Proceedings of ICBioMed 2025 Symposium: AI for Healthcare: Advanced Medical Data Analytics and Smart Rehabilitation
DOI: 10.54254/2753-8818/2025.AU25795

Risk Prediction and Intervention Modeling in the HIV Epidemic

Tianyou Xie

College of Literature and Science, University of Wisconsin, Madison, USA
qianyangllgz@gmail.com

Abstract. HIV epidemic is a persistent global health challenge, there are needs in effective tools which can provide assistance in risk prediction and intervention planning. Recent research has applied multiple mathematical models such as logistic regression, random forest, ODE with optimal control theory, and multi-state Markov models to tackle different aspects of the HIV disease. Nevertheless, few studies provide a general comparison across these modeling frameworks. The paper focuses on analyzing four HIV modeling methods: data requirements, interpretability, application areas, and strengths and limitation. The findings indicate that logistic regression and random forest models can support individual prediction, meanwhile ODE and Markov models offer strategic ideas in intervention, treatment. The study highlights the importance of choosing reasonable models based on the contexts and purposes. It further supports the future work to integrate diverse data sources, hybrid modeling frameworks, and realistic simulation of population behavior. The further research could address problems such as model generalizability, real-time adaptation, which are crucial to HIV modeling due to its complex settings.

Keywords: HIV risk prediction, Statistical modeling, Logistic regression, Random Forest, Optimal control theory

1. Introduction

Human immunodeficiency virus (HIV) remains a major public health concern in the world [1]. The virus can progress into a serious syndrome (AIDS) if medication assistance is not sought on time [2]. The virus causes deficiency in human immune system, leading to a more vulnerable human body [3]. Since its first appearance in the 1980s in the United States, the number of people living with HIV has continued to grow [3]. In 2020, there were around 38 million people living with HIV worldwide [4]. Although via the medical progress, there have been efforts to apply antiretroviral therapy (ART) which is able to suppress HIV into a more manageable chronic disease, but it is still not curative [5]. In recent years, increasing concentration has been focused on the development of statistical modeling techniques for a better understanding, prediction, and controlling the spread of HIV.

Among statistical modeling in HIV research, there could be two general categories in this paper: risk prediction models, which help identify individuals or groups at risk of HIV infection and guide interventions, and there is control models which evaluate intervention strategies based on a public

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2025.AU25795

health level. There are two types of statistical models discussed in the risk prediction model in the paper: logistic regression and random forest machine learning model. On the other hand, ODE models and Markov models will be discussed in the control model part. In the domain of risk prediction models, research has developed from conventional statistical models to more complex machine learning models. But logistic regression model still widely applies due to its simplicity and interpretability. The model has been used in create effective risk prediction models, such as for the study of Chinese MSM that successfully predict HIV infection risk and achieve a high AUC of 0.832 on testing data [4]. While the model results present decent predictive ability, logistic regression may be restricted by complex and nonlinear data. In order to address this, machine learning approach like random forest have gained favor. RF as a crucial component in the the "MySTRisk' web-based tool, is able to accurately predict HIV risk with consistency, scoring an AUC of 0.78 in risk prediction of HIV [1]. In contrast, control models evaluate and optimize public health intervention. One method is using ordinary differential equations (ODE) to simulate HIV transmission trend. For instance, a study on simulating the co-infection of HIV and malaria applied ODEs and optimal control theory in order to indicate the most effective strategy for minimizing number of infection while keeping intervention costs [6]. Alternatively, another method utilizing multi-state Markov models dynamically model the HIV care trajectories and identify possible ways to improve long-term outcomes for people living with HIV [7]. In this paper, the intent is to systematically review and compare two types of models in HIV research: risk prediction models (logistic regression and random forest) and control models (ODE and Markov models). By summing each general development, application, performance, and limitations, the review attempts to provide a general and basic understanding of current modeling practices in HIV field.

