Original Paper

# Web-Based Risk Prediction Tool for an Individual's Risk of HIV and Sexually Transmitted Infections Using Machine Learning Algorithms: Development and External Validation Study

Xianglong Xu[1,2,3], PhD; Zhen Yu[2,3,4], MSc; Zongyuan Ge[4], PhD; Eric P F Chow[1,2,5], PhD; Yining Bao[3], MSc; Jason J Ong[1,2,3], PhD; Wei Li[6], PhD; Jinrong Wu[7], MSc; Christopher K Fairley[1,2,3], PhD; Lei Zhang[1,2,3], PhD

[1]Melbourne Sexual Health Centre, Alfred Health, Melbourne, Australia

[2]Central Clinical School, Faculty of Medicine, Nursing and Health Sciences, Monash University, Melbourne, Australia

[3]China Australia Joint Research Center for Infectious Diseases, Xi'an Jiaotong University Health Science Centre, Xi'an, China

[4]Monash e-Research Centre, Faculty of Engineering, Airdoc Research, Nvidia AI Technology Research Centre, Monash University, Melbourne, Australia

[5]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, University of Melbourne, Melbourne, Australia

[6]School of Public Health, Southeast University, Nanjing, China

[7]Research Centre for Data Analytics and Cognition, La Trobe University, Melbourne, Australia

**Corresponding Author:**
Lei Zhang, PhD
Melbourne Sexual Health Centre
Alfred Health
580 Swanston Street
Melbourne, 3053
Australia
Phone: 61 3 9341 6230
Fax: 61 3 9341 6264
Email: lei.zhang1@monash.edu

## Abstract

**Background:** HIV and sexually transmitted infections (STIs) are major global public health concerns. Over 1 million curable STIs occur every day among people aged 15 years to 49 years worldwide. Insufficient testing or screening substantially impedes the elimination of HIV and STI transmission.

**Objective:** The aim of our study was to develop an HIV and STI risk prediction tool using machine learning algorithms.

**Methods:** We used clinic consultations that tested for HIV and STIs at the Melbourne Sexual Health Centre between March 2, 2015, and December 31, 2018, as the development data set (training and testing data set). We also used 2 external validation data sets, including data from 2019 as external "validation data 1" and data from January 2020 and January 2021 as external "validation data 2." We developed 34 machine learning models to assess the risk of acquiring HIV, syphilis, gonorrhea, and chlamydia. We created an online tool to generate an individual's risk of HIV or an STI.

**Results:** The important predictors for HIV and STI risk were gender, age, men who reported having sex with men, number of casual sexual partners, and condom use. Our machine learning–based risk prediction tool, named MySTIRisk, performed at an acceptable or excellent level on testing data sets (area under the curve [AUC] for HIV=0.78; AUC for syphilis=0.84; AUC for gonorrhea=0.78; AUC for chlamydia=0.70) and had stable performance on both external validation data from 2019 (AUC for HIV=0.79; AUC for syphilis=0.85; AUC for gonorrhea=0.81; AUC for chlamydia=0.69) and data from 2020-2021 (AUC for HIV=0.71; AUC for syphilis=0.84; AUC for gonorrhea=0.79; AUC for chlamydia=0.69).

**Conclusions:** Our web-based risk prediction tool could accurately predict the risk of HIV and STIs for clinic attendees using simple self-reported questions. MySTIRisk could serve as an HIV and STI screening tool on clinic websites or digital health platforms to encourage individuals at risk of HIV or an STI to be tested or start HIV pre-exposure prophylaxis. The public can use this tool to assess their risk and then decide if they would attend a clinic for testing. Clinicians or public health workers can use this tool to identify high-risk individuals for further interventions.

***(J Med Internet Res 2022;24(8):e37850)*** doi: 10.2196/37850

**KEYWORDS**

HIV; sexually transmitted infections; syphilis; gonorrhea; chlamydia; sexual health; sexual transmission; sexually transmitted; prediction; web-based; risk assessment; machine learning; model; algorithm; predictive; risk; development; validation

## Introduction

HIV and sexually transmitted infections (STIs) are major global public health concerns [1,2]. The World Health Organization (WHO) estimated that over 1 million curable STIs occur every day among people aged 15 years to 49 years worldwide [3]. An estimated 29,090 people have been infected with HIV in Australia as of the end of 2020, with an HIV prevalence rate of 0.14% among people over 15 years old [4]. The estimated undiagnosed HIV infection rate among all people living with HIV in Australia was about 9% in 2020 [4]. Gonorrhea, chlamydia, and early syphilis can be asymptomatic. There were large increases in STIs in Australia between 2013 and 2017. The notification rates of STIs for chlamydia increased from 302.2/100,000 to 394.9/100,000 in men and from 430.7/100,000 to 441.8/100,000 in women, gonorrhea increased from 91.1/100,000 to 174.2/100,000 in men and from 39.6/100,000 to 61.8/100,000 in women, and syphilis increased from 12.3/100,000 to 31.1/100,000 in men and from 1.4/100,000 to 5.5/100,000 in women [5]. In addition, STIs account for a large health and economic burden in limited-income countries [6].

In response to the rising rates of STIs, the WHO proposed the "Global health sector strategy on Sexually Transmitted Infections, 2016-2021," which aimed to end STI epidemics as public health concerns by 2030. This specifically includes a 90% reduction in gonorrhea incidence globally from the 2018 global baseline and achieving a rate of ≤50 congenital syphilis cases per 100,000 live births in 80% of countries [7]. In 2018, the United Nations proposed "The 2030 Agenda for Sustainable Development," which called for an end to the AIDS epidemic by 2030 [8]. Key to the effective control of these infections is accessible health care and, in particular, frequent testing because treated infections rapidly become noninfectious [2]. Screening of asymptomatic individuals is important for diagnosis, treatment, prevention, and control of HIV and STIs [9]. Barriers to testing include misjudgment of an individual's HIV and STI risk, limited availability of testing, and high cost of testing [10]. Therefore, developing innovative tools will help individuals accurately judge their risk of HIV and STIs, hence increasing screening in high-risk individuals.

An easily accessible and user-friendly tool that accurately identifies an individual's risk of infection could form part of a web-based risk prediction program and play a role in risk prediction and personalized risk management [11]. Providing the public with risk prediction tools to assist them in estimating the risk of HIV and STIs may encourage those individuals at high risk to test more regularly. A previous study showed that increased risk perceptions were associated with greater STI health care use (eg, testing) [12]. An HIV and STI risk prediction tool may increase risk perceptions and motivate individuals to seek HIV and STI testing or treatment. Another review study suggested that web-based screening apps can effectively increase the uptake of health screening in the general population [13]. However, there is no web-based tool we could identify that provides users with an individual's current quantitative risk of HIV and STIs (gonorrhea, chlamydia, and syphilis) using self-reported questions.

A number of mathematical techniques can be used to generate an individual's risk of HIV and STIs. Logistic regression has limitations in predictive analysis that uses complex and big data. Logistic regression methods require strong assumptions and cannot easily deal with nonlinear relationships, interactions, and multicollinearity [14,15]. In contrast, nonlinear machine learning approaches can address these limitations and have numerous advantages (eg, capturing nonlinear relationships and interactions) in predictive analysis using big data [16]. Machine learning also can identify rare health outcomes with high accuracy [17]. Ensemble learning is also a machine learning approach that combines multiple machine learning algorithms to improve the model's performance [18].

