


Prediction of new HIV infection in men who have sex with men based on machine learning: secondary analysis of a prospective cohort study from Western China

Kangjie Li^{a#}, Guiqian Shi^{a#}, Cong Zhang^a, Bing Lin^a, Yi Tao^{a,b}, Qian Wang^a, Haijiao Zeng^a, Jielian Deng^a, Lei Zou^{a,c}, Biao Xie^a and Xiaoni Zhong^a 

^aSchool of Public Health, Research Center for Medicine and Social Development, Chongqing Medical University, Chongqing, China; ^bThe First Affiliated Hospital of Chongqing Medical University, Chongqing, China; ^cCenter for Disease Control and Prevention of Jiulongpo District, Chongqing, China

ABSTRACT

Objective: This study aimed to construct a model based on machine learning to predict new HIV infections in HIV-negative men who have sex with men (MSM).

Methods: This is a secondary analysis of a previous random clinical trial aiming to evaluate the preventive effects of PrEP on new HIV infection in MSM. During 2013–2015, 1455 HIV-negative MSM were enrolled. Participants were divided into treatment group and control group and regularly followed up until they seroconverted to HIV positive or until the 2-year endpoint reached. Five machine-learning approaches were applied to predict the risk of HIV infection. Model performance was evaluated using Harrel's C-index and area under the receiver operator characteristic curve (AUC) and validated in an external validation cohort. To explain this model, shapley additive explanation (SHAP) values were calculated and visualized.

Results: During the observation period, 102 MSM developed HIV infection. Thirteen parameters are selected to construct the model. The random survival forest model showed the best performance in the validation cohort, with a C-index of 0.7013, and could significantly categorize MSM into three groups. Our model indicated that MSM with younger age, receptive anal intercourse, and multiple male sexual partners had an increased risk of HIV infection, and those with higher AIDS knowledge scores had a lower risk.

Conclusion: We presented a machine learning-based model to predict their risk of developing HIV infection. This model could be applied to recognize MSM who are at a higher risk of developing HIV infection.

Abbreviations: MSM: men who have sex with men; SHAP: shapley additive explanation; STI: sexually transmitted diseases; AUC: area under the receiver operator characteristic curve; PrEP: pre-exposure prophylaxis; CPH: Cox proportional hazards regression model; MICE: multiple imputation by chained equation; RSF: random survival forest; SSVM: survival support vector machine; GBM: gradient boosting survival model; XGBoost: extreme gradient boosting survival model; DeepSurv: deep learning-based survival model; IA: insertive anal intercourse; RA: receptive anal intercourse

ARTICLE HISTORY

Received 29 September 2024

Revised 7 January 2025

Accepted 6 February 2025

KEYWORDS

Men who have sex with men; HIV infection; machine learning; random survival forest


Introduction

The prevalence of HIV/AIDS among men who have sex with men (MSM) is a global public health concern. In recent years, there has been a high rate of HIV infection and its incidence among MSM. Until 2022, MSM

accounted for 25.6% of newly reported HIV cases in China, which is significantly higher than the 2.5% reported in 2006 [1]. The HIV infection rate among MSM in China is approximately 7% [2], while in the general population, it ranges from 5.61 cases/100 person-years

CONTACT Xiaoni Zhong  zhongxiaoni@cqmu.edu.cn; Biao Xie  kybiao@cqmu.edu.cn  School of Public Health, Research Center for Medicine and Social Development, Chongqing Medical University, Chongqing, China.

[#]These authors contributed equally to this work and are the first co-author.

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/07853890.2025.2476040>.

© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

to as high as 18.80 cases/100 person-years [3]. The risk of HIV transmission among MSM remains high.

To reduce the HIV/AIDS epidemic, the UNAIDS has advocated ending the HIV/AIDS epidemic by 2030. The first goal is to have 95% of individuals infected with HIV aware of their status. However, in China, more than 20% of HIV-infected individuals remain unaware of their status [4], and only 61.2% of MSM have undergone HIV testing and know their infection status [5]. The low rate of HIV testing is often attributed to a lack of awareness of the risk of HIV infection risk [6]. Additionally, pre-exposure prophylaxis (PrEP) has been globally recommended for populations at risk of HIV infection [7]. PrEP has not been widely utilized [8] despite its ability to reduce HIV infections by more than 90% [9]. The reason for this, in addition to lower risk awareness [10], is the inability of primary healthcare providers to identify potential candidates who would benefit from PrEP [11]. An effective tool for identifying high-risk MSM for developing HIV infection might help health workers provide preventive services, such as PrEP medicine, to those in urgent need.

Machine learning algorithms have been demonstrated to be useful for predicting the outcomes of different diseases. Unlike conventional methods such as logistic regression [12] or the Cox proportional hazards regression model (CPH) [6,13], machine learning algorithms do not rely on statistical inference or assumptions and exhibit superior performance in handling complex nonlinear relationships and capturing patterns of high-dimensional data [14]. Several studies have presented evidence of machine learning algorithms based on clinical data to predict HIV infection risk among MSM. A study conducted in Australia developed and validated an HIV diagnostic model using gradient boosting with an area under the receiver operator characteristic curve (AUC) of 0.763 [15]. Duthe et al. employed a combination model (least absolute shrinkage and selection operator, random forest, and generalized linear model) to identify MSM individuals at risk of HIV infection and to recognize potential PrEP candidates, achieving an AUC of 0.888 [16]. Another study in Zimbabwe showed the excellent performance of recursive neural networks in predicting HIV infection status among MSM with an AUC of 0.940 [17]. He et al. utilized sentinel surveillance data of MSM individuals in Zhejiang province to predict HIV infection risk, and the random forest algorithm achieved the best performance (AUC = 0.846) [18]. However, these were cross-sectional studies that emphasized screening of the infected individuals in the populations, focusing predominantly on diagnosis.

Additionally, these cross-sectional studies did not take the time which MSM became HIV-positive into account.