2. Logistic regression model

2.1. Basic concept

Logistic regression is mainly a statistical method used to predict a binary outcome. It is a foundational method in predicting HIV infection risk, and even widely adapted in biomedical research [8]. The method analyzes on the relationship between multiple independent variables (risk factors) and a single dependent variable (HIV status). By applying this method, logistic regression can map the relationship between predictor variables and log odds of HIV infection, generating a score between 0 and 1 to indicate the risk of everyone. And the result can help clinicians to identify patients who can get help from preventive interventions like the pre-exposure prohaylaxis (PrEP). There are many predictor variables like demographics: age, gender, and race, and these can be important factors to consider in the model. Another important factor is the behavioral factors like practices of intercourses, these can strongly associate with increased HIV risk. Finally, the history of sexually transmitted infections (STIs) will also be another powerful predictor variable in risk prediction of HIV [8].

2.2. Data requirement

For the data requirements, logistic regression requires diverse and complete datasets accumulated from various sources. For example, the candidate predictors can range from a few as 4 to 18, depending on the data source [8]. For example, electronic health records (EHR) can be a primary source of data providing key clinical information. The data includes diagnoses of STIs like syphilis, gonorrhea, and these can be strong predictors; prescription drug history like the medications used to

DOI: 10.54254/2753-8818/2025.AU25795

treat STIs; and moreover, laboratory results like previous recorded HIV screening tests. Besides from EHR, using surveys and questionnaires are a method for collecting necessary data on risk behaviors for the model. Some common data information for instance, includes sexual practices, the number of sexual partners, ways of intercourse, and having an HIV-positive partner. Symptoms like fever and weight loss can also be a part of data information. Last, substance use like methamphetamine and popper use.

2.3. Advantage and limitation

The widely use of logistic regression shows its advantages compared to other models in providing a clear framework for risk assessment. Logistic regression is one of the primary model used to create HIV prediction tools that help identify people at risk of exposing to HIV, enabling preventive interventions like PrEP. The model also provides also offer clear and highly interpretable outcomes like point values or odds ratios for each risk factors, which assists clinicians in analyzing the risk. Via flagging people exposing to high risk of HIV, the model can support early detection and facilitate timely access to care. While there are also key limitations for logistic regression. Social determinants of health such as socioeconomic status or education level are often excluded while they are important contextual factors influencing HIV risk. Also, most models focus on men who have sex with men (MSM), while the validation of cisgender women and transgender individuals is limited, reducing their generalizability. More importantly, many studies presented strong bias due to poor handling of missing data and lack of external validation. So its applicability in broader scale is questionable. And while the models can identify associations, the causation is not assured, then predictors cannot assumed to be intervention targets without further research.

2.4. Real world application

A typical practice of apply logistic regression is demonstrated in a research that focuses on developing a predictive model of HIV risk prediction among MSM in China. The principle of the study is to explore beyond just behavioral risks and examine the impact of psychosocial variables. In order to achieve the objective, the research conducted a cross-sectional study, collecting information from anonymous, face-to-face 547 MSM in Shanghai. The collected data included sociodemographic data, behavioral patterns, and responses to validated psychosocial scales. The scales measured constructs like involuntary subordination (IS), sexual compulsivity, and perceived social support, giving a comprehensive insight into psychological vulnerability. For the calculation method, the researchers applied LASSO regression, a machine learning technique to reduce data dimension and capture the most crucial predicators from the list of 15 candidate variables. The four selected variables selected by LASSO will then use to build the final prediction model by multivariate logistic regression analysis. For validation, the dataset is separated into training set (75%) and testing set (25%). The model performance will be assessed its consistency and generalizability by a series of metrics: Receiver Operating Characteristic (ROC) curves, calibration plots, and the Population Stability Index (PSI). A clear feature of the model is its identification of the psychosocial variable involuntary subordination (IS) as an independent and key risk factor for HIV infection. There are also other important behavioral risks: unprotected anal intercourse (UAI), multiple sexual partners (MSP), and drug use before sex. The logistic regression analysis indicates the odd ratios with 2.8 for MSP, 4.6 for drug use before sex. The model did present a decent predictive capability with an AUC of 0.832 in its testing set [4]. In order to strengthen the clinical

DOI: 10.54254/2753-8818/2025.AU25795

applicability, the model was presented as a nomogram which can help healthcare providers to better visualize and estimate individuals' risk of HIV infection by format.