Despite the advantages of machine learning approaches, there is an absence of individual risk prediction tools for HIV and STI risk using machine learning models. Existing studies using machine learning algorithms to predict HIV and STI acquisition mainly focus on HIV [19-30], and few focus on STIs [19,21,31]. Of these HIV prediction studies, 4 studies focused on high-risk individuals (such as men who have sex with men [MSM] [20,21,24,29]), 2 studies used imaging or clinical text data [22,30], 4 studies used more than 40 predictors [23,26-28], and 2 studies assessed future but not current HIV prediction [19,25]. Of the STI prediction studies, 1 study was conducted with MSM [21], and the other 2 studies focused on future STI prediction [19,31]. These studies also found that nonlinear machine learning models (eg, random forest [RF], gradient boosting machine [GBM], and neural networks) performed better than logistic regression in HIV and STI prediction [19,21,24,31]. These published studies highlight a lack of machine learning models that use simple self-reported questions, predict both the risk of HIV and STIs, and can be used by both men and women. Therefore, to address the current lack of studies that predict the risk of both STIs and HIV, particularly in lower-risk heterosexual individuals, we aimed to use a stacking ensemble learning framework and self-reported questions to predict HIV and 3 common STIs (gonorrhea, chlamydia, and syphilis) in both men and women and a subsequent web-based HIV and STI risk prediction tool.

## Methods

### Study Population

The Melbourne Sexual Health Centre (MSHC) is the largest public sexual health center in Victoria, Australia and offers free HIV and STI testing and management [32]. At the MSHC, individuals' demographic information and sexual practices are recorded using a computer-assisted self-interview (CASI) at each visit, at least 3 months apart [33]. We used clinical consultation data from the electronic health record (EHR) at MSHC to develop and validate the risk prediction model. We

chose March 2, 2015, as the commencement date because this date was when we adopted a new testing platform for gonorrhea and chlamydia (Aptima Combo, Hologic, Marlborough, MA). Our study data included men and women aged 18 years and older who was tested for HIV or an STI at the MSHC between March 2, 2015, and January 29, 2021. We excluded transgender people and individuals aged younger than 18 years.

We used data from March 2, 2015, to December 31, 2018, as the development data set (training and testing data set). The HIV study data set included training and testing data (88,642 consultations). The syphilis, gonorrhea, and chlamydia study data sets had 92,291, 97,473, and 115,845 consultations, respectively.

We used temporal validation as the external validation to evaluate the transportability and generalizability of our risk prediction models. The COVID-19 epidemic may potentially have changed the demographics of those who attend the MSHC [34]. We performed 2 temporal validations to validate our models further and reduce the possible bias caused by COVID-19. The 2 external validation data sets included data from 2019 as external "validation data 1" and data from January 2020 and January 2021 as external "validation data 2." For HIV, the first external validation data set contained 28,875 consultations, and the second external validation data set contained 18,052 consultations. For syphilis, the first external validation data set contained 30,302 consultations, and the second external validation data set contained 19,150 consultations. For gonorrhea, the first external validation data set contained 36,805 consultations, and the second external validation data set contained 22,886 consultations. For chlamydia, the first external validation data set contained 36,393 consultations, and the second external validation data set contained 22,615 consultations.

## Ethical Approval

Ethical approval was granted by the Alfred Hospital Ethics Committee, Melbourne, Australia (project number: 124/18). All methods were carried out following relevant guidelines and regulations of the Alfred Hospital Ethics Committee. As this was a retrospective study involving minimal risk to the privacy of the study participants, the need for informed consent was waived by the Alfred Hospital Ethics Committee. All identifying details of the study participants were removed before any computational analysis.

## Predictors

The data fields we selected for inclusion as predictors were informed by literature review, expert opinion, and prior work [21]. The predictors were self-reported questions from the EHR, including demographics, sexual practices, STI history, and STI contact history (summarized in Table 1 and Tables S1-S5 in Multimedia Appendix 1).

**Table 1.** Characteristics of clinic consultations in the training and testing data set.

| Variables | HIV (n=88,642 consultations) | Syphilis (n=92,291 consultations) | Gonorrhea (n=97,473 consultations) | Chlamydia (n=115,845 consultations) |
|---|---|---|---|---|
| **Gender, n (%)** | | | | |
| Female | 26,651 (30.1) | 27,134 (29.4) | 31,282 (32.1) | 38,548 (33.3) |
| Male | 61,991 (69.9) | 65,157 (70.6) | 66,191 (67.9) | 77,297 (66.7) |
| Age at consultation (years), median (IQR) | 29.0 (24.0-35.0) | 29.0 (25.0-35.0) | 28.0 (24.0-35.0) | 28.0 (24.0-34.0) |
| **Country of birth, n (%)** | | | | |
| Australia | 39,148 (44.2) | 40,990 (44.4) | 43,881 (45.0) | 51,162 (44.2) |
| Overseas | 46,003 (51.9) | 47,670 (51.7) | 49,835 (51.1) | 60,272 (52.0) |
| Missing | 3491 (3.9) | 3631 (3.9) | 3757 (3.9) | 4411 (3.8) |
| **STI[a] symptoms, n (%)** | | | | |
| No | 56,175 (63.4) | 57,413 (62.2) | 54,595 (56.0) | 68,584 (59.2) |
| Yes | 25,067 (28.3) | 27,150 (29.4) | 34,751 (35.7) | 38,930 (33.6) |
| Missing | 7383 (8.3) | 7728 (8.4) | 8127 (8.3) | 8331 (7.2) |
| **Men who have sex with men, n (%)** | | | | |
| Not applicable (female) | 26,651 (30.1) | 27,134 (29.4) | 31,282 (32.1) | 38,548 (33.3) |
| No | 16,508 (18.6) | 17,089 (18.5) | 15,245 (15.6) | 26,975 (23.3) |
| Yes | 45,483 (51.3) | 48,068 (52.1) | 50,946 (52.3) | 50,322 (43.4) |

[a]STI: sexually transmitted infection.

