



Published in final edited form as:

AIDS. 2025 March 01; 39(3): 290–297. doi:10.1097/QAD.0000000000004055.

Predictors of HIV seroconversion in Botswana: machine learning analysis in a representative, population-based HIV incidence cohort

Yifan CUI^{1,*}, Sikhulile MOYO^{2,3,4,*,#}, Molly PRETORIUS HOLME², Kathleen E HURWITZ⁵, Wonderful CHOGA^{3,4,6,7}, Kara BENNETT⁵, Unoda CHAKALISA³, James Emmanuel SAN^{8,9}, Kutlo MANYAKE³, Coulson KGATHI³, Ame DIPHOKO³, Simani GASEITSIWE³, Tendani GAOLATHE^{3,8}, M. ESSEX^{2,3}, Eric TCHETGEN TCHETGEN^{9,10}, Joseph M. MAKHEMA^{2,3}, Shahin LOCKMAN^{2,3,11,#}

¹Center for Data Science, Zhejiang University, Hangzhou, Zhejiang, China;

²Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Boston, MA, USA;

³Botswana Harvard Health Partnership, Gaborone, Botswana;

⁴Division of Medical Virology, Faculty of Medicine and Health Sciences, Stellenbosch University, Tygerberg, Cape Town, South Africa;

⁵Target RWE, Durham, North Carolina;

⁶School of Allied Health Sciences, Faculty of Health Sciences, University of Botswana, Gaborone, Botswana;

⁷Centre for Epidemic Response and Innovation (CERI), School of Data Science and Computational Thinking, Stellenbosch University, Stellenbosch, South Africa;

⁸Department of Internal Medicine, University of Botswana, Gaborone, Botswana;

⁹Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia, PA, USA;

¹⁰Department of Biostatistics and Epidemiology, The Wharton School, University of Pennsylvania, Philadelphia, PA, USA;

¹¹Division of Infectious Diseases, Brigham & Women's Hospital, Boston, MA, USA.

Abstract

Objective: To identify predictors of HIV acquisition in Botswana.

Design: We applied machine learning approaches to identify HIV risk predictors using existing data from a large, well-characterized HIV incidence cohort.

[#]**Corresponding Authors:** Sikhulile Moyo, MSc, MPH, PhD (sikhulilemoyo@gmail.com), Botswana Harvard Health Partnership, Gaborone, Botswana. Shahin Lockman, MD, MSc (shahin.lockman@gmail.com), Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Boston, MA, USA.

*Equal contribution, listed alphabetically

Methods: We applied machine learning (randomForestSRC) to analyze data from a large population-based HIV incidence cohort enrolled in a cluster-randomized HIV prevention trial in 30 communities across Botswana. We sought to identify the most important risk factors for HIV acquisition, starting with 110 potential predictors.

Results: During a median 29-month follow-up of 8,551 HIV-negative adults, 147 (1.7%) acquired HIV. Our machine learning analysis found that for females, the most important variables for predicting HIV acquisition were the use of injectable hormonal contraception, frequency of sex in the prior 3 months with the most recent partner and residing in a community with HIV prevalence of 29% or higher. For the small proportion (0.3%) of females who had all three risk factors, their estimated probability of acquiring HIV during 29 months of follow-up was 34% (approximate annual incidence of 14%). For males, non-long-term relationships with the most recent partner and community HIV prevalence of 34% or higher were the most important HIV risk predictors. The 6% of males who had both risk factors had a 5.1% probability of acquiring HIV during the follow-up period (approximate annual incidence of 2.1%).

Conclusions: Machine learning approaches allowed us to analyze a large number of variables to efficiently identify key factors strongly predictive of HIV risk. These factors could help target HIV prevention interventions in Botswana.

Clinical Trials Registration: [NCT01965470](https://clinicaltrials.gov/ct2/show/study/NCT01965470)

Keywords

HIV risk factors; HIV incidence; machine learning; Ya Tsie Study; BCPP; Botswana; Africa

BACKGROUND

Despite significant global progress in reducing HIV incidence and HIV-related mortality, new infections continue to undermine efforts to end the HIV epidemic by 2030 [1]. Sub Saharan Africa (SSA) is the region most severely affected by HIV/AIDS. Many SSA countries have provided antiretroviral treatment (ART) to all people diagnosed with HIV since approximately 2016 (“treat all” approach). Despite progress in ART coverage [2], global HIV incidence is only 43% lower than a decade ago [3]. HIV pre-exposure prophylaxis (PrEP) is an important and highly effective HIV prevention tool. However, in the context of generalized HIV epidemics (such as in SSA), it is challenging to identify individuals at highest risk for HIV who may benefit the most from PrEP. While adolescent girls and young women and sero-discordant couples remain at higher risk of HIV acquisition in SSA, exclusive focus of HIV prevention efforts on these groups may miss other individuals with elevated HIV risk. This is important especially important in generalized epidemic settings such as Botswana, which has an estimated HIV prevalence of 20.8% among persons aged 15–64 years [4].