Longitudinal studies possess the advantage of elucidating causal relationships, a capability that cross-sectional studies lack [19], and produce time-to-event data. In the context of time-to-event data, the exclusive application of conventional logistic regression or binary-class machine learning models may lead to a substantial loss of data information. With the advancement of artificial intelligence, applying machine learning to time-to-event data significantly enhances the accuracy of prediction models relative to traditional CPH. Thus, the prediction model based on longitudinal data was more convincing in identifying MSM at high risk of HIV infection. In the present study, we conducted a secondary analysis of our previous random clinical trial which aimed to evaluate the preventive effects of different PrEP medication strategies on new HIV infections in HIV-negative MSM. The enrolled MSM were treated with different PrEP strategies, and followed up until they seroconverted to HIV positive or until the 2-year endpoint reached. Here, we developed a machine learning model to predict new HIV infections in MSM based on our previously conducted prospective cohort study.

Material and methods

Study design and participants

This present study is a secondary analysis of a randomized clinical trial and the design of this study has been previously described [20]. Briefly, the multicentre, parallel controlled clinical trial was conducted to evaluate the preventive effect of different PrEP medication strategies on new HIV infections in HIV-negative MSM. All participants were recruited from 2013 to 2015 in four provinces of China: Chongqing, Sichuan, Guangxi, and Xinjiang (registration number: ChiCTR-TRC-13003849). Participants in Chongqing, Sichuan, and Xinjiang were labelled with a random number at the entry of the cohort and were then randomly divided into daily PrEP, event-driven, and blank control groups at a 1:1:1 ratio. Participants in Guangxi were recruited by the local CDC, not by our researchers, and were given the chance to select which group they would prefer. MSM in the daily PrEP group were administered 300mg TDF orally per day (Gilead Sciences, Inc. (Foster City, CA, USA), specifications: 300mg per tablet. Lot: A818213). The event-driven group was supposed to take 300mg TDF orally 48–24h before sexual activity and another 300mg TDF 2h after sexual activity. The blank control group did not receive any drugs or placebo. All enrolled MSM

in each group were followed up every 12 weeks. During each follow-up visit, we conducted a follow-up questionnaire survey and serum HIV antibody testing. The follow-up questionnaire for participants included information about sexual behaviours over the past two weeks and the medication adherence. Participants were regularly followed up until they became HIV positive or until the 2-year endpoint reached. More information for the primary study could be found in the published article [20].

The inclusion criteria for MSM were as follows: (1) 18–65 years, (2) HIV negative at enrolment, (3) had at least one sexual intercourse with men every two weeks, (4) one more male sex partner before their entry into this study, and (5) signed the informed consent form. In the current study, individuals who did not participate any of the follow-up visits were excluded for analysis. Written informed consent was obtained from all individual participants included in the primary study, which declared that they consented to the use of their data for research related to this study.

Data collection

At recruitment, the participants were required to complete a questionnaire investigating their demographic information, AIDS-related knowledge, and sexual behaviours. Demographic information included age, monthly income, education, career, and marital status. At each follow-up, blood was collected to test for the presence of HIV-1 and HIV-2 antibodies using ELISA. Once the HIV-1 or HIV-2 antibody became positive, the participant met our endpoint and we ceased follow-up for this participant. Medication adherence to PrEP was evaluated at each follow-up based on the latest two-week medication: medication rate = (number of pills that should have been taken, number of pills that were missed)/(number of pills that should have been taken).

Missing data handling

For missing data with missing proportion < 20%, multiple imputations using the chained equation (MICE) were conducted for numerical and category variables. The proportion of missing data for each predictor is shown in Figure S1. In our data, the missing proportion of all variables was <10%.

Derivation and validation cohort

The participants in the derivation cohort were enrolled from Chongqing, Sichuan, and Guangxi provinces. The

participants from Xinjiang were divided into validation cohorts. For numerical variables, Student's t-test was used to compare the differences in characteristics between the derivation and validation cohorts. A chi-square test was conducted to detect differences in categorical variables. $p < 0.05$ was considered statistically significant.

Predictors selection

We used a random survival forest (RSF) to identify important predictors of new HIV infection. First, all 19 predictors were included in the model, and the importance of the variables in this model was calculated using the permutation-based variable importance scores. If one variable negatively contributed to the model, it was removed from the next model construction. Second, all the predictors that positively contributed to the model in the last step were included to build the next model. This process was cycled for 500 iterations to ensure that all predictors included in the final model positively contributed to the model.

Outcome

New HIV infection during the follow-up visits was defined as the event of this analysis.

Development and validation of machine learning models

Currently, mainstream machine learning models are primarily based on the following theories: decision tree theory, ensemble learning theory, support vector machine theory, and neural network theory. In this study, five machine learning approaches were utilized to predict new HIV infections in MSM: random survival forest model (RSF), survival support vector machine (SSVM), gradient boosting survival model (GBM), extreme gradient boosting survival model (XGBoost), and deep learning-based survival model (DeepSurv). The random survival forest model, gradient boosting survival model, and extreme gradient boosting model are extensions of the random forest model, gradient boosting model, and extreme gradient boosting model, respectively, which were developed based on decision tree models and designed for survival data. DeepSurv is a deep neural network-based Cox proportional hazards approach that analyzes the effects of covariates on patient outcome. SSVM is an extension of a support vector machine for censored survival data. Additionally, we constructed a traditional multivariate CPH to predict new HIV infection in MSM.