3. Random forest model

3.1. Basic concept

Random Forest is an ensemble learning algorithm that is able to create a multitude of decision trees and aggregates the outputs to achieve a high degree of predictive accuracy [1]. Each decision tree is built based on a random subset of the data and features, then makes predictions by splitting input spaces based on the learned decision rules. The RF model then gathers the outcomes of all these trees and generate a more accurate and reliable prediction. The RF is powerful and prevalent in machine learning technique for medical fields of study, including the prediction of HIV. Unlike logistic regression, it is a non-linear, ensemble-based classifier that combines multiple models to achieve high level of predictive performance [9].

3.2. Data requirement

An effective RF model for risk prediction of HIV needs a comprehensive dataset with sufficient size and diversity to test multiple decision trees. The model can utilize various types of data from sources, for example, demographic, biological, and behavioral sources. Socio-demographic data includes age, gender, education level, country of birth, etc. In behavioral data, it includes important behaviors like injection (syringe use), and sexual behaviors, number of partners, etc. Lastly, biological and clinical data includes medical records like past medical history, presence of STI, and laboratory test [9].

3.3. Advantage and limitation

RF models show some key advantages in the aspect of medical prediction of HIV. One strength is the high predictive accuracy, RF is powerful in predicting outcomes with relatively high degree of accuracy. One study presents the studied RD model achieved 82% accuracy in predicting HIV status, it has a 10-15% of improvement compared to classifiers like SVM, J48, PART [9]. Another study applied an ensemble model including RF and reached an acceptable or excellent level with AUC of 0.78 in HIV prediction [1]. Another trait is its excellent performance to linear models: RF is a non-linear model which can understand complex relationships in data. During the studies, non-linear machine learning models like RF has a better ability in HIV prediction than logistic regression [1]. Also, the decision trees used in the structure of RF allow it to handle a large size of selected feature in a manageable and efficient way [9]. The use of ensemble learning, which multiple models are combined, enhance the generalizability of the model, and helps tackle problems like overfitting [1]. Despite its power, RF modeling also has some limitations that should be considered. One limitation is the problem about data collection. It is challenging to gather accurate data since self-reported information may not be accurate and individuals may be reluctant to provide accurate information due to concerns in various aspects. Another limitation is the generalizability. The prediction of the model is restricted by specific population that is used to train on. Then the result from a particular region cannot be generalize to other regions or other settings, so further validation is necessary when the model is used in different settings [1]. Another limitation is the challenge of overfitting, the model presents decent performance on training data while perform poorly when

DOI: 10.54254/2753-8818/2025.AU25795

facing new data. So, there needs to be techniques that help tackle the problem like applying cross-validation.

3.4. Real world application

A real-world example is the study on developing a predictive model to classify individuals at high risk of future HIV infection within high-risk groups in Pakistan [9]. The research group applied RF with dataset from electronic health records, combining socio-demographic, behavioral, and biological features to construct the model with highly accurate predictive capability. The development process takes from 47110 people who inject drugs. The data underwent significant preprocessing like removing outliers and noise, filling missing values, and transforming attributes with excessive categories into concise ones. Also, the model is processed by feature selection: a combination of domain expert opinion and automated machine learning approaches such as Select-K-Best, was used to identify the most predictive features. Some key predictors include 'SyringeAcquiredFrom', 'Age', 'FrequencyOfDailyInjecting', and ‘MaritalStatus' [9]. For the training and balancing of the model, the dataset is split into three parts: training, testing, and validation set. Given the major class imbalance, more HIV negative than positive cases, researchers employed Synthetic Minority Oversampling Technique (SMOTE) to balance the data and minimize the risk of based model predictions. By utilizing the technique, the model’s predictive accuracy rises to 82.36%, a major improvement over the 76.5% accuracy observed when the model is trained on the previous dataset. The data result presents high sensitivity that the model shows a strong True Positive Rate for its efficiency at identifying HIV-positive individuals. The case demonstrates that RF can be suitable for predicting HIV risk by using diverse, real-world data. The strong performance highlights the capacity of the model in facing complex behavioral and demographic variables. Through using SMOTE to address the class imbalance, the model can make accurate predictions and further identify meaningful risk factors and giving insights for public health interventions [9].