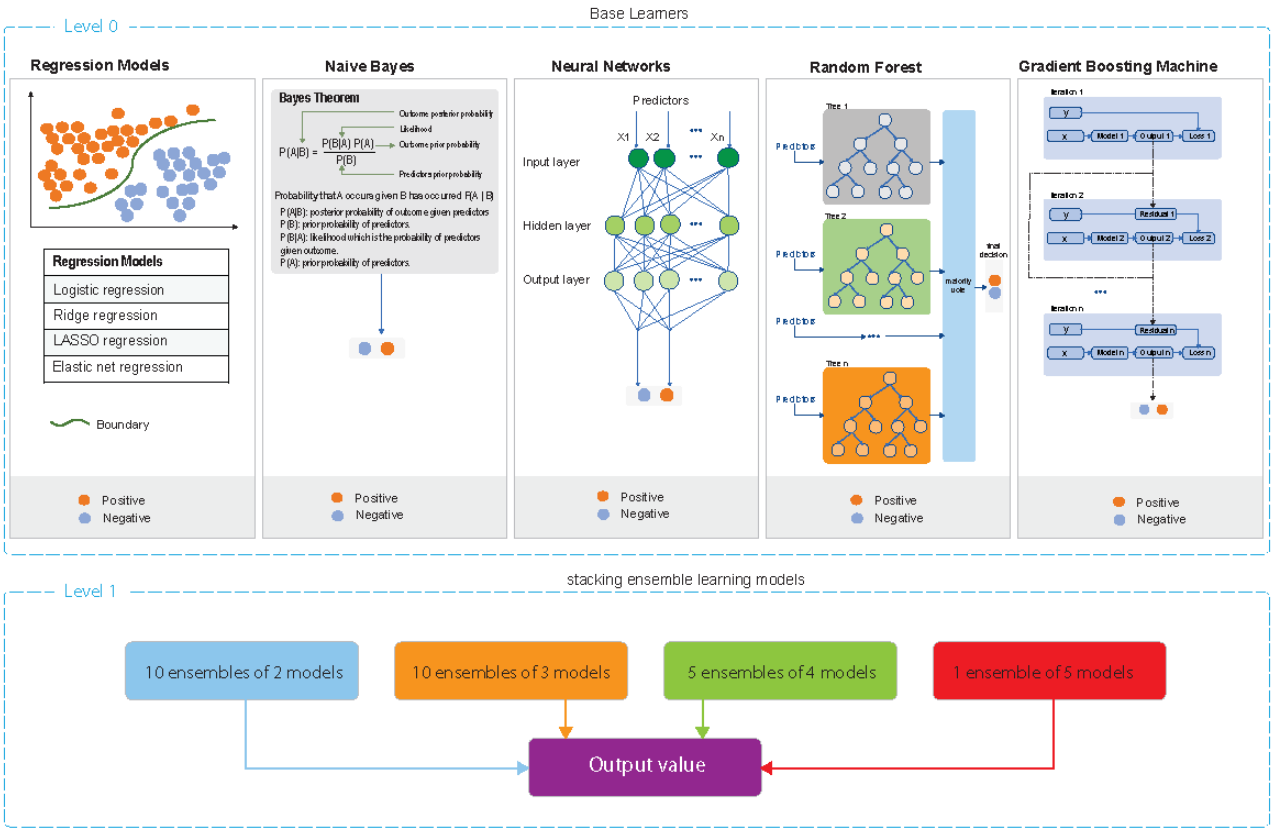
## Measurement of Outcomes

HIV infection was defined as a new diagnosis of HIV based on serology. Syphilis infection was defined as a new diagnosis of early syphilis (primary, secondary, and early latent [<2 years]) using a blood test or nucleic amplification test (NAAT). Gonorrhea infection was defined as a new diagnosis of gonorrhea using culture or NAAT at any anatomical site. In the clinic, gonorrhea testing initially occurs with NAAT, and culture is mostly used after a positive NAAT. Chlamydia infection was defined as a new diagnosis using NAAT at any anatomical site. Our previous publications report the diagnostic methods in detail [19,21].

## Risk Assessment Model Development

We developed 34 machine learning models to assess the risk of acquiring HIV, syphilis, gonorrhea, and chlamydia (details in Figure 1).

**Figure 1.** Development of machine learning algorithms. The architecture of the gradient boosting machine was adapted from Feng et al [35]. LASSO: least absolute shrinkage and selection operator.




### *Base Learner*

Logistic regression has been widely used to predict the risk of incident STIs and HIV [36,37]. GBM uses boosting based on decision trees by adjusting the parameters to minimize a loss function and determine the optimal point with the smallest error [38]. RF comprises an ensemble of decision trees using bootstrap aggregation and randomization of predictors to achieve a high degree of predictive accuracy [39]. Naive Bayes (NB) is simple, has high accuracy and speed in large databases, and has been widely used for disease classification [40]. Deep learning (DL) has effectively solved many medical problems and utilizes a hierarchical level of an artificial neural network to perform the classification process [41].

We first established 4 regression models, including logistic regression, ridge regression, least absolute shrinkage and selection operator (LASSO) regression, and elastic net regression (ENR). Based on the preliminary results of the 4 regression analyses, we found that ENR was better than the other 3 regression analyses (details in Multimedia Appendix 1). Considering our previous machine learning study among MSM [21] and the advantages of NB (eg, high accuracy and speed in large databases), we developed 5 base models, including the aforementioned ENR, NB, DL (neural networks), RF, and GBM.

### *Stacking Ensemble Learning*

Stacking ensemble learning is an ensemble learning method that trains a new model based on the combined predictions of 2 (or more) previous machine learning models. Stacking ensemble learning often performs better than individual machine learning techniques [42]. We systematically established 26 ensemble learning models by combining the aforementioned 5 base models to improve the performance of predicting HIV and STIs. Details are in Multimedia Appendix 1 (summarized in Table S6).

### *Machine Learning Training Techniques*

Our models used a one-hot encoding scheme for data classification. We did not impute missing data but created a binary feature vector indicating missing values. The data were considered "imbalanced" given that each of the 4 infections was <10%. Imbalanced data may cause either overfitted or underperformed predictive results [43]. We used 5 x 10 (5 outer folds, 10 inner folds) nested cross-validation (CV) for model selection and training [21,44]. The outer 5-fold CV was used to address the selection bias caused by using a single data set. The inner 10-fold CV was used on the training data set to perform the hyperparameter tuning of machine learning models. We used the area under the curve (AUC) to select the best model. An AUC of 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and >0.9 is considered outstanding [45]. Machine learning models were built using the *h2o* package (version 3.32.1.2) in R software (3.6.1 and R studio 1.2.5019).

### Estimating the Risk of HIV and STIs

Our machine learning models predicted the probability of HIV or an STI with a normalized distribution between values 0 and 1. The model-predicted probability was calibrated to the actual prevalence level of HIV and STIs. We used a logistic function to provide a fitting curve for each model-predicted probability and infection prevalence. We regarded the estimated infection prevalence as the "calibrated risk" of infection and presented it in the risk report. We used MATLAB R2019a (MathWorks, Natick, MA) to calibrate the model-predicted probability to the actual prevalence level. The method is described in detail in our previous paper [19]. We classified the calibrated risk of HIV or an STI into 3 risk levels: HIV (low, <0.1%; medium, 0.1%-1.0%; and high, >1.0%), syphilis (low, <0.2%; medium, 0.2%-5.0%; and high, >5.0%), gonorrhea (low, <0.1%; medium, 0.1%-1.0%; and high, ≥1.0%), and chlamydia (low, <2.0%; medium, 2.0%-15.0%; and high, >15.0%).

### Establishment of a HIV and STI Risk Prediction Tool

To investigate the effect of predictors, we used the best base machine learning model to calculate the variable importance for HIV, syphilis, gonorrhea, and chlamydia infection. We identified and selected predictors that accounted for more than 80.0% of the overall model performance for each infection. We retrained, retested, and revalidated the best performing model based on these predictors. We compared the AUC, sensitivity, and specificity to re-evaluate the model performance with the shortlisted predictors. We also used the AUC to evaluate the change in performance in the best machine learning model before and after predictor shortlisting (details in Multimedia Appendix 1). We formed a new questionnaire by pooling the important predictors to develop a web-based tool for HIV and STI risk prediction.

## Results

### Characteristics of the Study Data

Our training and testing data included 216 (0.2% of 88,642 consultations) HIV infections, 787 (1.9% of 92,291 consultations) syphilis infections, 7581 (7.8% of 97,473 consultations) gonorrhea infections, and 10,217 (8.8% of 115,845 consultations) chlamydia infections. The proportion of each of the 4 infection data sets that was men was between 66.7% (77,297/115,845) and 70.6% (65,157/92,291). Further details are provided in Table 1 and Table S1 in Multimedia Appendix 1. The characteristics of the external validation data are shown in Tables S2-S5 in Multimedia Appendix 1.

