Developing and validating a Bayesian clinical risk prediction model for three sexually transmitted infections in key populations from two Canadian provinces

Fio Vialard1,2, Qihuang Zhang1,3, Duncan Webster4, Stefanie Materniak4, Alexandre Dumont Blais5, Suma Nair6, Susan Bartlett2,7, Nitika Pant Pai1,2,7

1School of Population and Global Health, McGill University, Montreal, Canada.

2Centre for Outcome Research and Evaluation, Research Institute of the McGill University Health Centre, Montreal, Canada.

3Quantitative Life Sciences, McGill University, Montreal, Canada.

4Infectious Disease Research Unit, Saint John Regional Hospital, Saint John, Canada, Saint John, Canada.

5RÉZO, Montreal, Canada.

6School of Public Health, DY Patil University, Navi Mumbai, India.

7Department of Medicine, McGill University, Montreal, Canada.

Corresponding author: Dr. Nitika Pant Pai

Key words: risk prediction models; sexually transmitted and blood-borne infections;

key populations; Bayesian statistics; diagnostics.

Running title: Bayesian clinical risk prediction model for sexually transmitted and blood-borne infections.

Word count: 3000 (Max 3000)

Reference number: 30 (Max 30)

# Abstract

Word count: 300 (Max 300)

## Objectives

Across Canada, in the last decade, incidence rates of sexually transmitted and blood-borne infections (STBBI) have peaked (syphilis) or plateaued (hepatitis C virus [HCV] and human immunodeficiency virus [HIV]). Key populations (gay, bisexual, and other men who have sex with men, trans- and gender-diverse people, and people who use injection drugs) are at greater risk for these STBBIs, so correctly predicting risk before screening potentially infected individuals is crucial. We developed and validated a diagnostic clinical risk prediction model (CRPM) that estimates HIV, HCV, and syphilis risk for two key populations in two Canadian provinces.

## Methods

We used 20 variables and STBBI test results from a cross-sectional study evaluating multiplex testing (detection of co-infections) in New Brunswick and Quebec ( 400) to develop our CRPM. We randomly split the data into development ( 300) and validation ( 100) datasets using clinic-stratified sampling. We used Bayesian predictive projection with development data to select ranked STBBI predictors. We obtained the odds ratios (OR) of the highest performing model measured as area under the receiver operating curve (AUC), sensitivity, and specificity with 89% Credible Intervals (89%CrI) using validation data. Analyses were performed in R (v4.2.3).

## Results

Out of 400 participants, 73 were infected with HIV (n=16), HCV (n=60), and/or syphilis (n=5). An internally validated submodel with two predictors (*Past drug injection*, *Type of past sexually transmitted infection*) displayed the highest AUC (0.82 (89%CrI: 0.74-0.82)), sensitivity (0.83 (89%CrI: 0.76-0.88)), and specificity (0.36 (89%CrI: 0.21-0.57)). The predictor contributing most to STBBI risk was *Past drug injection* (OR = 7.62 (89%CrI: 4.41-13.07)).

## Conclusions

This Bayesian-based CRPM is the first to identify high-risk individuals for HIV, HCV, and syphilis with an overall good performance that minimizes case missing. After additional validation, it could serve as a promising novel tool for pre-screening key populations and improve Canadian STBBI multiplexed screening strategies.

# Key messages

## What is already known on this topic

* Clinical risk prediction models (CRPM) to identify individuals at higher risk of sexually transmitted and blood-borne infections (STBBI) have been developed and validated mostly for human immunodeficiency virus (HIV) and have shown variable performance.
* In Canada, where STBBI incidence rates are currently rising or plateauing, only two frequentist statistics-based STBBI CRPMs have been developed to date, and neither targeted people who use injection drugs, gay, bisexual, and other men who have sex with men, or trans and gender diverse individuals of Quebec or New Brunswick.

## What this study adds

* To our knowledge, we have developed and validated the first data-driven Bayesian statistics-based hepatitis C virus (HCV), HIV, and syphilis CRPM, which allowed us to obtain easily interpretable measures of uncertainty in the form of Credible Intervals (CrI).
* The final model had good overall performance (AUC: 0.82 (89%CrI: 0.74-0.82)) that achieved a high sensitivity (0.83 (89%CrI: 0.76-0.88)), despite having low specificity (0.36 (89%CrI: 0.21-0.57)). If implemented, it would reduce the probability of erroneously identifying high-risk individuals as low-risk, which is key in STBBI screening programs.

## How this study might affect research, practice, or policy

* We present a promising novel tool for pre-screening individuals at high-risk for HIV, syphilis, and/or HCV that, after additional validation, has the potential to improve STBBI screening strategies if it is implemented in combination with targeted testing strategies in two distinct key populations.