4. Ordinary differential equation model

4.1. Basic concept

Ordinary differential equation (ODE) models are a type of fundamental mathematical framework used to simulate the dynamics of infectious diseases like HIV. In these models, the population is separated into several compartments like susceptible, infected, treated, and others. The transition between compartments will be controlled by differential equations based on its biological and behavioral parameters. ODE models can be described as deterministic, meaning that their prediction to the average behavior of the epidemic by time without randomness [10]. In order to improve further for public health results, ODE models will often extended with optimal control theory, a mathematical method for determining the optimal intervention strategies to reduce infection, and many other public health concerns or burdens [6]. For HIV, the control variables may include treatment rates, screening efforts, or behavioral interventions like media campaigns. The main idea is to minimize the objective function such as the total number of infected individuals or a cost-effectiveness ratio, subject to the dynamics of the disease model [10]. This process often depends on Pontryagin’s Maximum Principle to derive the necessary conditions for the optimal strategy which then solved with numerical methods just like the forward-backward sweep method [10].

DOI: 10.54254/2753-8818/2025.AU25795

4.2. Data requirement

ODE based optimal control models rely on comprehensive quantitative data in order to simulate disease transmission and evaluate intervention strategies. For data requirements, the models require starting values for the population size for each defined compartment. For instance, a model uses approximated data from 2019, Ethiopia to set up initial values for the compartments such as susceptible people, diagnosed cases, and those receiving treatment [10]. By comparison, a more complex model involves twelve compartments in model structure and thus requires a larger set of initial values [6]. Also, numerous biological, behavioral, and epidemiological rates are needed. These rates include transmission rates, treatment efficacy, natural and ALDS-related morality, screening coverage, and awareness transition via media campaigns. Where empirical data were not available, parameter values were either cited from existing literature or estimated based on similar cases [6]. For the optimal control problem, the models require weight coefficient for the objective function. This coefficient represents the relative cost of different interventions like prevention, screening, and treatment, same as the broader societal cost of each infection. But there is one paper indicate the accurate data for the cost factors can be difficult to obtain. It could be challenging then sometimes researchers decide to rely on reasonable assumptions or numerical simulation [10].

4.3. Advantage and limitation

By the application of ODEs with optimal control, there are some advantages for researching and managing HIV. One benefit is the models are able to provide a systemic format to simulate and compare the effectiveness of different public heath interventions such as prevention, screening, and treatment [10]. And via applying methods like cost-effectiveness analysis, the models can help determine which combination of strategies can offer the best option, the lowest cost makes the greatest impact. This offers valuable insights for policy makers making resource allocation decisions [10]. Another advantage is the ability models to capture complex features, dynamics of an epidemic. For example, by incorporating variable-order fractional differential equations helps capture how past events can affect current disease trends [6]. And another model includes the simulation of social dynamics. By dividing susceptible population into two categories “aware” and “unaware” groups, the model can learn the differences in behavior and risk perception [10]. In terms of theoretical aspect, the use of optimal control theory is a strong tool to understand the principles of handling epidemic. The theory finds the optimal way to tackle and control the disease within real-world limits such as cost [10]. Despite the advantages, the models have some noticeable limitations. First, the performance depends heavily on data quality. In many cases especially about estimation of cost, reliable data can be difficult to obtain. Then, researchers often have no choice but to make approximate assumption [10]. For more complicate model that use multiple parameters, collecting and validating important real-world data can be even more challenging. Second, it is essentially impossible to simulate reality as many factors are left out in modeling. All models are simplification of the reality. For instance, some models choose to exclude mother to child transmission, the effects of other infections coexisting with HIV, or the movement of infected individuals between regions. In one study, the authors note that these aspects could be added in future versions of the model [10]. Last, due to the deterministic nature of ODE models, they only present how disease might behave on average in a population. The model is unable to capture random events or individual differences which can make impacts on disease spread in real world context [10].