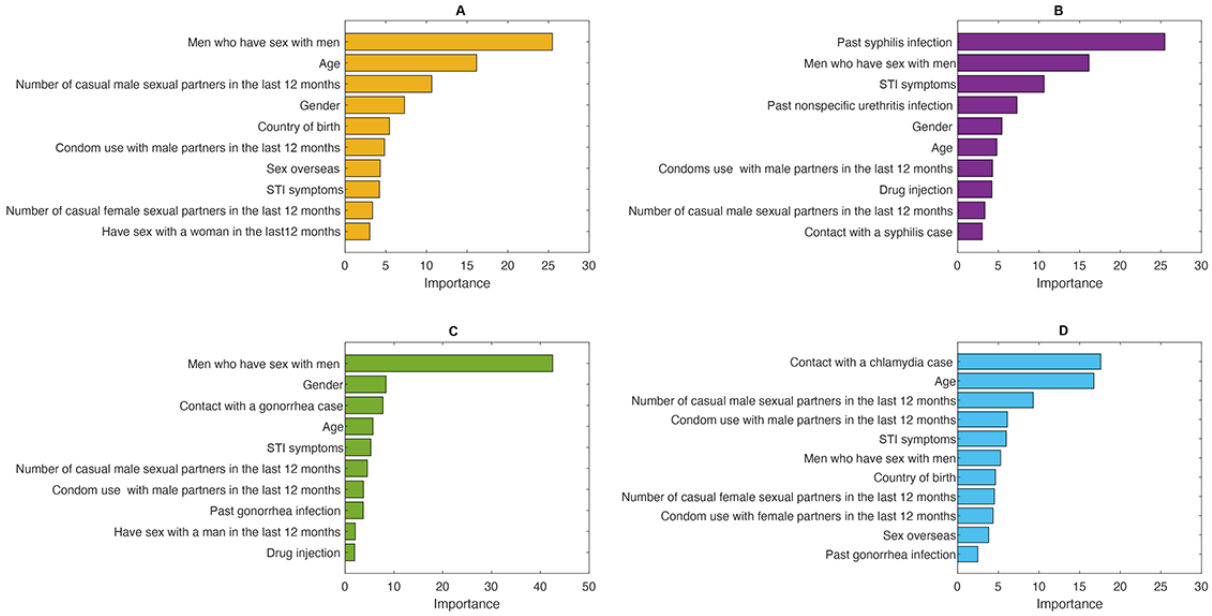
### Selecting the Best ML Model for the HIV and STI Risk Prediction Tool

Our results demonstrated that the ensemble learning models performed better than individual machine learning models. Of all 34 models, our best model (ensemble ENR+GBM+RF) provided acceptable or excellent performance on testing data for predicting HIV (AUC=0.78), syphilis (AUC=0.84), gonorrhea (AUC=0.78), and chlamydia (AUC=0.70; Figures S1-S3 in Multimedia Appendix 1). Details on the testing data analysis are provided in Tables S7-S22 in Multimedia Appendix 1. Our external validation results showed very comparable AUCs (0.69-0.85) to the testing data analysis. Details on the external validation analysis are provided in Tables S7-S22 in Multimedia Appendix 1.

### Selecting the Most Important Predictors for the HIV and STI Risk Prediction Tool

The top 10 predictors for each of the 4 infections accounted for >80.0% of the overall HIV and STI model performance. These predictors included gender, presence of STI symptoms, MSM, age, country of birth, having sex with a man in the last 12 months, the number of casual male sexual partners in the last 12 months, condom use with male partners in the last 12 months, the number of casual female sexual partners in the last 12 months, drug injection in the last 12 months, sex overseas in the last 12 months, past gonorrhea infection, past nonspecific urethritis infection, past syphilis infection, contact with a gonorrhea case, contact with a chlamydia case, and contact with a syphilis case (Figure 2). We formed the final HIV and STI risk prediction questionnaire with the top 10 predictors for each infection.

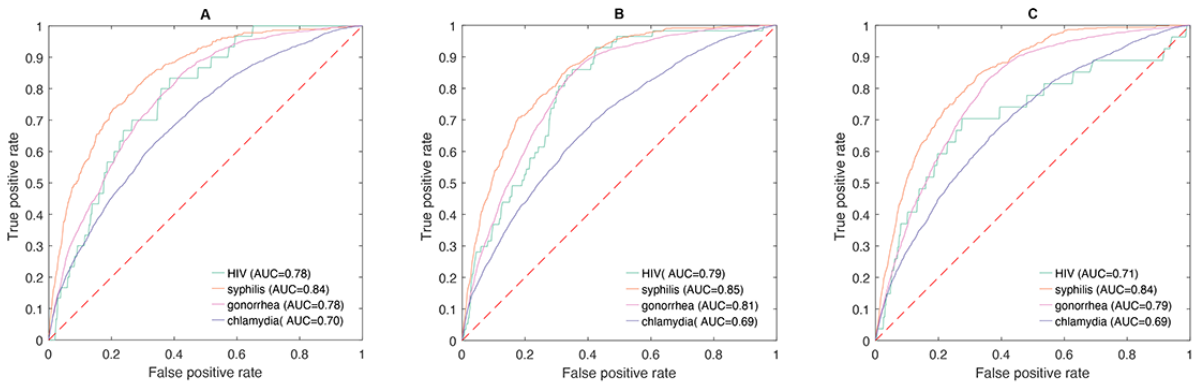
**Figure 2.** Importance of the top 10 predictors in the prediction of HIV or sexually transmission infections (STIs) using a gradient boosting machine, for detecting (A) HIV, (B) syphilis, (C) gonorrhea, and (D) chlamydia.




## Establishment and Evaluation of the HIV and STI Risk Prediction Tool, MySTIRisk

Based on the selected most important predictors and the best model (ensemble ENR+GBM+RF), we built a HIV and STI risk prediction tool, named *MySTIRisk*. We examined *MySTIRisk* and demonstrated its performance on testing to be acceptable or excellent (AUC for HIV=0.78; AUC for syphilis=0.84; AUC for gonorrhea=0.78; AUC for chlamydia=0.70), similar to its original model based on predictors. Our risk prediction tool obtained stable performance on external validation data from 2019 (AUC for HIV=0.79; AUC for syphilis=0.85; AUC for gonorrhea=0.81; AUC for chlamydia=0.69). Our risk prediction tool also achieved stable performance on external validation data from 2020-2021 (AUC for HIV=0.71; AUC for syphilis=0.84; AUC for gonorrhea=0.79; AUC for chlamydia=0.69; Figure 3 and Tables S23-S26 in Multimedia Appendix 1). Using the selected predictors, our risk prediction tool showed comparable AUCs to the best machine learning model using all predictors (Table S27 in Multimedia Appendix 1).