In recent years, machine learning algorithms, with their ability to capture complex interactions within data, have shown promise for identifying predictors of HIV infection risk [5–7], compared with traditional mathematical models. They have been shown to be both efficient and sensitive in classifying and quantifying HIV risk in both high-income and low/middle-income settings [7–9] and to be useful in decision-making around prioritizing

PrEP [10–12]. Some of the advantages of the random forest approach include the ability to handle highly correlated variables; robustness to data noise, missing data, and outliers; and lack of assumptions regarding distribution of data [13, 14]. Random forest models are free from the assumptions of generalised linear models. Traditional models are not efficient in handling large numbers of variables and hence assumptions are made or an *a priori* selection of variables is required [13–15].

We applied machine learning techniques to data from a nationwide community-randomized HIV prevention trial in Botswana that prospectively followed a large population-based HIV incidence cohort in which a large number of potential predictor variables were systematically collected. Our goal was to identify the most important individual-level and community-level risk factors for HIV acquisition in the context of high community-level HIV diagnosis and ART rates.

METHODOLOGY

Study population

We used data from the completed Botswana Combination Prevention Project (BCPP), also known as Ya Tsie [16, 17] ([NCT01965470](#)). BCPP was a pair-matched, cluster-randomized trial in communities across Botswana that tested whether a combination of measures in the intervention communities reduced population-level cumulative incidence of HIV infection [17]. The intervention included intensive HIV testing campaigns, rapid universal ART initiation with active linkage to care, and enhanced voluntary male circumcision services. The trial was conducted from 2013–2018 in 30 rural and peri-urban communities (15 intervention, 15 standard of care) across Botswana. BCPP enrolled a prospective cohort of all consenting persons aged 16–64 years residing in a random 20% sample of households in each of the 30 communities. A baseline participant survey collected extensive sociodemographic information, self-reported HIV risk behavior, HIV testing and treatment history, and health history. All participants without documentation of a prior positive HIV test underwent HIV testing. We followed each cohort participant prospectively at approximately 12 and 29 months after enrollment, and re-tested HIV-negative persons for HIV at each visit.

Ethics

Participants 16–18 years of age provided written assent (with parent/guardian written permission). Adult participants provided written informed consent. The study was approved by the Botswana Health Research and Development Committee and US Centers for Disease Control and Prevention IRB.

Machine Learning Modelling

We used information on individual HIV infection status over time in the HIV incidence cohort to build an interpretable prediction model for predictors of HIV seroconversion based on 110 potential predictors collected in the baseline household survey in the study (Supplemental Table S1). We excluded persons documented as living with HIV at baseline from this analysis. Missing data were processed by multiple imputation using missForest

[18] with maximum number of iterations of 10 and 50 trees to grow in each forest. We stacked 5-fold imputed datasets together to build the random survival forest model, resulting in 25,995 females and 16,760 males included in the machine learning analysis.

We then considered decision trees and random forest to build risk scores, using `randomForestSRC` [19] and following `rpart` function in the `rpart` [20] R package to grow our decision tree model. A natural risk score was defined as the estimated probability of HIV seroconversion during the follow-up period. We used a two-step tree-based algorithm to build our prediction model. The first step was to run a random survival forest to obtain the variable importance of each covariate. We first screened the variables by selecting the top-K variables (10% of the variables with highest variable importance for predicting the risk of HIV acquisition). The variable importance is defined as the difference between prediction error when a variable is “permuted” versus the prediction error otherwise [21]. The second step was to build a classification tree based on the top-K important variables to predict the risk score. The obtained tree is interpretable and provides a flowchart-like structure in which each internal node represents a homogeneous group.

Measures of risk score

We stratified analyses by sex, as the applicability and importance of some potential HIV incidence predictor variables differs by sex (e.g., male circumcision is relevant only to men, and hormonal contraception only to women). We defined the risk of an individual instance as the percentage of participants within the same terminal node who acquired HIV.

Logistic regression

We then performed univariate logistic regression to further validate our tree model, using `glm` function [22], including variables that were identified by our tree models (Table 1, Supplemental Tables S2 and S3). The regression analysis was done using the full dataset.

RESULTS

Study population

We enrolled 12,610 individuals in the prospective cohort (in randomly selected households from the 30 communities): 3,596 (29%) people with HIV and 8,974 (71%) people without HIV at enrollment (the HIV incidence cohort). Most HIV incidence cohort participants were female (60%) with median age 29.6 years (IQR 22.5, 43.2). Retention of the HIV incidence cohort was high: 8551 (95%) of the 8974 HIV-negative participants were-tested for HIV at 12 and/or 29 months after enrollment [16].