DOI: 10.54254/2753-8818/2025.AU25795

4.4. Real world application

For real-world example, research focusing on the application of optimal control theory to address the HIV epidemic in case study of Ethiopia. The goal is to identify the most effective public health strategies in order to reduce new infections and the related economic burden. Researchers applies deterministic ordinary differential equation model that divides the population into socially related compartments like “aware” and “unaware” susceptible individuals. The structure enables the model to reflect the influence of media campaigns and educational efforts on behavior. For the methodology of the study, the model focuses on minimizing a cost function which analyzes both the burden of disease and the spending of intervention strategies. The measures include three specific control directions: preventive education efforts, screening programs, and treatment interventions. The optimal combination of these strategies is first explored via using pontryagin’s Maximum Principle theoretically. For numerical analysis, forward-backward sweep method is utilized to solve the system. The cost-effectiveness is assessed via the Incremental Cost-Effectiveness Ratio (ICER), highlighting a direct economic comparison of the intervention strategies [10]. The model was initialized with estimated data from Ethiopia for 2019. The simulation results displayed that using any combination of the three interventions can derive to a great reduction in number of infected individuals compared to scenarios without intervention. And among the tested strategies, the combination of preventive education and screening appeared to be the most efficient and cost-efficient approach. The research provides a great insight based on real world data and its practical focus. By applying country-specific estimates and giving a clear economic emulation of various intervention options, the model combines complex mathematical analysis and public health policy. As a result, it gives pragmatic suggestions that can be valuable for supporting decision making process by the public health department in Ethiopia [10].

5. Markov model

5.1. Basic concept

The multi-state Markov model is a statistical tool used to analyze HIV patient transition between different antiretroviral therapy (ART) regimens over time. Each treatment combination is defined as a “state.” The central concept of the Markovian assumption is that it assumes that a patient’s future treatment path depends only on their current regimen, not on the past history [7]. The model framework allows researchers to estimate the transition probability, calculating how possible it is for patients to switch from one regimen to another. The model also calculates the average time spent in each regimen before switching to another. In order to track patients’ data, it uses data like CD4 count, TB status, weight, and age to identify risk factors for treatment changes. Overall, the Markov model tracks HIV treatment patterns, identifies unstable regimens, and supports public health planning of drug supply and patient care.

5.2. Data requirement

The model requires detailed longitudinal data for each patient, capturing how patients move between ART regimens over time. The data components include demographics: age, sex, marital status, and educational background are collected; another important element is the clinical profiles which include detailed record like weight, WHO clinical stage, presence of opportunistic infections, TB status, and functional status like working, bedridden; test results from laboratory is another one:

DOI: 10.54254/2753-8818/2025.AU25795

CD4 cell counts and HIV viral load were consistently under record, also along with liver function and Hepatitis B/C serology status; ART treatment history is also included: initial regimen of patients, subsequent treatment changes, and timing of the changes were tracked; lastly, the outcomes like the reasons of regimen change like toxicity, pregnancy, etc. Follow up status will also be included.

5.3. Advantage and limitation

The multi-state model offers some advantages over the other traditional survival analysis in HIV aspect of study. The model provides a comprehensive and complete treatment mapping. It models the full sequence of treatment regimens a patient has over time, not just the first or last. Also, the model is able to capture the relationships between types of regimen changes and allows for analysis of sequential events. The model can also be very flexible, allowing to make accommodations to various types of change like single-drug substitutions, full regimen switches, and reversible transitions, changing back to the previous regimen. Aside from its flexibility, the model can output useful information about clinical and policy decisions. For example, estimating probability of transitioning regimens, predictions about future treatment status of patients, etc. Lastly, the model can use patient's feature (covariates) to analyze risk factors like TB status or baseline CD4 affect the risk of switching treatments. On the other hand, there are constraints primarily related to the core assumptions and data source of the model. The model assumes that only current treatment State affect further transitions, ignoring the full treatment history. This may not be compatible to real-world clinical way of decision-making. There are also concerns about the data quality. For example, the study has to exclude 5824 patients in total from analysis due to missing information in drug, and this could impact the results. Besides, the application of continuous-time, homogenous multi-state Markov model assumes constant transition rates over time. The simplification can potentially fail to simulate how actual regimen changes happen [7].