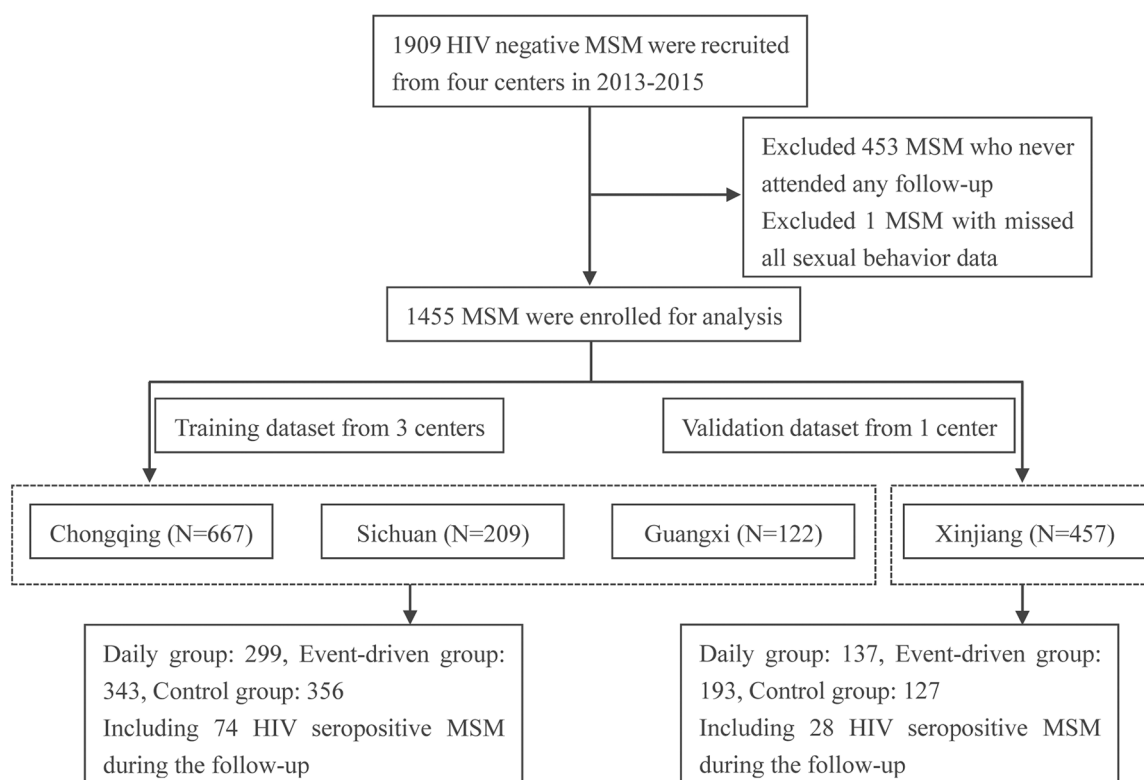


Figure 1. Flowchart of this study.

For the development of the machine learning model, 5-folds cross-validation was used to train the model, which enabled it to confirm the best hyperparameters. The 5-fold cross-validation split the derivation cohort into five folds, in which four folds were defined as the training datasets and the remaining as the test dataset. Each proportion of the five folds was used as a test dataset. The performance of each model was evaluated using Harrel's c-index. When the averages of the C-index reached a peak, these hyperparameters were selected to construct the machine-learning model. The performance of the model was validated in an external validation cohort, which was also evaluated using the C-index. In addition, we employed calibration curve and Brier score to further illustrate the performance of the final model.

To interpret the machine learning model, the SHAP value was calculated to explore the contribution of each variable to HIV infection.

Statistical analysis

R software (version 4.3.0) was used to construct the machine learning models, including the RSF, GBM, XGBoost, and SSVM models. The DeepSurv model was constructed using the deepsurv package in Python software. In addition, the python package shap v0.44.0, was applied to calculate the SHAP values. For the sensitivity analysis, we used a model with missing data to

verify the stability of the model. $p < 0.05$ was considered significant differences.

Results

Baseline characteristics of enrolled MSM

During 2013–2015, we recruited 1909 qualified HIV-negative MSM. A total of 453 MSM were excluded from this study because of the absence of any of follow-up visits, and one participant was excluded due to the lack of all sexual behaviour data. Eventually, 1455 MSM were enrolled for analysis, of whom 998 were in the derivation cohort and 457 were in the validation cohort (Figure 1). The baseline characteristics of the enrolled MSM from the two cohorts are stated in Table 1. The mean age of MSM in the derivation cohort was not statistically different from that of MSM in the validation cohort ($p = 0.716$). Additionally, sex role, number of sex partners, use of condoms, diagnosis of sexually transmitted diseases, and frequency of commercial sex were also found to be comparable between the derivation and validation cohorts ($p > 0.05$). In the derivation cohort, a higher proportion of MSM came from the countryside ($p < 0.001$) and the AIDS-related knowledge score was significantly lower ($p = 0.004$). In the derivation cohort, more MSM reported using illicit drugs in the past six months ($p = 0.019$). The PrEP medication adherence was lower

Table 1. Comparison of baseline characteristics between derivation cohort and validation cohort.

	Validation cohort (N = 457)	Derivation cohort (N = 998)	P-value
New HIV infection			
Negative	429 (93.9%)	924 (92.6%)	0.439
Positive	28 (6.1%)	74 (7.4%)	
Age			
Mean (SD)	30.3 (7.62)	30.1 (9.12)	0.716
PrEP strategy			
Daily	137 (30.0%)	299 (30.0%)	0.004
Event driven	193 (42.2%)	343 (34.4%)	
None	127 (27.8%)	356 (35.7%)	
Residence			
City	386 (84.5%)	679 (68.0%)	<0.001
Countryside	71 (15.5%)	317 (31.8%)	
Missing	0 (0%)	2 (0.2%)	
Nationality			
Han	401 (87.7%)	946 (94.8%)	<0.001
Minority	55 (12.0%)	52 (5.2%)	
Missing	1 (0.2%)	0 (0%)	
Education			
Illiteracy	2 (0.4%)	8 (0.8%)	<0.001
Junior	5 (1.1%)	28 (2.8%)	
Middle	27 (5.9%)	110 (11.0%)	
Senior	95 (20.8%)	308 (30.9%)	
Junior college	102 (22.3%)	241 (24.1%)	
College	224 (49.0%)	303 (30.4%)	
Missing	2 (0.4%)	0 (0%)	
Career			
Employed	365 (79.9%)	748 (74.9%)	0.009
Retired	0 (0%)	12 (1.2%)	
Student	62 (13.6%)	135 (13.5%)	
Unemployed	29 (6.3%)	101 (10.1%)	
Missing	1 (0.2%)	2 (0.2%)	
Marriage			
Spinsterhood	360 (78.8%)	721 (72.2%)	0.050
Married	62 (13.6%)	188 (18.8%)	
Divorced	35 (7.7%)	88 (8.8%)	
Widow	0 (0%)	1 (0.1%)	
Monthly income			
<1 k	64 (14.0%)	171 (17.1%)	<0.001
1–3 k	108 (23.6%)	427 (42.8%)	
3–5 k	198 (43.3%)	296 (29.7%)	
5–10 k	71 (15.5%)	72 (7.2%)	
>10 k	9 (2.0%)	19 (1.9%)	
Missing	7 (1.5%)	13 (1.3%)	
AIDS knowledge score			
High	202 (44.2%)	360 (36.1%)	0.004
Low	255 (55.8%)	638 (63.9%)	
HIV test			
No	39 (8.5%)	242 (24.2%)	<0.001
Yes	418 (91.5%)	753 (75.5%)	
Missing	0 (0%)	3 (0.3%)	
HIV consultation			
No	96 (21.0%)	432 (43.3%)	<0.001
Yes	360 (78.8%)	561 (56.2%)	
Missing	1 (0.2%)	5 (0.5%)	
Sex role			
Insertive only	116 (25.4%)	264 (26.5%)	0.220
Insertive mostly	102 (22.3%)	214 (21.4%)	
Equal to be insertive or receptive	106 (23.2%)	277 (27.8%)	
Receptive mostly	80 (17.5%)	143 (14.3%)	
Receptive only	52 (11.4%)	96 (9.6%)	
Missing	1 (0.2%)	4 (0.4%)	
Number of sex partner			
≤1	344 (75.3%)	743 (74.4%)	0.502
2–5	111 (24.3%)	245 (24.5%)	
6–9	2 (0.4%)	5 (0.5%)	
≥10	0 (0%)	5 (0.5%)	
Condom			
Always use	263 (57.5%)	497 (49.8%)	0.060
Sometimes use	137 (30.0%)	295 (29.6%)	
Never use	31 (6.8%)	97 (9.7%)	
Missing	26 (5.7%)	109 (10.9%)	
Sexually transmitted diseases (STI)			