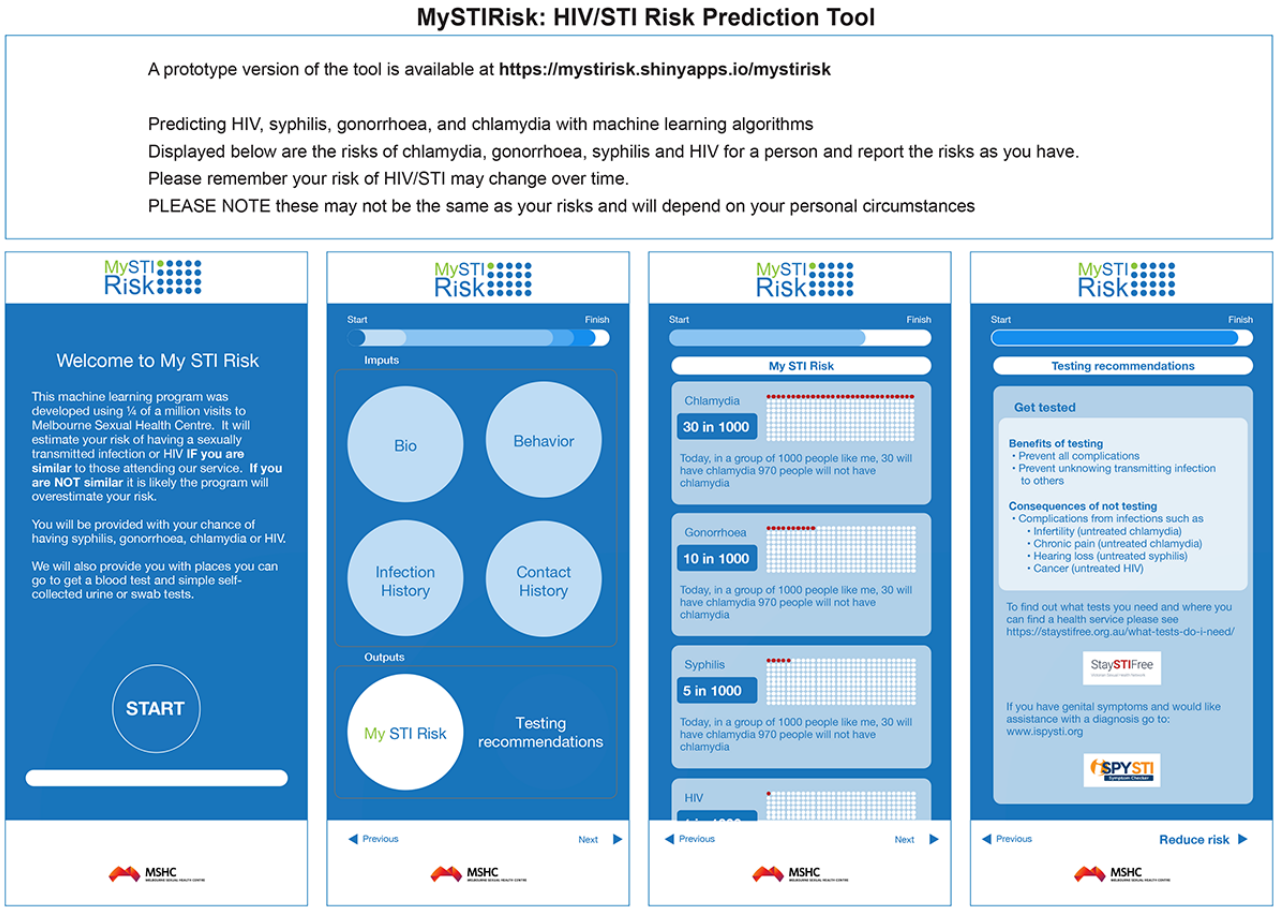
**Figure 3.** Receiver operating characteristic curve performance of the HIV and sexually transmitted infection (STI) risk prediction tool on (A) testing data analysis from 2015-2018, (B) external data validation analysis from 2019, and (C) external data validation analysis from 2020-2021. AUC: area under the curve.




To estimate the risk of HIV or an STI, we fitted the data using a logistic function to provide a fitting curve for each model-predicted probability and infection prevalence (Figures S4-S7 in Multimedia Appendix 1). Then, a prototype version of the tool was created with R Shiny [46,47] to allow for individual input and HIV and STI risk computation. A prototype version of the tool is available online [48]. The graphical user interface elements of the tool are summarized in Figure 4. The web application collects individual characteristics, processes the collected characteristics, loads the trained machine learning models, calculates a quantitative HIV and STI risk, and displays the results of the risk and recommendations. The web application's input was designed using previous successful websites or internal CASI questionnaires (60,000 entries a year)

that operate at MSHC and used individual characteristic data, including demographics, sexual practices, STI history, and STI contact history. The web application's output includes HIV and STI risk prediction results and recommendations that were developed in consultation with Professor Jon Emery at the University of Melbourne, who is an expert in the communication of risk (see the Acknowledgments section). We acknowledge that this is a prototype and that further development will take place in optimizing this output for accurate risk communication.

**Figure 4.** Graphical user interface elements of the HIV and sexually transmitted infection (STI) risk prediction tool, called MySTIRisk. A prototype version of the tool is available at [48]. Machine learning algorithms are used to predict a person's risk of chlamydia, gonorrhea, syphilis, and HIV.




These are examples of the HIV and STI risk prediction results:

> *Your HIV risk is about 2/1000. In a group of 1000 people like me, 2 will have HIV. 998 people will not have HIV.*
>
> *Your syphilis risk is about 10/1000. In a group of 1000 people like me, 10 will have syphilis. 990 people will not have syphilis.*
>
> *Your gonorrhea risk is about 30/1000. In a group of 1000 people like me, 30 will have gonorrhea. 970 people will not have gonorrhea.*
>
> *Your chlamydia risk is about 50/1000. In a group of 1000 people like me, 50 will have chlamydia. 950 people will not have chlamydia.*

The following examples describe testing recommendations:

- Benefits of testing: Prevent all complications and prevent unknowingly transmitting infection to others.
- Consequences of not testing: Complications from infections such as infertility (untreated chlamydia), chronic pain (untreated chlamydia), hearing loss (untreated syphilis), and cancer (untreated HIV).

# Discussion

## Principal Findings

This is the first web-based risk prediction tool based on machine learning algorithms and self-reported data to accurately identify HIV and syphilis, gonorrhea, and chlamydia infection in men and women and was stable on external validation. Our findings showed that machine learning algorithms could predict HIV and STIs in clinic attendees. Our results also showed that stacking ensemble learning algorithms perform better than individual machine learning models to predict HIV and STIs. We then developed a web-based application to provide an immediate and individualized assessment for the risk of a positive diagnosis of HIV and 3 STIs. Our application could be a part of clinic websites or digital health platforms to identify individuals with a higher risk of HIV and STIs or potential candidates for HIV pre-exposure prophylaxis (PrEP). Further validation studies in other countries can assess the usefulness of this risk prediction tool, which helps reduce HIV and STI incidence and the cost of HIV and STI screening, which requires expensive equipment and specialized expertise.

### Comparison With Prior Work

Our results showed that nonlinear machine learning algorithms provided better performance than the conventional logistic regression for predicting HIV and STIs in men and women. Our findings are consistent with the results of previous machine learning predictive models for HIV and STIs [19,21,24,31]. Bao et al [21] showed that a GBM model performed better than logistic regression in MSM. Our study suggests that nonlinear machine learning models (eg, GBM, RF) could provide better performance than conventional logistic regression even without ensemble learning.

Our results showed that the stacking ensemble machine learning techniques outperform individual machine learning models. We systematically developed and tested 34 machine learning models and found that stacking ensemble learning technology outperformed individual machine learning models [18]. Previous studies have used ensemble learning models to predict an individual's HIV risk [19,25]; however, no study has looked at the risk of gonorrhea, chlamydia, or syphilis using ensemble learning models. The only study we could identify was one that had predicted the risk of a repeat STI with ensemble learning. Elder et al [31] showed that an ensemble of models could perform better for 2 or more repeat STIs within 730 days of follow-up than the individual classifiers (AUC=0.76). Our results found that stacking ensemble techniques could also be applied to enhance the performance of HIV prediction. The AUC of our ensemble HIV model (AUC=0.78, 95% CI 0.74-0.83) was higher than that in a similar study in Kenya and Uganda for HIV risk prediction (AUC=0.73, 95% CI 0.71-0.76) [25]. We also found that the combinations of more individual machine learning models do not necessarily lead to a better stacking ensemble model. For example, in our study, the stacking ensemble learning of 4 models for syphilis was not higher than a stacking ensemble learning of 3 models. We also found that a better performing stacking ensemble model always included GBM. The findings of our stacking ensemble learning strategies may have implications for future stacking ensemble learning frameworks.

Our models have several strengths compared with previous machine learning models for predicting HIV and STIs. First, our predictive models were not limited to high-risk groups (such as MSM). HIV and STI risk prediction models have been published previously but mainly for high-risk individuals, such as MSM [20,21,24,29]. Our models could predict HIV and STI acquisition in both men and women, including homosexual and heterosexual individuals. Second, our predictive models only used self-reported and simple questions to develop models. Previously published studies used numerous predictors for their models [23,26-28]. Third, we systematically developed 26 ensemble models. In our study, we tested all possible combinations of 5 base models. The final strength of our research is that we performed 2 external validation analyses of each model.

