-024-04350-4

50. Atkinson A, Zwahlen M, Barger D, d'Arminio Monforte A, De Wit S, Ghosn J, et al. Withholding primary pneumocystis pneumonia prophylaxis in virologically suppressed patients with human immunodeficiency virus: an emulation of a pragmatic trial in COHERE. *Clin Infect Dis Off Publ Infect Dis Soc Am*. (2021) 73:195–202. doi: 10.1093/cid/ciaa615

51. Li Y, Feng Y, He Q, Ni Z, Hu X, Feng X, et al. The predictive accuracy of machine learning for the risk of death in HIV patients: a systematic review and meta-analysis. *BMC Infect Dis*. (2024) 24:474. doi: 10.1186/s12879-024-09368-z