# Introduction

## Context and rationale

In Canada, sexually transmitted and blood-borne infection (STBBI) incidence rates have either plateaued (e.g. human immunodeficiency [HIV], hepatitis C virus [HCV]) or increased (e.g. syphilis) in the last decade, with an estimated 63,000 people currently living with HIV and 246,000 people with HCV.[1] Syphilis, caused by *Treponema pallidum*, is experiencing an epidemic with 30 cases per 100,000, corresponding to a 178% increase over 10 years.[1] Although survival attributed to treatment has improved, HIV was still responsible for 104 deaths in 2018 and HCV had a case fatality rate of 7.4 per 100,000.[1,2]

Additionally, individuals belonging to key populations are at greater risk of contracting STBBIs because they are often stigmatized and discriminated against, including when accessing healthcare services.[3] This situation increases their likelihood of avoiding healthcare services, where they would normally receive appropriate prevention, testing and treatment.[3] Consequently, up to 14 and 44% of people infected with HIV and HCV, respectively, are unaware of their status.[1] Gay, bisexual, and other men who have sex with men (gbMSMs) accounted for 43.8% of new HIV cases in 2020 and 71% of all male syphilis cases in 2021.[2,4] People who use injection drugs (PWID) accounted for 46.1% of HCV seroprevalence in 2019.[1] No such data is available for trans- and gender-diverse (TGD) individuals. In addition, co-infections, during which these pathogens interact in ways that influence disease susceptibility, progression, and treatment outcomes, are common.[1]

Standard asymptomatic (i.e. screening) and symptomatic testing strategies in Canada have not yet successfully reached these at-risk populations due to long turnaround time (TAT) and reliance on laboratory labour.[5–8] Although point-of-care (POC) technologies designed to facilitate comfort with testing or decrease TAT by accurately detecting multiple infections at once (e.g. multiplexed testing) are being introduced into screening programs, limitations based on linkage to care and reliance on confirmatory testing still remain.[5–8] Calls to develop, improve, and deploy these technologies have specified that to ensure effectiveness, they must be combined with methods to target those most at risk.[1]

In this context, we used secondary data from a study evaluating a HIV, HCV, and syphilis multiplexed screening strategy to develop and validate a diagnostic clinical risk prediction model (CRPM) PWID and gbMSMs, and TGD invidivuduals from two distinct provinces.[9] CRPMs are statistical tools that can be employed by physicians or patients to estimate the probability of a specific health outcome based on patient characteristics.[10] Current STBBI screening strategies would benefit from diagnostic CRPMs used as pre-screening tools to assist physicians in identifying those most at-risk.[11] Acceptable CPRM performance, corresponding to an area under the receiver operating curve (AUC; a combined measure of sensitivity and specificity) above 0.80, is crucial to minimize the number of high-risk individuals erroneously not encouraged to test (i.e. false negative) and of low-risk individuals erroneously encouraged to do so (i.e. false positive).[11] Existing STBBI CRPMs have mostly focused on HIV prediction and have shown variable performance (AUC: 0.49–0.89).[11–13] Researchers described CRPMs developed and/or validated for other STBBIs (i.e. CT/NG and HCV) that showed variable performance (AUC:0.49–0.99).[14–16] The vast majority of existing CRPMs have been developed using frequentist statistical methods with data-driven methods (i.e. Bayesian statistics and machine learning [ML]) being increasingly used.[16,17] Despite the predictive power of ML, its black-box nature often prevents intuitive result interpretation.[18]. We pursued a Bayesian framework that allows for a more intuitive interpretation of uncertainty through Credible Intervals (CrI).[19] We previously developed a Bayesian-based HIV CRPM for individuals living in townships in South Africa with an AUC of 0.71 (89%CrI: 0.68-0.72).[20] To our knowledge, only two STBBI CRPMs have been developed using Canadian data. Both employed frequentist methods, neither targeted key populations from these provinces, and only one was validated, highlighting limitations in CRPM data and methods.[21,22]

## Objective

Therefore, we aimed to develop and validate a diagnostic pre-screening CRPM that, if combined with multiplexed testing, could improve STBBI testing equity in Canadian key populations, especially PWID, gbMSMs, and TGD individuals..