(Continued)

Table 1. Continued.

	Validation cohort (N=457)	Derivation cohort (N=998)	P-value
No	414 (90.6%)	918 (92.0%)	0.398
Yes	42 (9.2%)	77 (7.7%)	
Missing	1 (0.2%)	3 (0.3%)	
Commercial sex			
No	438 (95.8%)	940 (94.2%)	0.330
Yes	19 (4.2%)	55 (5.5%)	
Missing	0 (0%)	3 (0.3%)	
Illicit drug			
No	429 (93.9%)	967 (96.9%)	0.019
Yes	19 (4.2%)	19 (1.9%)	
Missing	9 (2.0%)	12 (1.2%)	

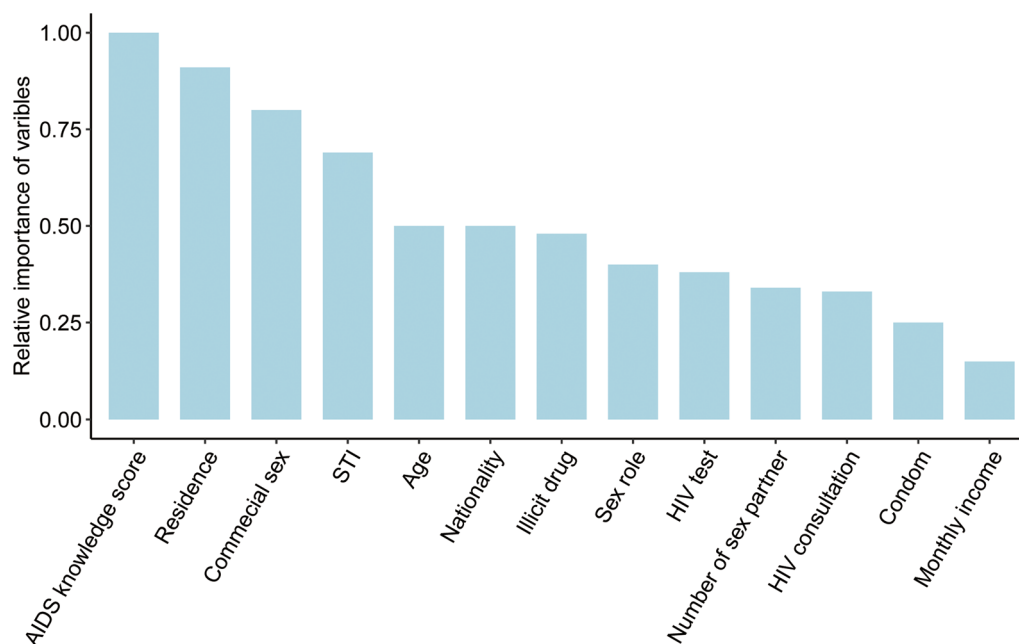


Figure 2. Relative variable importance of the RSF model for the prediction of HIV infection in the derivation cohort using the permutation-based parameter importance scores.

than 80%, with mean medication adherence of 57% and 39% in the daily PrEP and event-driven groups, respectively (Figure S2).

HIV seroconversion rate

In the derivation cohort, during the mean observation period of 1.18 years, 74 MSM (7.4%) were newly infected with HIV. In the validation cohort, the mean observation period was 1.01 years and 28 MSM were infected with HIV, with an infection rate of 6.1%. HIV seroconversion rates between the derivation and validation cohorts were comparable ($p=0.439$).

RSF model performs best for predicting new HIV infection in MSM

After 500 iterations, 13 parameters were selected to construct the models, including age, AIDS-related

knowledge score, sex role, and number of sex partners. The relative importance of the included parameters was calculated using a permutation-based importance score, as depicted in Figure 2.

Among these machine learning models, extreme overfitting was observed in the SSVM model, with a C-index of 0.9741 for the derivation cohort and 0.4956 for the validation cohort (Figure 3A,B; Table S2). The discriminative ability of the DeepSurv model was poor in both cohorts. In the derivation cohort, the RSF, GBM, and XGBoost models achieved good performance, with a C-index exceeding 0.75. For the CPH model, fair agreement was achieved, with a C-index of 0.7075. In the validation cohort, RSF outperformed other machine learning approaches and had superior discriminative ability for predicting new HIV infection in MSM, with a C-index of 0.7013, compared to GBM (C-index: 0.6732), CPH (C-index: 0.6454), and XGBoost (C-index: 0.6261). Moreover, we evaluated the