During the median 29 months of follow-up of the HIV incidence cohort, 147 HIV seroconversions were observed (i.e. 1.72% of the HIV incidence cohort members seroconverted during follow-up). Annual HIV incidence was lower in the intervention (0.59%) than in the standard of care arm (0.92%), and higher in females (0.71%) than in males (0.40%) [16].

Top-K important variables

The top 11 (10%) variables, ranked by variable importance, are shown in Figures 1A and 1B for females and males, respectively.

For females, the top 11 variables that most strongly predicted HIV acquisition risk were: education level, age category, alcohol use during last sexual encounter, community HIV prevalence, use of injectable hormonal contraception, type of relationship with 2nd most recent partner in past 12 months, energy source, frequency of alcohol use in past 12 months, marital status, perception about whether people work together to solve neighborhood problems, and frequency of sex in last 3 months with most recent sexual partner (Figure 1A).

For males, the top 11 variables that most strongly predicted HIV acquisition risk were: type of relationship with most recent sexual partner, community HIV prevalence, education level, type of relationship with 3rd most recent sexual partner in past 12 months, region of country, frequency of sex in past 3 months with second most recent partner, age of most recent partner, reason for most recent HIV test, whether second most recent partner lived within community, age, and timing of most recent HIV test (Figure 1B).

Interpretable classification tree model

Figures 2A and 2B present the tree models for females and males, respectively. The decimal numbers refer to the estimated probability of acquiring HIV over the follow-up period in the node, and percentages refer to the percentage of the population that is represented in that node. For example, in Figure 2A, the root node includes 100% of female participants, and the estimated probability of overall HIV seroconversion in females is 2.3%.

For females, a combination of using injectable hormonal contraception, frequency of sex with the most recent partner in the past 3 months, and community HIV prevalence of 29% or higher were the most important variables for predicting risk of HIV acquisition.

In the tree model for females, the estimated probability of acquiring HIV was 5.1% for females using injectable hormonal contraception (who comprised 10% of female participants) (Figure 2A). Among the 90% of women who did not use injectable hormonal contraceptives, who had a lower risk of HIV, the most important variables identified were energy source (not gas or electricity) and being less than 35 years of age (Supplemental Figure S1). Among females using injectable hormonal contraception who reported having sex less than 1.5 times in the past 3 months with their most recent sexual partner, their estimated probability of acquiring HIV was 13%. Among females using injectable hormonal contraception who reported having sex less than 1.5 times in the past 3 months with their most recent sexual partner and who lived in a community with HIV prevalence of at least 29%, their estimated probability of acquiring HIV was 34% over the 29-month study period (which translates to approximately 14.1% annual HIV incidence). It is important to note that this latter group that had all 3 of these characteristics comprised only 0.28% of female participants.

For males, reporting that they were in a non-long-term relationship with their most recent partner and living in a community with HIV prevalence >34% were the most important variables for predicting HIV risk.

In the male group tree model, the estimated probability of acquiring HIV was 1.7% for males reporting that they were in a non-long-term relationship with their most recent partner (Figure 2B). Among males reporting that they were in a non-long-term relationship with their most recent partner, their estimated probability of acquiring HIV during the 29-month follow-up period was 5.1% if their community HIV prevalence was at least 34% (approximate annual incidence 2.1%).

Factors associated with HIV acquisition—In the univariate logistic regression analysis of the variables identified in our tree models, we confirmed that female participants using injectable hormonal contraception (odds ratio (OR) = 2.4; 95% confidence interval (CI): 1.9–2.9) and community HIV prevalence of >30% (OR = 1.7; 95% CI: 1.3–2.3) were associated with increased odds of acquiring HIV during the follow-up period (Table 1). While the frequency of sex with the most recent partner was important in identifying the next most important factor in the tree model, this variable had a wide confidence interval (high uncertainty) and was not itself a significant predictor of HIV risk in the logistic model.

We also confirmed that males whose most recent sexual partner was a boyfriend/girlfriend (OR = 4.1; 95% CI: 2.6–6.4), most recent sexual partner was non-long-term (OR = 9.4; 95% CI: 5.5–16.0), and living in a community with HIV prevalence of >30% (OR = 2.2; 95% CI: 1.3–3.9) were each associated with increased odds of acquiring HIV during the follow-up period (Table 1).

Among female participants, HIV acquisition was associated with being aged 16–24 years, using an energy source other than gas or electricity, and a partner using alcohol in the last sex encounter. Married female participants had reduced odds of HIV acquisition as compared to single or never-married participants (OR = 0.36; 95% CI: 0.31–0.67) (Supplemental Table S2).