We were unable to locate any web-based, publicly available tool to quantify STI risk. We identified some available web-based HIV prediction tools, such as the "HIV risk prediction tool" [49], "HIV/AIDS Risk Calculator" [50], and "Online Risk Assessment" [51]. We also identified some available web-based STI prediction tools, such as "Find out if you need to get tested for an STD" [52], "Online STI Testing" [53], and "Take a free test" [54]. These HIV and STI prediction tools provide only subjective terms such as "high" risk or "You are advised to take an HIV/STI test." Our risk prediction tool could quantify the risk of HIV and STIs. In addition, our artificial intelligence (AI)–based risk prediction tool can simultaneously provide risk scores for HIV and 3 common STIs (gonorrhea, chlamydia, and syphilis) for men and women aged 18 years and older.

### Implications

Our web-based HIV and STI risk prediction tool can be used as a screening tool to potentially increase HIV and STI testing and encourage access to testing and health care (Figure S8 in Multimedia Appendix 1). The tool could be used on clinic websites so the public could assess their risk and then decide if they would attend a clinic for testing. It may also be used within a clinic to identify and triage those at higher risk of HIV and STIs if the demand in the clinic is too great to see everyone who attended. However, an AI-based risk prediction tool cannot replace formal HIV and STI testing and treatment in clinical settings, but it would allow individuals to understand their own risks and increase testing uptake. Our tool could increase risk perception and concern about infection, thus increasing HIV and STI testing. A study in the British population showed that increased risk perceptions are associated with greater STI health care use [12]. Further external validation of our AI-based risk prediction tool in other countries or regions, such as low- and middle-income countries, may provide an opportunity to reduce the cost of HIV and STI screening by better focusing testing on those at highest risk [55].

There are many possible ways that our web-based risk prediction tool could be potentially used, including as part of a behavioral intervention to control HIV and STIs or to help clinicians or public health workers identify high-risk individuals for risk management or further interventions. An example of this exists in adolescent health risk behaviors. Researchers used an individual's risk behavior scores and personalized feedback as part of an intervention for health behaviors, including nutritional behaviors, physical activity, and sleep [56]. In this randomized clinical trial, the youths in the intervention group significantly reduced their risk behavior scores at 3 months compared with the control group [56]. Our web-based risk prediction tool could serve as a behavioral intervention tool in the same way.

Future work will investigate the effectiveness of this web-based HIV and STI risk prediction tool for behavioral change (ie, uptake of PrEP or condom promotion) and STI service utilization behaviors (timely clinic attendance and HIV and STI testing uptake) after receiving risk prediction results and testing recommendations. Implementing this web-based HIV and STI prediction tool may encourage individuals with STI symptoms or those at high risk without symptoms to attend health services for timely testing and regular testing. Since February 2009, the MSHC has offered MSM regular SMS reminders for STI screening [57]. For example, providing an estimated risk of HIV and STIs and risk reduction advice (ie, uptake of PrEP or

condom promotion) among high-risk populations (eg, MSM) in an SMS reminder message may encourage testing and behavioral changes.

### Limitations

This study has some limitations. First, the predictive factors depend on self-reported information from the CASI system, which is subject to the participants' recall, nonresponse, and social desirability bias. For example, MSM who declined to report the number of male partners were at a higher risk of chlamydia [58]. There has been substantial work undertaken on the CASI system's validity and accuracy [59]. Second, machine learning models may suffer from overfitting. We used repeated CV to tackle the overfitting problem. We also used ensemble learning methods to enhance the model's generalizability. Third, the generalizability of our models to those not attending the clinic or to other countries or regions is limited because it was derived from a single sexual health service. Thus, if it is used in other countries and regions, further validation is required. Finally, the risks of HIV have changed rapidly over this time by introducing PrEP, so future models will need to include this question, given how the potency of this single preventive strategy.

### Conclusions

This is the first web-based risk assessment tool using machine learning algorithms and self-reported data to identify HIV, syphilis, gonorrhea, and chlamydia in men and women. Our online risk prediction tool could accurately predict the risk of HIV and STIs in clinic attendees with a simple self-administered questionnaire. Our risk prediction tool could be part of clinic websites or digital health platforms. The public can use this risk prediction tool to assess their HIV and STI risk to inform testing. Clinicians or public health workers can use this risk prediction tool to identify high-risk individuals for further interventions.

## Acknowledgments

EC and JJO are supported by Australian National Health and Medical Research Council Emerging Leadership Investigator Grants (GNT1172873 and GNT1193955, respectively). CKF is supported by an Australian National Health and Medical Research Council Leadership Investigator Grant (GNT1172900). LZ is supported by the National Natural Science Foundation of China (Grant number: 81950410639); Outstanding Young Scholars Support Program (Grant number: 3111500001); Xi'an Jiaotong University Basic Research and Profession Grant (Grant number: xtr022019003, xzy032020032); Epidemiology modeling and risk assessment (Grant number: 20200344); and Xi'an Jiaotong University Young Scholar Support Grant (Grant number: YX6J004). The authors want to acknowledge Afrizal Afrizal from the Melbourne Sexual Health Centre (MSHC) for data extraction. The authors thank Glenda Fehler for her contribution to data cleaning. The authors also would like to acknowledge Jon Emery from the University of Melbourne for an insightful discussion on risk prediction tools (eg, Figure 4). We thank Mark Chung at the MSHC for his assistance in preparing Figure 4.

## Authors' Contributions

XX, CKF, and LZ conceived and designed the study. XX cleaned the data, established the models and coding, wrote the first draft, and edited the manuscript. WL, EC, CKF, and LZ contributed to data cleaning. XX, ZG, ZY, YB, and LZ contributed to establishing the models and coding. JW and XX developed the web-based application. CKF and LZ contributed to establishing the web-based application. EC, CKF, and LZ contributed to data verification and supervision. EC, YB, ZY, ZG, JJO, WL, CKF, and LZ contributed to the interpretation of data and manuscript revision. All authors contributed to the preparation of the manuscript and approved the final manuscript.

## Conflicts of Interest

None declared.

## Multimedia Appendix 1

Supplementary tables and figures.
[DOCX File , 728 KB-Multimedia Appendix 1]

## References

1. Ramchandani MS, Golden MR. Confronting rising STIs in the era of PrEP and treatment as prevention. Curr HIV/AIDS Rep 2019 Jun;16(3):244-256 [FREE Full text] [doi: 10.1007/s11904-019-00446-5] [Medline: 31183609]
2. Chow EPF, Grulich AE, Fairley CK. Epidemiology and prevention of sexually transmitted infections in men who have sex with men at risk of HIV. Lancet HIV 2019 Jun;6(6):e396-e405. [doi: 10.1016/S2352-3018(19)30043-8] [Medline: 31006612]
3. Report on global sexually transmitted infection surveillance 2018. World Health Organization. 2018. URL: https://apps.who.int/iris/bitstream/handle/10665/277258/9789241565691-eng.pdf [accessed 2019-05-04]
4. HIV, viral hepatitis and sexually transmissible infections in Australia Annual surveillance report 2021. Kirby Institute. 2021. URL: https://kirby.unsw.edu.au/sites/default/files/kirby/report/Annual-Suveillance-Report-2021_HIV.pdf [accessed 2022-04-06]