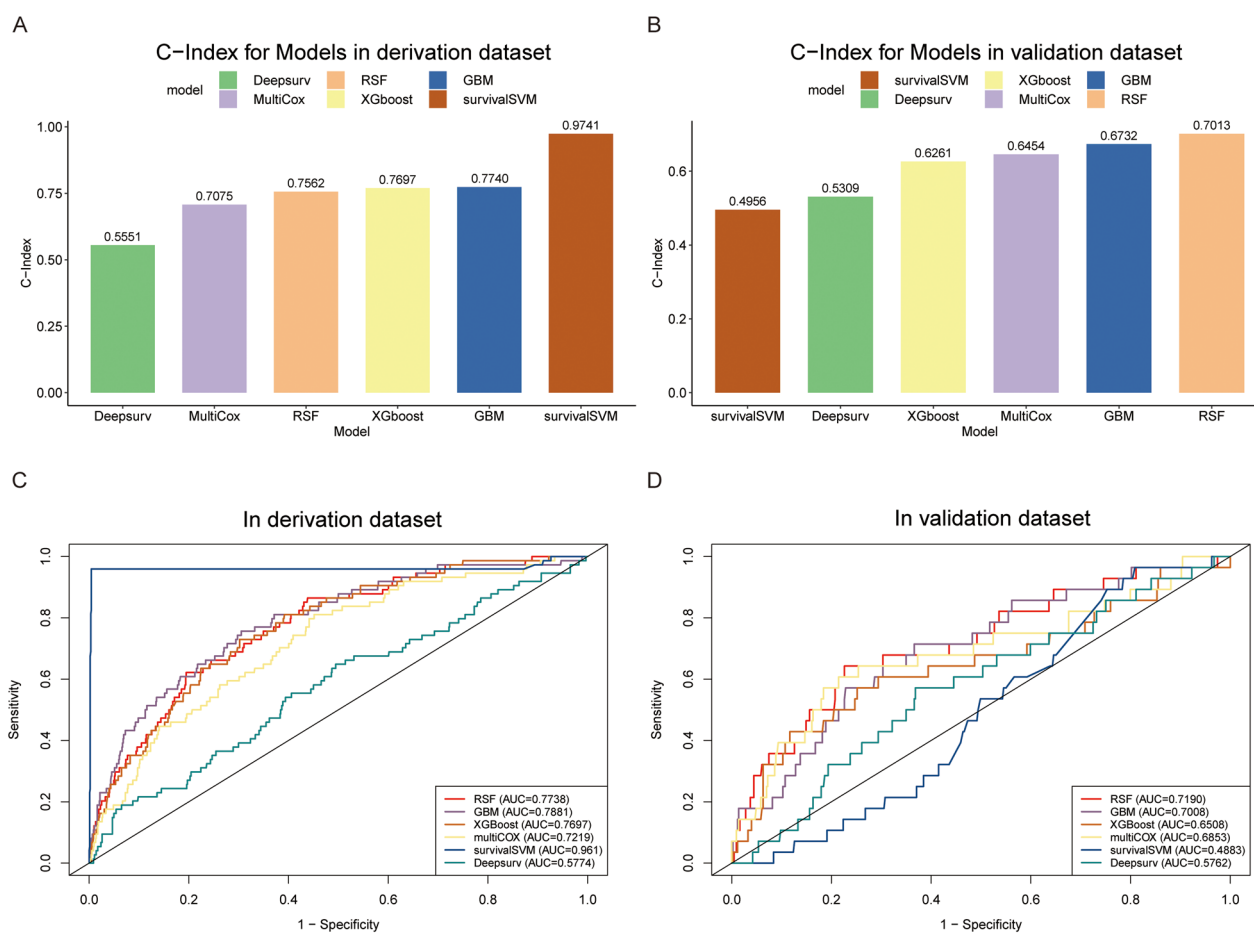


Figure 3. Performance assessment of the six models by c-index and AUC. Random survival forest model performed best in the validation cohort.

prediction accuracy of these models by using the receiver operating characteristic curve. The results demonstrated that the RSF model outperformed the CPH model and other machine-learning models (Figure 3C,D). The calibration curve and Brier score indicated the RSF model performed well in both derivation and validation cohorts (Figure S3).

Overall, the RSF model outperformed other machine learning and CPH models. Thus, the RSF model was used for further analysis.

RSF could significantly discriminate high-risk MSM of developing HIV

Based on the tertiles of the risk score in derivation dataset, we categorized the samples in both derivation cohort and validation cohort into three risk groups: low-risk, middle-risk and high-risk. The results of Kaplan–Meier curves showed that the cumulative HIV infection probability was significantly different in both the derivation and validation cohorts ($p < 0.01$, log-rank

test). MSM in the high-risk group had a higher probability of developing an HIV infection (Figure 4A,B).

Interpretation of the RSF model

To determine how these predictors modulate HIV infection risk, we implemented SHAP values to explain the output of the RSF model. The SHAP summary plot described below describes the impact of these variables on HIV infection risk (Figure 5). According to the results, younger MSM were at higher risk of HIV infection. Compared with MSM who only engaged in insertive anal intercourse (IAI), the HIV infection risk of MSM who only engaged in receptive anal intercourse (RAI) increased significantly. In addition, MSM with higher AIDS-related knowledge scores, without a history of sexually transmitted diseases (STI), living in a city, fewer sex partners, always using condoms, and never engaged in commercial sex had a decreased risk of HIV infection. On the other hand, MSM who had never attended any HIV consultation or HIV test were found to be at a

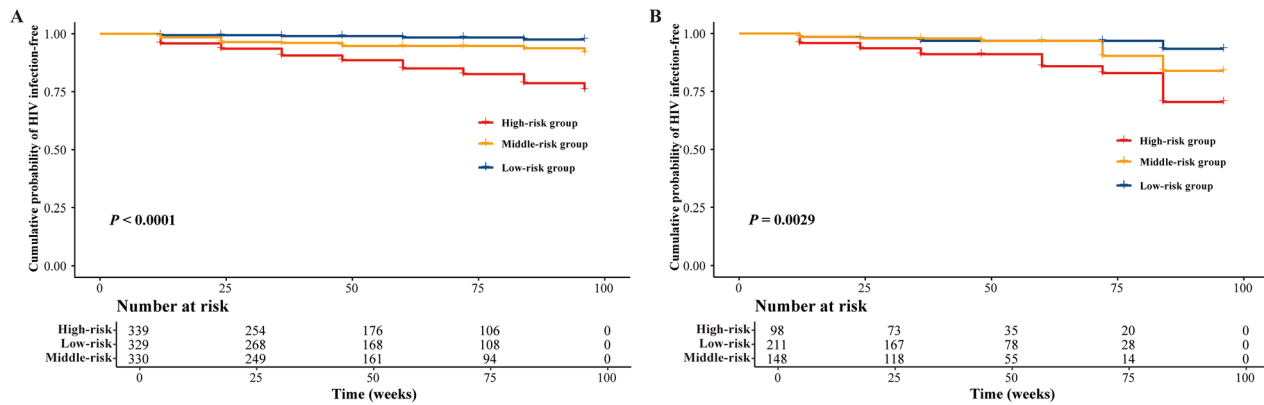


Figure 4. Kaplan–Meier curve of HIV infection in derivation cohort (A) and validation plot (B). Patients in both cohorts were divided into high-risk group, middle-risk group and low-risk group based on the tertile of risk score in derivation cohort.