Among male participants, HIV acquisition was associated with having either non-formal education (OR = 5.64; 95% CI: 1.82–17.46) or junior secondary education (OR = 3.15; 95% CI: 1.22–8.11), as compared to having primary education; with being aged 25–34 years (OR = 2.02; 95% CI: 1.34–3.05), as compared to 16–24 years; and with having an unknown casual partner (OR = 13.24; 95% CI: 3.14–55.88), as compared to having a long-term partner (Supplemental Table S3). However, among female participants, those aged 25–34 years (OR = 0.81; 95% CI: 0.66–0.99) and greater than 35 years (OR = 0.55; 95% CI: 0.40–0.75) had a reduced likelihood of HIV acquisition compared to those aged 16–24 years (Supplemental Table S2).

DISCUSSION

Using machine learning to analyze 110 diverse potential HIV risk predictors in a large, population-level prospective HIV incidence cohort from across Botswana, we identified several factors that were strongly predictive of actual HIV acquisition. In females, the

combination of using injectable hormonal contraception, reporting less frequent sex with their most recent sexual partner, and living in a community with high HIV prevalence was associated with HIV seroconversion. And for females, reporting less frequent sex with the most recent partner may be a proxy for that partner being a casual sexual relationship. In males, the combination of having a non-long-term relationship with their most recent sexual partner and high community HIV prevalence was most predictive of HIV risk. These predictors could conceivably be used by clinicians and HIV prevention programs in Botswana, using readily available and regularly updated nationwide and community-specific prevalence data, to help target PrEP to HIV-negative women and men who have a higher risk for HIV infection.

Machine learning is a powerful approach to evaluating predictors of HIV infection risk when large numbers of individual-level sociodemographic and behavioral factors and community-level factors are available. Tree-based methods in machine learning, including algorithms like decision trees, random forests, and boosting, offer several advantages. Unlike parametric models, tree-based methods can capture nonlinear relationships between features and the response [23]. In addition, random forests and boosting leverage ensemble learning by combining multiple trees to improve predictive accuracy. This ensemble approach helps to reduce overfitting and enhance generalizability. Furthermore, decision trees are interpretable models in the sense that they mimic human decision-making processes by partitioning the feature space into segments that are easy to understand [24]. This makes it easier to explain the machine learner's predictions to practitioners.

While both random forests and boosting usually perform similarly in terms of predictive accuracy, random forests might be more robust to overfitting, especially when dealing with noisy data or datasets with a large number of predictors. Moreover, random forests provide a built-in feature, variable importance, which measures the importance of variables in making predictions [15].

Few prior studies have used machine learning approaches to construct risk profiles or risk scores [7, 25]. These analyses compared the performance of several machine learning algorithms to conventional logistic regression methods and found machine learning methods to be superior for identifying individuals at risk for HIV. Previous observational evidence provided mixed evidence showing a higher risk of HIV acquisition in women using injectable hormonal contraceptive [26–30] (although the randomized ECHO trial did not observe increased risk [31]). The other risk factors identified by our machine learning analysis are generally also plausible in the context of the regional HIV epidemic.

Our study had several strengths. We collected a rich and broad complement of sociodemographic and behavioral factors and followed a large, population-based, representative sample of adults from across Botswana to ascertain HIV incidence. We had high (95%) follow-up and HIV re-testing rates and high data completeness for relevant variables. Our study also has limitations: it was focused on rural/peri-urban areas and enrolled more females than males.

The top 11 variables and the final variables selected by our tree model were consistent with known or potential risk factors for HIV acquisition. However, some of the detailed variables would not likely have been included *a priori* in conventional logistic regression models. In summary, this machine learning approach allowed efficient identification of risk factors that could conceivably be ascertained in a clinical setting in Botswana and even used to help direct and offer interventions such as PrEP.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

The authors thank study participants, dikgosi and other community leaders, and the clinic staff, District Health Management Teams, and Community Health Facilities at study sites; the *Ya Tsie* Study Team at the Botswana Harvard AIDS Institute Partnership, the Harvard T.H. Chan School of Public Health, the CDC Botswana, and the Botswana Ministry of Health; Ria Madison and Lars Colson for administrative support; our implementing partners, Tebelopele Voluntary Counseling and Testing Centers, and Jhpiego - Botswana; and the NIAID Prevention-Africa Data and Safety Monitoring Board. The authors also thank all the members of the Ya Tsie Community Advisory Board.

FUNDING

YC was supported by the National Natural Science Foundation of China. SM was supported by the Fogarty International Center (FIC) at the US National Institutes of Health (K43 TW012350, D43 TW009610). This work was partially supported by grants from FIC (D43 TW009610-09S1) and the National Institute of Allergy and Infectious Diseases (K24 AI131928, K24 AI131924). The BCPP was supported by the United States President's Emergency Plan for AIDS Relief (PEPFAR) through the Centers for Disease Control and Prevention (CDC) (U01 GH000447, U2G GH001911). The views expressed in this publication are those of the authors and not necessarily those of the funding agencies. The funders had no role in the study design, data collection, decision to publish, or preparation of this manuscript.