5. HIV, viral hepatitis and sexually transmissible infections in Australia: Annual surveillance report 2018. Kirby Institute. 2018. URL: https://kirby.unsw.edu.au/report/hiv-viral-hepatitis-and-sexually-transmissible-infections-australia-annual-surveillance [accessed 2019-05-08]
6. Mayaud P, Mabey D. Approaches to the control of sexually transmitted infections in developing countries: old problems and modern challenges. Sex Transm Infect 2004 Jun 01;80(3):174-182 [FREE Full text] [doi: 10.1136/sti.2002.004101] [Medline: 15169997]
7. Global health sector strategy on Sexually Transmitted Infections, 2016-2021. World Health Organization. 2016 Oct 03. URL: https://www.who.int/publications/i/item/WHO-RHR-16.09 [accessed 2021-04-13]
8. UNAIDS data 2018. UNAIDS. 2018. URL: https://www.unaids.org/sites/default/files/media_asset/unaids-data-2018_en.pdf [accessed 2021-07-12]
9. Levy SB, Gunta J, Edemekong P. Screening for sexually transmitted diseases. Prim Care 2019 Mar;46(1):157-173. [doi: 10.1016/j.pop.2018.10.013] [Medline: 30704656]
10. Vermund SH, Wilson CM. Barriers to HIV testing-where next? The Lancet 2002 Oct;360(9341):1186-1187. [doi: 10.1016/s0140-6736(02)11291-8]
11. Collins IM, Bickerstaffe A, Ranaweera T, Maddumarachchi S, Keogh L, Emery J, et al. iPrevent®: a tailored, web-based, decision support tool for breast cancer risk assessment and management. Breast Cancer Res Treat 2016 Feb;156(1):171-182 [FREE Full text] [doi: 10.1007/s10549-016-3726-y] [Medline: 26909793]
12. Clifton S, Mercer CH, Sonnenberg P, Tanton C, Field N, Gravningen K, et al. STI risk perception in the British population and how it relates to sexual behaviour and STI healthcare use: findings from a cross-sectional survey (Natsal-3). EClinicalMedicine 2018 Aug;2-3:29-36 [FREE Full text] [doi: 10.1016/j.eclinm.2018.08.001] [Medline: 30320305]
13. Ooi CY, Ng CJ, Sales AE, Lim HM. Implementation strategies for web-based apps for screening: scoping review. J Med Internet Res 2020 Jul 20;22(7):e15591 [FREE Full text] [doi: 10.2196/15591] [Medline: 32706655]
14. Eftekhar B, Mohammad K, Ardebili HE, Ghodsi M, Ketabchi E. Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data. BMC Med Inform Decis Mak 2005 Feb 15;5:3 [FREE Full text] [doi: 10.1186/1472-6947-5-3] [Medline: 15713231]
15. Rajula HSR, Verlato G, Manchia M, Antonucci N, Fanos V. Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. Medicina (Kaunas) 2020 Sep 08;56(9):1 [FREE Full text] [doi: 10.3390/medicina56090455] [Medline: 32911665]
16. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. Nat Methods 2018 Apr 3;15(4):233-234 [FREE Full text] [doi: 10.1038/nmeth.4642] [Medline: 30100822]
17. Garg R, Dong S, Shah S, Jonnalagadda S. A bootstrap machine learning approach to identify rare disease patients from electronic health records. arXiv 2016 Sep 06:1-8.
18. Latha CBC, Jeeva SC. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked 2019;16:100203. [doi: 10.1016/j.imu.2019.100203]
19. Xu X, Ge Z, Chow EPF, Yu Z, Lee D, Wu J, et al. A machine-learning-based risk-prediction tool for HIV and sexually transmitted infections acquisition over the next 12 months. J Clin Med 2022 Mar 25;11(7):1818 [FREE Full text] [doi: 10.3390/jcm11071818] [Medline: 35407428]
20. 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 Jan 17;19(2):1010 [FREE Full text] [doi: 10.3390/ijerph19021010] [Medline: 35055826]
21. Bao Y, Medland NA, Fairley CK, Wu J, Shang X, Chow EPF, 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 Jan;82(1):48-59. [doi: 10.1016/j.jinf.2020.11.007] [Medline: 33189772]
22. Turbé V, Herbst C, Mngomezulu T, Meshkinfamfard S, Dlamini N, Mhlongo T, et al. Deep learning of HIV field-based rapid tests. Nat Med 2021 Jul;27(7):1165-1170 [FREE Full text] [doi: 10.1038/s41591-021-01384-9] [Medline: 34140702]
23. Duthe J, Bouzille G, Sylvestre E, Chazard E, Arvieux C, Cuggia M. How to identify potential candidates for HIV pre-exposure prophylaxis: an AI algorithm reusing real-world hospital data. Stud Health Technol Inform 2021 May 27;281:714-718. [doi: 10.3233/SHTI210265] [Medline: 34042669]
24. Xiang Y, Fujimoto K, Li F, Wang Q, Del Vecchio N, Schneider J, et al. Identifying influential neighbors in social networks and venue affiliations among young MSM: a data science approach to predict HIV infection. AIDS 2021 May 01;35(Suppl 1):S65-S73 [FREE Full text] [doi: 10.1097/QAD.0000000000002784] [Medline: 33306549]
25. Balzer LB, Havlir DV, Kamya MR, Chamie G, Charlebois ED, Clark TD, et al. Machine learning to identify persons at high-risk of human immunodeficiency virus acquisition in rural Kenya and Uganda. Clin Infect Dis 2020 Dec 03;71(9):2326-2333 [FREE Full text] [doi: 10.1093/cid/ciz1096] [Medline: 31697383]
26. Gruber S, Krakower D, Menchaca J, Hsu K, Hawrusik R, Maro J, et al. Using electronic health records to identify candidates for human immunodeficiency virus pre-exposure prophylaxis: An application of super learning to risk prediction when the outcome is rare. Stat Med 2020 Oct 15;39(23):3059-3073 [FREE Full text] [doi: 10.1002/sim.8591] [Medline: 32578905]