5.4. Real world application

For the real-world application, the study by Kumsa et al. used national-scale data from Ethiopia, analyzing 33716 cases of HIV patients to understand the durability of first-line ART and explore the facts that lead to patients changing their treatment regimens. The study applied a multi-state Markov model, and the model is built on Markovian assumption, indicating the future treatment transition of a patient will depend only on current regimen, not by past history. And each ART combination is treated as a "state", and treatments are "transition" between the States. The model was used to estimate the transition probability between different ART regimens and then calculate the average length of stay in each state. The calculation method is conducted through R package "msm". The data results show that 14.4% of patients changed their treatment at least once. And the primary reasons were toxicity or other clinical problems. The regimen with drug d4T is most likely to be changed, with a median duration of 24 months compared to 42 months for AZT-based regimens. The key risk factors for treatment modification like positive TB status, opportunistic disease, low body weight, etc. Overall, the characteristic of the model is its capacity to capture multiple treatment pathways, handling reversible transitions, incorporating patient specific risk factors, and further providing clinical insights like transition probability and expected duration in each regimen, etc.

DOI: 10.54254/2753-8818/2025.AU25795

6. Conclusion

The models serve different purposes, but they are complementary. Logistic Regression and Random Forest focus on risk prediction at the individual level, while ODE and Markov Models support broader intervention planning and long-term strategy. In future development of HIV modeling, there is a lot of potential to make these models even more useful. One direction is the comprehensive combination of data types, enabling the picture to be more complete. Another is blending machine learning methods with traditional public health frameworks to get the best use. To make predictions more realistic, future models could simulate how people interact through agent-based or network-based approaches. By incorporating factors like cost and quality of life, this would further boost the practicality of these tools in real-world public health context.

References

[1] Xu, Xianglong, et al. Web-Based Risk Prediction Tool for an Individual's Risk of HIV and Sexually Transmit ted Infections Using Machine Learning Algorithms: Development and External Validation Study. Journal of Medical Internet Research, vol. 24, no. 8, 2022, article e37850.
[2] Cheneke, Kumama Regassa. Optimal Control and Bifurcation Analysis of HIV Model. Computational and Mathematical Methods in Medicine, vol. 2023, no. 1, 2023, article ID 4754426.
[3] Boukary, Ouedraogo, Zorom Malicki, and Gouba Elisée. Mathematical Modeling and Optimal Control of an HIV/AIDS Transmission Model. International Journal of Analysis and Applications, vol. 22, 2024, pp. 234–234.
[4] Dong, Yinqiao, et al. Prediction Model for the Risk of HIV Infection among MSM in China: Validation and Stability. International Journal of Environmental Research and Public Health, vol. 19, no. 2, 2022, article 1010.
[5] Landovitz, Raphael J., Hyman Scott, and Steven G. Deeks. Prevention, Treatment and Cure of HIV Infection. Nature Reviews Microbiology, vol. 21, no. 10, 2023, pp. 657–670.
[6] Sweilam, N. H., et al. Optimal Control for Variable Order Fractional HIV/AIDS and Malaria Mathematical Models with Multi-Time Delay. Alexandria Engineering Journal, vol. 59, no. 5, 2020, pp. 3149–3162.
[7] Kumsa, Tsegaye Hailu, et al. Multi-State Markov Model for Time to Treatment Changes for HIV/AIDS Patients: A Retrospective Cohort National Datasets, Ethiopia. BMC Infectious Diseases, vol. 24, no. 1, 2024, article 627.
[8] Albernas, Adrian, et al. HIV Risk Score and Prediction Model in the United States: A Scoping Review. AIDS and Behavior, 2025, pp. 1–20.
[9] Nisa, Sehar Un, et al. HIV/AIDS Predictive Model Using Random Forest Based on Socio-Demographical, Biological and Behavioral Data. Egyptian Informatics Journal, vol. 24, no. 1, 2023, pp. 107–115.
[10] Ayele, Tigabu Kasia, Emile Franc Doungmo Goufo, and Stella Mugisha. Mathematical Modeling of HIV/AIDS with Optimal Control: A Case Study in Ethiopia. Results in Physics, vol. 26, 2021, article 104263.