References

1. Govender K, Nyamaruze P, Cowden RG, Pillay Y, Bekker L-G. Children and young women in eastern and southern Africa are key to meeting 2030 HIV targets: time to accelerate action. *Lancet HIV* 2023; 10:e343–e350. [PubMed: 37149293]
2. Frescura L, Godfrey-Faussett P, Feizzadeh AA, El-Sadr W, Syarif O, Ghys PD, et al. Achieving the 95 95 95 targets for all: A pathway to ending AIDS. *PLOS ONE* 2022; 17:e0272405. [PubMed: 35925943]
3. Stannah J, Soni N, Lam JKS, Giguère K, Mitchell KM, Kronfli N, et al. Trends in HIV testing, the treatment cascade, and HIV incidence among men who have sex with men in Africa: a systematic review and meta-analysis. *Lancet HIV* 2023:e528–e542. [PubMed: 37453439]
4. Mine M, Stafford KA, Laws RL, Marima R, Lekone P, Ramaabya D, et al. Progress towards the UNAIDS 95-95-95 targets in the Fifth Botswana AIDS Impact Survey (BAIS V 2021): a nationally representative survey. *The Lancet HIV* 2024; 11:e245–e254. [PubMed: 38467135]
5. Fieggen J, Smith E, Arora L, Segal B. The role of machine learning in HIV risk prediction. *Front Reprod Health* 2022; 4:1062387. [PubMed: 36619681]
6. Majam M, Segal B, Fieggen J, Smith E, Hermans L, Singh L, et al. Utility of a machine-guided tool for assessing risk behaviour associated with contracting HIV in three sites in South Africa. *Inform Med Unlocked* 2023; 37:101192. [PubMed: 36911795]
7. He J, Li J, Jiang S, Cheng W, Jiang J, Xu Y, 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. [PubMed: 36091522]

8. Mutai CK, McSharry PE, Ngaruye I, Musabanganji E. Use of machine learning techniques to identify HIV predictors for screening in sub-Saharan Africa. *BMC Med Res Methodol* 2021; 21:159. [PubMed: 34332540]
9. 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; 11:1818. [PubMed: 35407428]
10. 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. *Lancet HIV* 2019; 6:e688–e695. [PubMed: 31285183]
11. 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; 6:e696–e704. [PubMed: 31285182]
12. Zheng W, Balzer L, van der Laan M, Petersen M. Constrained binary classification using ensemble learning: an application to cost-efficient targeted PrEP strategies. *Stat Med* 2018; 37:261–279. [PubMed: 28384841]
13. Esteve M, Aparicio J, Rodriguez-Sala JJ, Zhu J. Random Forests and the measurement of super-efficiency in the context of Free Disposal Hull. *Eur J Oper Res* 2023; 304:729–744.
14. Sarker IH. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Comput Sci* 2021; 2:160. [PubMed: 33778771]
15. Aldrich C Process Variable Importance Analysis by Use of Random Forests in a Shapley Regression Framework. *Minerals* 2020; 10:420.
16. Makhema J, Wirth KE, Pretorius Holme M, Gaolathe T, Mmalane M, Kadima E, et al. Universal Testing, Expanded Treatment, and Incidence of HIV Infection in Botswana. *N Engl J Med* 2019; 381:230–242. [PubMed: 31314967]
17. Wirth KE, Gaolathe T, Pretorius Holme M, Mmalane M, Kadima E, Chakalisa U, et al. Population uptake of HIV testing, treatment, viral suppression, and male circumcision following a community-based intervention in Botswana (Ya Tsie/BCPP): a cluster-randomised trial. *Lancet HIV* 2020; 7:e422–e433. [PubMed: 32504575]
18. Stekhoven DJ. Using the missForest package. 2011:11pp. https://stat.ethz.ch/education/semesters/ss2012/ams/paper/missForest_2011.2012.pdf.
19. Ishwaran H, Kogalur U, Blackstone E, Lauer M. Random survival forests. *Ann Appl Stat* 2008; 2:841–860.
20. Therneau T, Atkinson B, Ripley B. rpart: Recursive Partitioning and Regression Trees. 2023:34pp. <https://cran.r-project.org/package=rpart>.
21. Ishwaran H Variable importance in binary regression trees and forests. *Electron J Statist* 2007; 1:519–537.
22. Calcagno V, de Mazancourt C. glmulti: an R package for easy automated model selection with (generalized) linear models. *J Stat Software* 2010; 34:1–29.
23. Athey S, Tibshirani J, Wager S. Generalized random forests. *Ann Statist* 2019; 47:1148–1178.
24. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: A conditional inference framework. *J Comput Graph Stat* 2006; 15:651–674.
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; 71:2326–2333. [PubMed: 31697383]
26. Balkus JE, Brown ER, Hillier SL, Coletti A, Ramjee G, Mgodini N, et al. Oral and injectable contraceptive use and HIV acquisition risk among women in four African countries: a secondary analysis of data from a microbicide trial. *Contraception* 2016; 93:25–31. [PubMed: 26519646]
27. Byrne EH, Anahtar MN, Cohen KE, Moodley A, Padavattan N, Ismail N, et al. Association between injectable progestin-only contraceptives and HIV acquisition and HIV target cell frequency in the female genital tract in South African women: a prospective cohort study. *Lancet Infect Dis* 2016; 16:441–448. [PubMed: 26723758]
28. Wall KM, Kilembe W, Vwalika B, Htee Khu N, Brill I, Chomba E, et al. Hormonal contraception does not increase women's HIV acquisition risk in Zambian discordant couples, 1994–2012. *Contraception* 2015; 91:480–487. [PubMed: 25708502]