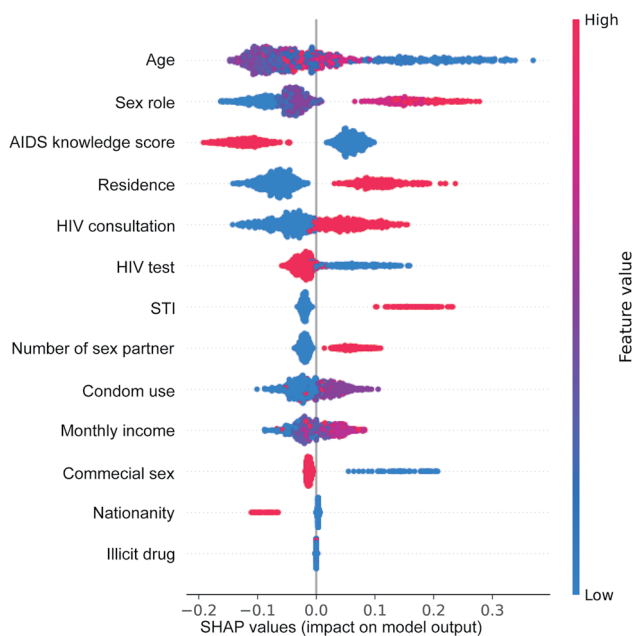


Figure 5. SHAP summary plot of RSF model. The colour of the dots indicated the feature value. The higher the SHAP value, the higher the risk of HIV infection.

higher risk of HIV infection. These results suggested that participants who were more concerned about their health had a lower risk of HIV infection.

Furthermore, we noticed that younger MSM were at higher risk of HIV infection; therefore, we compared the differences in sexual behaviours between younger MSM and older MSM, grouped by median age. We found that younger MSM tended to engage in RAI and had more sex partners (Table S1), which might explain why younger MSM were at a higher risk of HIV acquisition.

Sensitivity analysis of RSF model

We also assessed the performance of the RSF model using missing data to evaluate its stability. The results

indicated that RSF performed well in the presence of missing data, with a C-index of 0.7553.

Discussion

In this study, we aimed to develop and validate a machine learning algorithm to refine the accuracy of HIV risk prediction in the MSM population, which could help identify individuals at an increased risk of HIV infection, improve the utilization of preventive measures, and help reduce HIV infection. Our findings demonstrate that machine learning algorithms improve the prediction accuracy of new HIV infections in MSM. The predictive performance of the RSF model surpassed that of the other machine learning models and the conventional CPH model, yielding the highest C-index. Sensitivity analysis revealed that the RSF model maintained robust predictive performance, even in scenarios involving missing data. Hence, compared to traditional models, machine learning algorithms have the potential to enhance the performance of HIV prediction models, which is consistent with previous research outcomes [8,15]. Besides, our study used the time-to-event data to construct the model based on the prospective study in West China, which was also different from previous cross-sectional study. Longitudinal studies possess the advantage of elucidating causal relationships [19]. Therefore, our study, as compared to previous cross-sectional or retrospective studies, is better to demonstrate the causal link between sexual behaviour and HIV infection. Specifically, our model is useful for practical application by public health workers. They can utilize the model to calculate the HIV infection risk for MSM, take timely intervention measures for those at high risk, and prioritize interventions for those at higher risk of HIV acquisition. However, the predictive ability of our model was not perfect, with C-index of 0.7013.

Therefore, we recommend combining the results of our model with other HIV risk assessment tools to determine the HIV infection risk level of MSM. One Chinese researcher developed an HIV risk assessment tool for HIV-negative MSM in west China based on Delphi methods [21], with higher score indicating a higher risk of HIV infection. However, the researchers did not provide a cut-off value to distinguish risk levels. The absence of a clear threshold can lead to uncertainty in clinical decision-making. It would be helpful for the public health workers to assess the HIV infection risk of MSM by combining this tool and our model.

In the RSF model, the AIDS knowledge score, residence, commercial sexual behaviour, and STI history were identified as the most critical predictive variables for HIV infection. MSM with lower AIDS-related knowledge had a higher risk of acquiring HIV. Another study has also shown that MSM with higher AIDS-related knowledge scores have a 0.61 times lower likelihood of HIV infection [18], consistent with our conclusion. Most research reports have indicated that individuals residing in urban areas have a higher likelihood of HIV infection, primarily due to greater wealth, which provides them with more opportunities for sexual partners [22–25]. Conversely, we found that rural MSM were more susceptible to HIV infection, which might be because rural MSM are more likely to engage in high-risk behaviour and primarily meet sexual partners through dating applications [26]. Commercial sexual behaviour is an important risk indicator for predicting new HIV infections. A meta-analysis showed that the risk of new HIV infections was greater in the MSM population engaged in commercial sex, with an HR of 4.11 [27]. Similarly, a history of STI diagnosis is a major risk factor, as evidenced by various studies [15,18]. A diagnosis of STI can serve as a predictive factor for HIV due to shared high-risk behaviours [28].

We also found that sex role is an important factor for HIV infection. Our results suggest that MSM engaging only in RAI have a higher risk of developing HIV infection compared to those who only engage in IAI. The existing meta-analysis supported our findings, which showed that the risk of HIV infection for MSM practicing RAI was 6.2 times higher than for those practicing only IAI [29]. One possible explanation for this may be that the anal mucosa is more susceptible to HIV infection than the keratinized squamous epithelium of the penis [29]. Second, engaging in RAI may drive changes in the gut microbiota [30], causing immune activation and consequently increasing the risk of HIV infection [31]. This suggests that MSM engaging in RAI need to pay more attention to safe

sexual behaviour, and health workers are expected to prioritize effective interventions for MSM who often conduct receptive anal intercourse.