27. Marcus JL, Hurley LB, Krakower DS, Alexeeff S, Silverberg MJ, Volk JE. Use of electronic health record data and machine learning to identify candidates for HIV pre-exposure prophylaxis: a modelling study. The Lancet HIV 2019 Oct;6(10):e688-e695. [doi: 10.1016/s2352-3018(19)30137-7]
28. Krakower DS, Gruber S, Hsu K, Menchaca JT, Maro JC, Kruskal BA, et al. Development and validation of an automated HIV prediction algorithm to identify candidates for pre-exposure prophylaxis: a modelling study. Lancet HIV 2019 Oct;6(10):e696-e704 [FREE Full text] [doi: 10.1016/S2352-3018(19)30139-0] [Medline: 31285182]
29. Xiang Y, Fujimoto K, Schneider J, Jia Y, Zhi D, Tao C. Network context matters: graph convolutional network model over social networks improves the detection of unknown HIV infections among young men who have sex with men. J Am Med Inform Assoc 2019 Nov 01;26(11):1263-1271 [FREE Full text] [doi: 10.1093/jamia/ocz070] [Medline: 31197365]
30. Feller DJ, Zucker J, Yin MT, Gordon P, Elhadad N. Using clinical notes and natural language processing for automated HIV risk assessment. J Acquir Immune Defic Syndr 2018 Feb 01;77(2):160-166 [FREE Full text] [doi: 10.1097/QAI.0000000000001580] [Medline: 29084046]
31. Elder HR, Gruber S, Willis SJ, Cocoros N, Callahan M, Flagg EW, et al. Can machine learning help identify patients at risk for recurrent sexually transmitted infections? Sex Transm Dis 2021 Jan;48(1):56-62. [doi: 10.1097/OLQ.0000000000001264] [Medline: 32810028]
32. Vodstrcil LA, Fairley CK, Williamson DA, Bradshaw CS, Chen MY, Chow EPF. Immunity to hepatitis A among men who have sex with men attending a large sexual health clinic in Melbourne, Australia, 2012-2018. Sex Transm Infect 2020 Jun;96(4):265-270. [doi: 10.1136/sextrans-2019-054327] [Medline: 32169881]
33. Misson J, Chow EPF, Chen MY, Read TRH, Bradshaw CS, Fairley CK. Trends in gonorrhoea infection and overseas sexual contacts among females attending a sexual health centre in Melbourne, Australia, 2008-2015. Commun Dis Intell (2018) 2018;42:1 [FREE Full text] [Medline: 30626294]
34. Chow E, Hocking J, Ong J, Phillips T, Fairley C. Sexually transmitted infection diagnoses and access to a sexual health service before and after the national lockdown for COVID-19 in Melbourne, Australia. Open Forum Infect Dis 2021 Jan;8(1):ofaa536 [FREE Full text] [doi: 10.1093/ofid/ofaa536] [Medline: 33506064]
35. Feng J, Xu Y, Jiang Y, Zhou ZH. Soft gradient boosting machine. arXiv 2020 Jun 07:1-16 [FREE Full text]
36. Caron M, Allard R, Bédard L, Latreille J, Buckeridge DL. Enteric disease episodes and the risk of acquiring a future sexually transmitted infection: a prediction model in Montreal residents. J Am Med Inform Assoc 2016 Nov;23(6):1159-1165. [doi: 10.1093/jamia/ocw026] [Medline: 27026613]
37. Powers KA, Price MA, Karita E, Kamali A, Kilembe W, Allen S, et al. Prediction of extended high viremia among newly HIV-1-infected persons in sub-Saharan Africa. PLoS One 2018;13(4):e0192785 [FREE Full text] [doi: 10.1371/journal.pone.0192785] [Medline: 29614069]
38. Na K. Prediction of future cognitive impairment among the community elderly: A machine-learning based approach. Sci Rep 2019 Mar 04;9(1):3335 [FREE Full text] [doi: 10.1038/s41598-019-39478-7] [Medline: 30833698]
39. Rigatti SJ. Random forest. J Insur Med 2017;47(1):31-39. [doi: 10.17849/insm-47-01-31-39.1] [Medline: 28836909]
40. Venkata Ramana B, Babu MP, Venkateswarlu N. A critical study of selected classification algorithms for liver disease diagnosis. IJDMS 2011 May 31;3(2):101-114. [doi: 10.5121/ijdms.2011.3207]
41. Jang H, Cho K. Applications of deep learning for the analysis of medical data. Arch Pharm Res 2019 Jun;42(6):492-504. [doi: 10.1007/s12272-019-01162-9] [Medline: 31140082]
42. Pari R, Sandhya M, Sankar S. A multitier stacked ensemble algorithm for improving classification accuracy. Comput. Sci. Eng 2020 Jul 1;22(4):74-85. [doi: 10.1109/mcse.2018.2873940]
43. Menardi G, Torelli N. Training and assessing classification rules with imbalanced data. Data Min Knowl Disc 2012 Oct 30;28(1):92-122. [doi: 10.1007/s10618-012-0295-5]
44. Shehzad A, Rockwood K, Stanley J, Dunn T, Howlett SE. Use of Patient-Reported Symptoms from an Online Symptom Tracking Tool for Dementia Severity Staging: Development and Validation of a Machine Learning Approach. J Med Internet Res 2020 Nov 11;22(11):e20840 [FREE Full text] [doi: 10.2196/20840] [Medline: 33174853]
45. Mandrekar J. Receiver Operating Characteristic Curve in Diagnostic Test Assessment. Journal of Thoracic Oncology 2010 Sep;5(9):1315-1316 [FREE Full text] [doi: 10.1097/jto.0b013e3181ec173d]
46. Beeley C. Web Application Development with R Using Shiny: Build stunning graphics and interactive data visualizations to deliver cutting-edge analytics. Birmingham, UK: Packt Publishing Ltd; 2016.
47. Gregorich M, Heinzel A, Kammer M, Meiselbach H, Böger C, Eckardt K, et al. A prediction model for the decline in renal function in people with type 2 diabetes mellitus: study protocol. Diagn Progn Res 2021 Nov 18;5(1):19 [FREE Full text] [doi: 10.1186/s41512-021-00107-5] [Medline: 34789343]
48. MySTIRisk. URL: https://mystirisk.shinyapps.io/mystirisk [accessed 2022-03-08]
49. HIV Risk Reduction Tool. Centers for Disease Control and Prevention. URL: https://hivrisk.cdc.gov/risk-estimator-tool/#-sb [accessed 2022-04-06]
50. HIV/AIDS Risk Calculator. URL: https://www.medindia.net/patients/calculators/hiv-risk-calculator.asp [accessed 2022-04-06]
51. Online Risk Assessment. URL: https://aidsconcern.org.hk/en/testing-service/assess/
52. Find out if you need to get tested for an STD. URL: https://stdwizard.com/#/home
53. Online STI Testing. URL: https://www.getthefacts.health.wa.gov.au/online-sti-testing

54. Take a free test. URL: https://www.couldihaveit.com.au/Take-a-free-test
55. Alami H, Rivard L, Lehoux P, Hoffman SJ, Cadeddu SBM, Savoldelli M, et al. Artificial intelligence in health care: laying the foundation for responsible, sustainable, and inclusive innovation in low- and middle-income countries. Global Health 2020 Jun 24;16(1):52 [FREE Full text] [doi: 10.1186/s12992-020-00584-1] [Medline: 32580741]
56. Richardson LP, Zhou C, Gersh E, Spielvogle H, Taylor JA, McCarty CA. Effect of electronic screening with personalized feedback on adolescent health risk behaviors in a primary care setting: a randomized clinical trial. JAMA Netw Open 2019 May 03;2(5):e193581 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.3581] [Medline: 31074815]
57. Zou H, Fairley CK, Guy R, Bilardi J, Bradshaw CS, Garland SM, et al. Automated, computer generated reminders and increased detection of gonorrhoea, chlamydia and syphilis in men who have sex with men. PLoS One 2013;8(4):e61972 [FREE Full text] [doi: 10.1371/journal.pone.0061972] [Medline: 23613989]
58. Chow E, Carlin J, Read T, Chen M, Bradshaw C, Sze J, et al. Factors associated with declining to report the number of sexual partners using computer-assisted self-interviewing: a cross-sectional study among individuals attending a sexual health centre in Melbourne, Australia. Sex. Health 2018;15(4):350 [FREE Full text] [doi: 10.1071/sh18024]
59. Fairley CK, Sze JK, Vodstrcil LA, Chen MY. Computer-assisted self interviewing in sexual health clinics. Sex Transm Dis 2010 Nov;37(11):665-668. [doi: 10.1097/OLQ.0b013e3181f7d505] [Medline: 20975481]

## Abbreviations

**AI:** artificial intelligence
**AUC:** area under the curve
**CASI:** computer-assisted self-interview system
**CV:** cross-validation
**DL:** deep learning
**EHR:** electronic health records
**ENR:** elastic net regression
**GBM:** gradient boosting machine
**LASSO:** least absolute shrinkage and selection operator
**MSHC:** Melbourne Sexual Health Centre
**MSM:** men who have sex with men
**NAAT:** nucleic amplification tests
**NB:** Naive Bayes
**PrEP:** pre-exposure prophylaxis
**RF:** random forest
**RR:** ridge regression
**STI:** sexually transmitted infection
**WHO:** World Health Organization

*Edited by R Kukafka; submitted 09.03.22; peer-reviewed by X Zou, X Ma, S El kefi; comments to author 01.04.22; revised version received 13.04.22; accepted 28.07.22; published 25.08.22*

*Please cite as:*
*Xu X, Yu Z, Ge Z, Chow EPF, Bao Y, Ong JJ, Li W, Wu J, Fairley CK, Zhang L*
*Web-Based Risk Prediction Tool for an Individual's Risk of HIV and Sexually Transmitted Infections Using Machine Learning Algorithms: Development and External Validation Study*
*J Med Internet Res 2022;24(8):e37850*
*URL: https://www.jmir.org/2022/8/e37850*
*doi: 10.2196/37850*
*PMID:*