29. Wand H, Ramjee G. The effects of injectable hormonal contraceptives on HIV seroconversion and on sexually transmitted infections. *AIDS* 2012; 26:375–380. [PubMed: 22156970]
30. Curtis KM, Hannaford PC, Rodriguez MI, Chipato T, Steyn PS, Kiarie JN. Hormonal contraception and HIV acquisition among women: an updated systematic review. *BMJ Sex Reprod Health* 2020; 46:8–16.
31. Evidence for Contraceptive Options and HIV Outcomes (ECHO) Trial Consortium. HIV incidence among women using intramuscular depot medroxyprogesterone acetate, a copper intrauterine device, or a levonorgestrel implant for contraception: a randomised, multicentre, open-label trial. *Lancet* 2019; 394:303–313. [PubMed: 31204114]

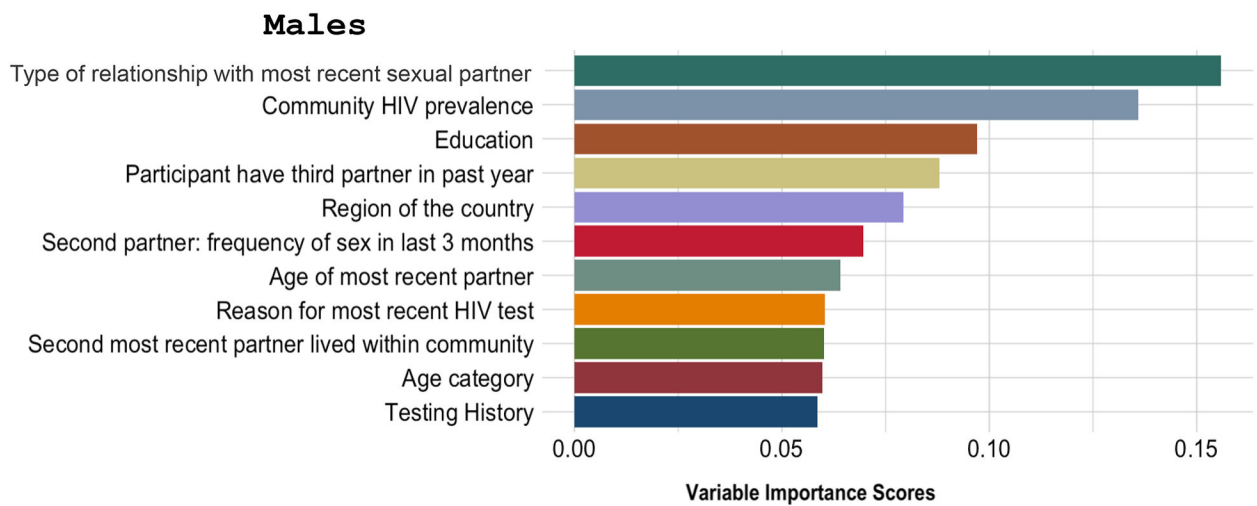
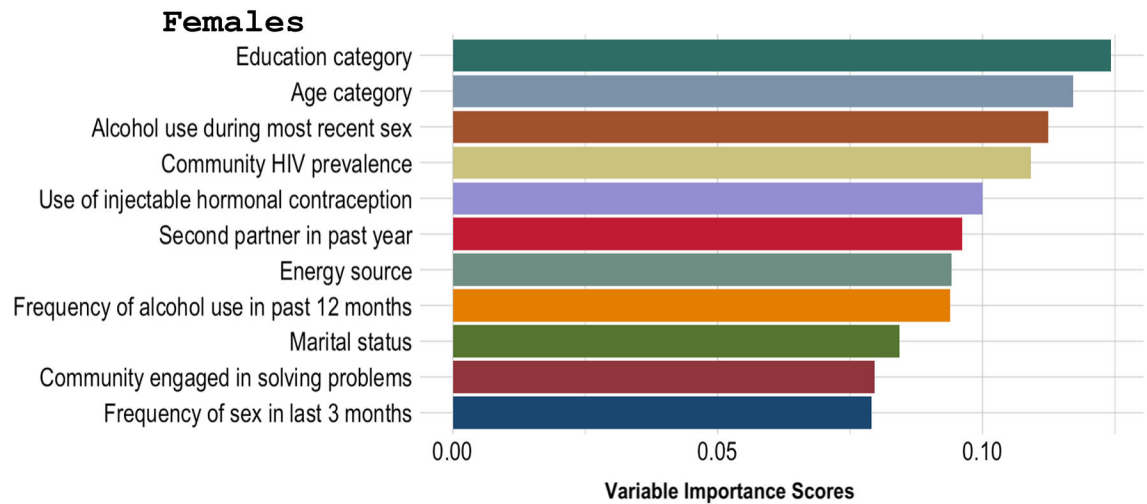
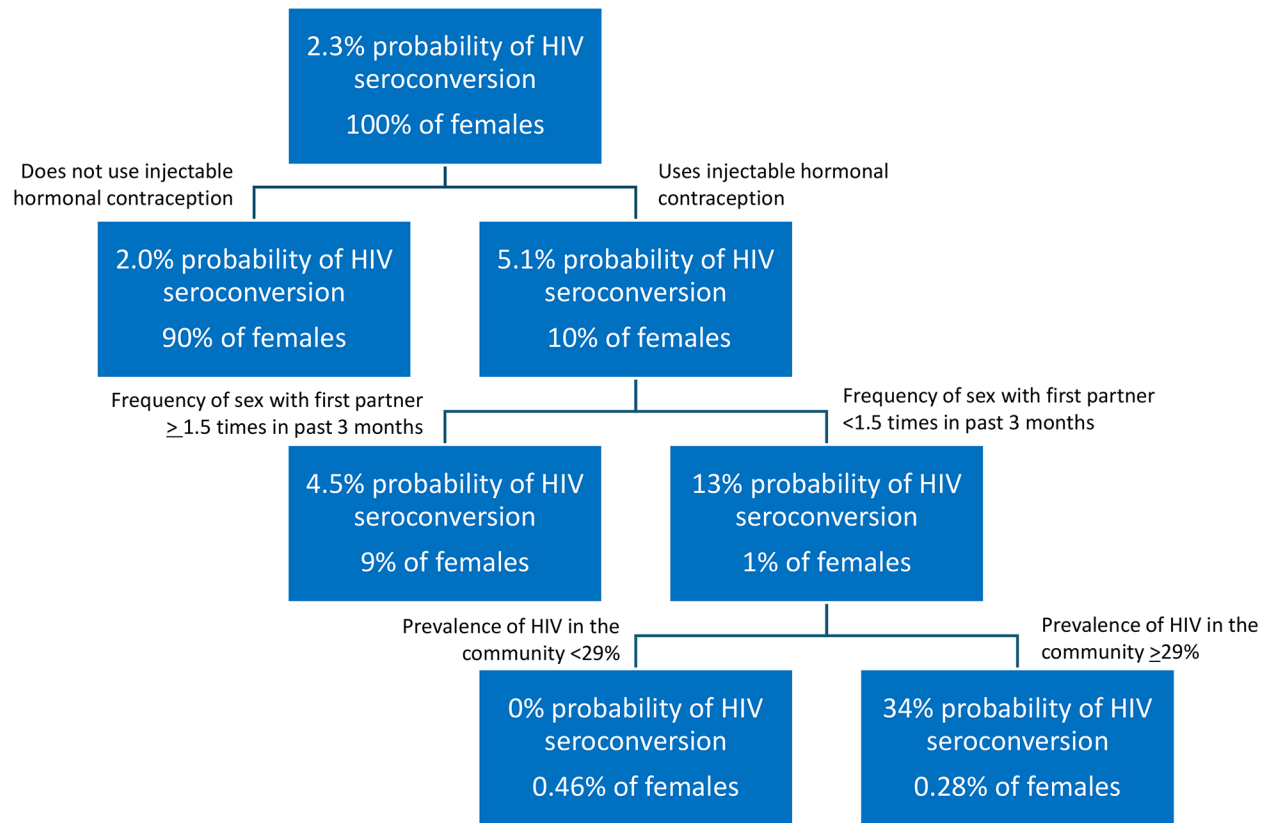


Figure 1.
Top 11 most important variables for females (A) and males (B)