Notably, the PrEP strategy did not emerge as a significant predictive factor in our predictive model largely because of the low medication adherence of the study participants. In fact, the efficacy of oral PrEP depends heavily on medication adherence [32]. In this cohort, mean medication adherence was approximately 57% and 39% in the daily PrEP and event-driven groups, respectively. Our previously published research showed that the differences in the HIV incidence rates among the daily PrEP, event-driven, and blank control groups were not statistically significant, and the HIV incidence rates in the daily PrEP and event-driven groups were significantly lower than those in the blank control group, but when adherence to medication was $\geq 80\%$ [20]. Therefore at this perspective, the treatment and blank control groups of the current study can be considered homogeneous populations. Thus, our model is suitable for HIV-negative MSM populations who have not used PrEP medication. It is also applicable when MSM have taken PrEP drugs but with low adherence, such as below 80%. It is worth noting that this result does not suggest that PrEP could not prevent HIV infection. In contrast, high medication adherence is the key to achieving effective HIV prevention; thus, we advocate improving medication adherence to PrEP. We have also performed a lot of work to improve medication adherence to PrEP. For instance, we developed a reminder system based on WeChat to remind MSM to take pill on time [33].

However, our study had several limitations. First, the data we investigated stemmed from self-reported information, potentially harbouring a degree of reporting bias. Second, we did not consider the changes of the sexual behaviours in the final machine learning model. In our primary study, we investigated participants' sexual behaviours over a two-week period during each follow-up visits, which was different from the behavioural information in the baseline. Therefore we did not include follow-up behaviour data in our model. Future studies are encouraged to utilize the dynamic behavioural data to construct the prediction model. Finally, our model was developed and validated based on data from Western China, thus necessitating further validation of its applicability in different settings.

Conclusion

In summary, compared to the conventional CPH model, the RSF model, as a machine learning algorithm, exhibited better performance in predicting new HIV

infections among the MSM population. The developed and validated RSF model can be applied to stratify HIV infection risks among the MSM population, aiming to improve risk awareness and identify potential PrEP candidates, ultimately facilitating effective utilization of preventive measures.

Acknowledgements

Not available.

Authors contributions

Conceptualization: Zhong and Xie; Funding acquisition: Zhong; Data curation: Li, Shi, Lin, and Tao; Formal analysis: Li, Shi, Zeng, and Wang; Methodology: Zhang, Deng, and Zou; Writing-original draft: Li, and Shi; Writing-review & editing: Zhong, Xie, Li, and Shi. All the authors have read and approved the final manuscript.

Ethics approval

The current study was a secondary analysis of our previous clinical trial. Our primary study was conducted in accordance with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of Chongqing Medical University (Ethical Approval code: 2012010).

Consent to participate

Informed consent was obtained from all enrolled individual participants in the primary study. All participants declared that they consented to the use of their data for related researches.

Disclosure statement

No potential conflict of interest was reported by the authors.

Data availability

The data reported in this work is available upon request from the lead contact, Prof. Zhong (zhongxiaoni@cqmu.edu.cn).

Funding

This study was supported by the National Key Project for Infectious Diseases of the Ministry of Science and Technology of China (2012ZX10001007-007 and 2018ZX10721102-005).

ORCID

Xiaoni Zhong  <http://orcid.org/0000-0002-8035-1841>

References

- [1] Han MJ. Analysis of the situation of the AIDS epidemic and prospects for prevention and treatment in China. *Chin J AIDS STD*. 2023;29(03):247–250.
- [2] Zhao T, Chen G, Sun C, et al. The epidemic of HIV and syphilis and the correlation with substance abuse among men who have sex with men in China: a systematic review and meta-analysis. *Front Public Health*. 2023;11:1082637. doi: [10.3389/fpubh.2023.1082637](https://doi.org/10.3389/fpubh.2023.1082637).
- [3] Zhang W, Xu J-J, Zou H, et al. HIV incidence and associated risk factors in men who have sex with men in Mainland China: an updated systematic review and meta-analysis. *Sex Health*. 2016;13(4):373. doi: [10.1071/SH16001](https://doi.org/10.1071/SH16001).
- [4] Zhao Y, Han M, Ma Y, et al. Progress towards the 90-90-90 targets for controlling HIV – China, 2018. *China CDC Wkly*. 2019;1(1):5–7.
- [5] UNAIDS. UNAIDS data 2023; 2023; Available from: https://www.unaids.org/en/resources/documents/2023/2023_unaids_data.
- [6] 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. doi: [10.2196/13475](https://doi.org/10.2196/13475).
- [7] WHO. Consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection: recommendations for a public health approach. 2nd ed.; 2016. [cited 2024; Available from: <https://www.who.int/publications/i/item/9789241549684>.
- [8] 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. *Lancet HIV*. 2019;6(10):e696–e704. doi: [10.1016/S2352-3018\(19\)30139-0](https://doi.org/10.1016/S2352-3018(19)30139-0).
- [9] Riddell J, Amico KR, Mayer KH. HIV preexposure prophylaxis: a review. *JAMA*. 2018;319(12):1261–1268. doi: [10.1001/jama.2018.1917](https://doi.org/10.1001/jama.2018.1917).
- [10] Sun Z, Gu Q, Dai Y, et al. Increasing awareness of HIV pre-exposure prophylaxis (PrEP) and willingness to use HIV PrEP among men who have sex with men: a systematic review and meta-analysis of global data. *J Int AIDS Soc*. 2022;25(3):e25883. doi: [10.1002/jia2.25883](https://doi.org/10.1002/jia2.25883).
- [11] Marcus JL, Hurley LB, Krakower DS, et al. Use of electronic health record data and machine learning to identify candidates for HIV pre-exposure prophylaxis: a modelling study. *Lancet HIV*. 2019;6(10):e688–e695. doi: [10.1016/S2352-3018\(19\)30137-7](https://doi.org/10.1016/S2352-3018(19)30137-7).
- [12] Lin TC, Gianella S, Tenenbaum T, et al. A simple symptom score for acute human immunodeficiency virus infection in a San Diego community-based screening program. *Clin Infect Dis*. 2018;67(1):105–111. doi: [10.1093/cid/cix1130](https://doi.org/10.1093/cid/cix1130).
- [13] Tordoff DM, Barbee LA, Khosropour CM, et al. 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. *J Acquir Immune Defic Syndr*. 2020;85(3):263–271. doi: [10.1097/QAI.0000000000002438](https://doi.org/10.1097/QAI.0000000000002438).
- [14] Brandao-de-Resende C, Melo M, Lee E, et al. A machine learning system to optimise triage in an adult ophthalmic emergency department: a model development and