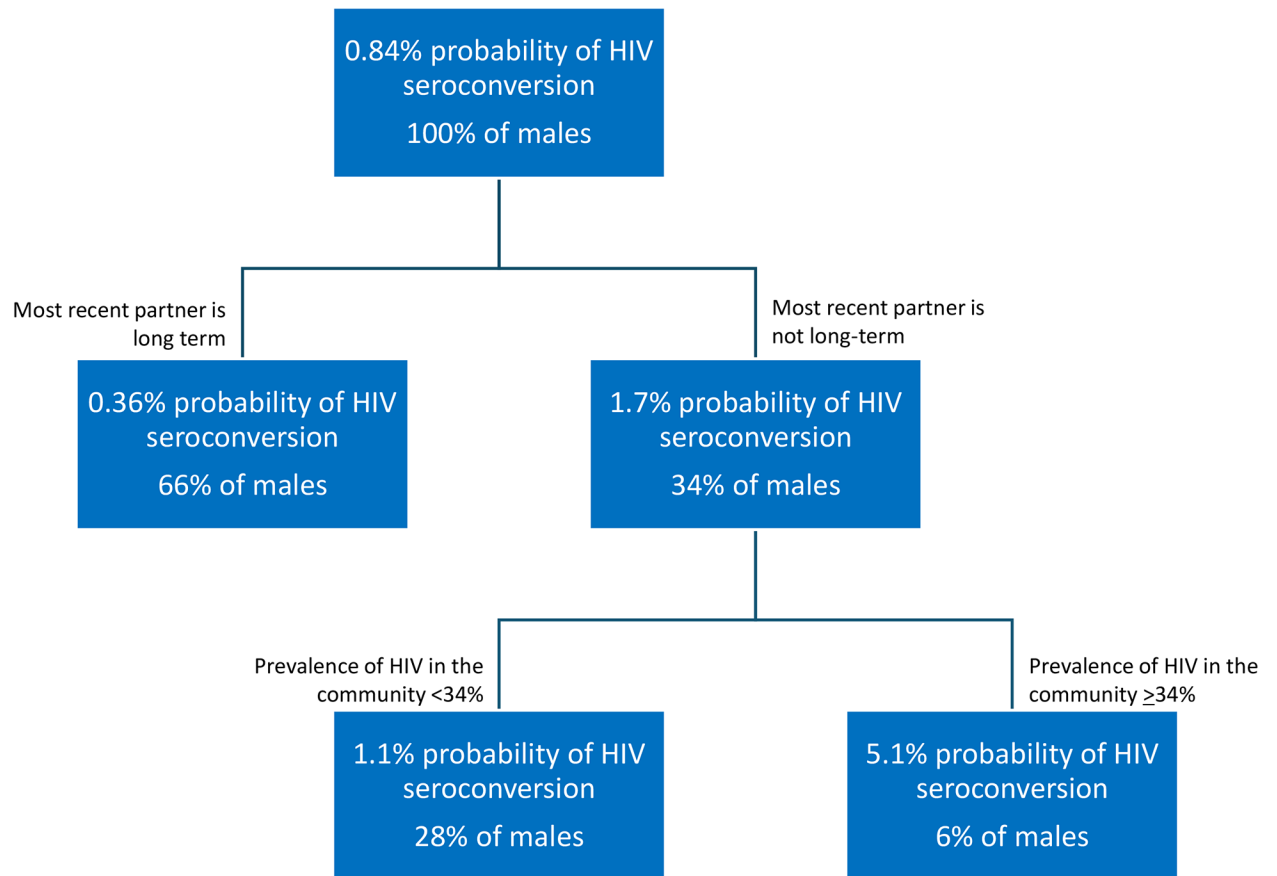


Figure 2.
Tree models for estimated probability of acquiring HIV for females (A) and males (B)

Table 1.

Odds ratios for variables predicting HIV seroconversion in females and males (with a focus on variables included in the tree models)

Predictor variable	Odds Ratio	95% Confidence Interval	P-value
Females			
Using injectable hormonal contraception (reference: not using injectable hormonal contraception)	2.4	1.9–2.9	<0.001
Had sex with their most recent partner more than 1.5 times in last 3 months, on average (reference: all others *)	0.9	0.7–1.2	0.37
Community HIV prevalence 20%–30% (reference <20%)	1.3	0.98–1.8	0.07
Community HIV prevalence 30%–50% (reference <20%)	1.7	1.3–2.3	<0.001
Males			
Most recent sexual partner was boyfriend/girlfriend (reference: most recent sexual partner was a long-term partner)	4.1	2.6–6.4	<0.001
Most recent sexual partner was casual ** (reference: most recent sexual partner was a long-term partner)	9.4	5.5–16.0	<0.001
Community HIV prevalence 20%–30% (reference <20%)	0.7	0.4–1.3	0.3
Community HIV prevalence 30%–50% (reference <20%)	2.2	1.3–3.9	.005

* “All others” reference group includes people having sex with their most recent partner ≤ 1.5 times in last 3 months and people who did not report having a sexual partner in the past year.

** Casual sexual partner defined in analysis as: casual partner who is known to the respondent; one-time partner; commercial sex worker; or other type of casual partner (predictor and reference variables both exclude boyfriend/girlfriend for this specific comparison).