- validation study. *EClinicalMedicine*. 2023;66:102331. doi: [10.1016/j.eclinm.2023.102331](https://doi.org/10.1016/j.eclinm.2023.102331).
- [15] 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. doi: [10.1016/j.jinf.2020.11.007](https://doi.org/10.1016/j.jinf.2020.11.007).
- [16] Duthe J-C, Bouzille G, Sylvestre E, et al. How to identify potential candidates for HIV pre-exposure prophylaxis: an AI algorithm reusing real-world hospital data. *Stud Health Technol Inform*. 2021;281:714–718. doi: [10.3233/SHTI210265](https://doi.org/10.3233/SHTI210265).
- [17] Chingombe I, Dzinamarira T, Cuadros D, et al. Predicting HIV status among men who have sex with men in Bulawayo & Harare, Zimbabwe using bio-behavioural data, recurrent neural networks, and machine learning techniques. *Trop Med Infect Dis*. 2022;7(9):231. doi: [10.3390/tropicalmed7090231](https://doi.org/10.3390/tropicalmed7090231).
- [18] 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. doi: [10.3389/fpubh.2022.967681](https://doi.org/10.3389/fpubh.2022.967681).
- [19] Kim, S. Cross-sectional and longitudinal studies. In: Gu D, Dupre ME, editors. *Encyclopedia of gerontology and population aging*. Cham: Springer International Publishing; 2021. p. 1251–1255.
- [20] Wu D, Tao H, Dai J, et al. Study on pre-exposure prophylaxis regimens among men who have sex with men: a prospective cohort study. *Int J Environ Res Public Health*. 2019;16(24):4996. doi: [10.3390/ijerph16244996](https://doi.org/10.3390/ijerph16244996).
- [21] 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):e039557. doi: [10.1136/bmjopen-2020-039557](https://doi.org/10.1136/bmjopen-2020-039557).
- [22] Singh RK, Patra S. What factors are responsible for higher prevalence of HIV infection among urban women than rural women in Tanzania? *Ethiop J Health Sci*. 2015;25(4):321–328. doi: [10.4314/ejhs.v25i4.5](https://doi.org/10.4314/ejhs.v25i4.5).
- [23] Mutai CK, McSharry PE, Ngaruye I, et al. Use of unsupervised machine learning to characterise HIV predictors in sub-Saharan Africa. *BMC Infect Dis*. 2023;23(1):482. doi: [10.1186/s12879-023-08467-7](https://doi.org/10.1186/s12879-023-08467-7).
- [24] Hajizadeh M, Sia D, Heymann SJ, et al. Socioeconomic inequalities in HIV/AIDS prevalence in sub-Saharan African countries: evidence from the Demographic Health Surveys. *Int J Equity Health*. 2014;13(1):18. doi: [10.1186/1475-9276-13-18](https://doi.org/10.1186/1475-9276-13-18).
- [25] Birri Makota RB, Musenge E. Predicting HIV infection in the decade (2005–2015) pre-COVID-19 in Zimbabwe: a supervised classification-based machine learning approach. *PLOS Digit Health*. 2023;2(6):e0000260. doi: [10.1371/journal.pdig.0000260](https://doi.org/10.1371/journal.pdig.0000260).
- [26] He L, Pan X, Yang J, et al. HIV risk behavior and HIV testing among rural and urban men who have sex with men in Zhejiang Province, China: a respondent-driven sampling study. *PLoS One*. 2020;15(4):e0231026. doi: [10.1371/journal.pone.0231026](https://doi.org/10.1371/journal.pone.0231026).
- [27] Feng Y, et al. Meta-analysis of HIV infection incidence and risk factors among men who have sex with men in China. *Chinese Journal of Epidemiology*. 2015;36(7):752–758.
- [28] Cornelisse VJ, Chow EPF, Latimer RL, et al. Getting to the bottom of it: sexual positioning and stage of syphilis at diagnosis, and implications for syphilis screening. *Clin Infect Dis*. 2020;71(2):318–322. doi: [10.1093/cid/ciz802](https://doi.org/10.1093/cid/ciz802).
- [29] Meng X, Zou H, Fan S, et al. Relative risk for HIV infection among men who have sex with men engaging in different roles in anal sex: a systematic review and meta-analysis on global data. *AIDS Behav*. 2015;19(5):882–889. doi: [10.1007/s10461-014-0921-x](https://doi.org/10.1007/s10461-014-0921-x).
- [30] Vujkovic-Cvijin I, Sortino O, Verheij E, et al. HIV-associated gut dysbiosis is independent of sexual practice and correlates with noncommunicable diseases. *Nat Commun*. 2020;11(1):2448. doi: [10.1038/s41467-020-16222-8](https://doi.org/10.1038/s41467-020-16222-8).
- [31] Li SX, Sen S, Schneider JM, et al. Gut microbiota from high-risk men who have sex with men drive immune activation in gnotobiotic mice and in vitro HIV infection. *PLoS Pathog*. 2019;15(4):e1007611. doi: [10.1371/journal.ppat.1007611](https://doi.org/10.1371/journal.ppat.1007611).
- [32] Ambrosioni J, Petit E, Liegeon G, et al. Primary HIV-1 infection in users of pre-exposure prophylaxis. *Lancet HIV*. 2021;8(3):e166–e174. doi: [10.1016/S2352-3018\(20\)30271-X](https://doi.org/10.1016/S2352-3018(20)30271-X).
- [33] Lin B, Liu J, He W, et al. Effect of a reminder system on pre-exposure prophylaxis adherence in men who have sex with men: prospective cohort study based on WeChat intervention. *J Med Internet Res*. 2022;24(8):e37936. doi: [10.2196/37936](https://doi.org/10.2196/37936).