# Methods

We followed the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement (Supplementary Checklist).[10]

## Data source

We recruited 401 from a cross-sectional study conducted at a community health clinic for gbMSMs and TGD individuals who have sex with men in Quebec (RÉZO; 237) and a harm-reduction clinic for PWID in New Brunswick (RECAP; 164), using convenience sampling.[9] We evaluated an app-assisted multiplexed testing strategy through which we screened participants for HIV, HCV, and syphilis using two immunological multiplexed tests and conducted confirmatory laboratory tests as per Canadian guidelines.[23] We extracted all available de-identified variables in the app’s pre-test questionnaire (predictors) and confirmatory test results (outcome). Predictors corresponded to relevant STBBI risk factors collected by nurses in the primary study. Details on inclusion/exclusion criteria, key study dates, sample size calculations, app interface, and laboratory testing, are reported elsewhere.[9]

## Model specifications

We used participant self-reported categorical data as predictors without applying variable pre-selection for model development. We cross-checked testing history data with clinic records and when we found discrepancies, clinic records information was used. The outcome of interest was STBBI status (positive or negative). It was determined by positive confirmatory test results for HIV, HCV, and/or syphilis (incident or existing acute or chronic infection) and negative confirmatory test results for all three. We identified *new syphilis, HCV,* or *HIV infections* in participants testing positive who had no or unknown HIV, syphilis, or HCV infection status history, excluding those that had spontaneously cleared or had been treated for their HCV infections, when the information was available. Study staff were blinded to the outcome during predictor collection and laboratory staff were blinded to predictors and outcomes during confirmatory testing. All model metadata used in our analysis are described in [Table 1](#thm-1). Additional variable handling details are in Supplementary methods and sensitivity analysis.

**Table 1** Clinical risk prediction model variables by category.

| **Parameters** | **Transformation** | **Imputation** |
| --- | --- | --- |
| **Category: Location** | | |
| Clinic.Location | None | None |
| **Category: Testing history** | | |
| TypePastSTD | 3-factor | Bayesian polytomous regression |
| PastSTD | Binary | Logistic regression |
| PastHIVTest, PastHCVTest, PastSyphilisTest, PastSyphilis | None | Bayesian polytomous regression |
| **Category: Individual** | | |
| Gender | None | Bayesian polytomous regression |
| Age | Binary | Logistic regression |
| **Category: Behavioural** | | |
| CurrentPartner | None | Logistic regression |
| SexualActivity, CondomUse, SexualPartners, PastInjectDrugs, ShareNeedles, Alcohol | Binary | Logistic regression |
| **Category: Socio-demographic** | | |
| Descent, EducationStatus, WorkStatus, MonthlyIncomeStatus | Binary | Logistic regression |
| **Category: Test results** | | |
| hcv.pos, hiv.pos, syphilis.pos | Binary (STBBI positive) | None |

## Missing data

We imputed 446 missing data from 77 individuals using *multiple imputation by chained equation* (MICE) implemented within the *mice* package.[24] We found no missing data for STBBI status. We excluded one individual from model development due to >50% missing data. We observed that missing data pattern for some variables depended on clinic ([Figure S 1](#lem-1)). Thus, we included *Clinic Location* as a dummy variable in the analysis and imputed data with the appropriate Bayesian regression ([Table 1](#thm-1)).[24]We imputed data in five datasets, evaluated its validity, and pooled values for the remainder of the analysis ([Figure S 2](#lem-2) & Supplementary methods and sensitivity analysis).

## Predictive projection

With the pooled imputed datasets, we conducted variable selection by predictive projection (i.e. projection of the posterior of the full model onto each candidate submodel), a robust method for high-dimensional datasets (i.e. relatively small sample sizes compared to the number of variables).[25] We separated our data into development (i.e. training; 300) and validation (i.e. testing; 100) datasets using clinic-stratified random sampling to ensure representation of both sites in each set. We then fit the development data with a full Bayesian logistic regression reference model containing all predictors ( 20) and STBBI status as the outcome variable.[25] By including *Clinic.Location* in the full model, we adjusted it to account for differences and confounding in each key population. We assigned horseshoe priors ( =0.02) for all coefficients, as described by Piironen et al.[25] We then selected the sparsest submodel with a performance equivalent to the full model, and reduced the CRPM complexity (Supplementary methods and sensitivity analysis).

## Model evaluation

We evaluated the predictive accuracy of the nine submodels with performances (i.e. estimated log predictive density [ELPD]) closest to the full model using a method independent from accuracy value thresholding. We computed sensitivity and specificity from 2000 draws of submodels posterior predictive distribution for *STBBI status*, and AUC from expected values of these draws. We quantified uncertainty using 89%CrIs to increase stability with effective sample sizes below 10,000 and to avoid confusion with 95% confidence intervals.[26] We selected the final model with the best AUC obtained with the validation dataset. We then computed the odds ratio (OR) and associated 89% and 95%CrIs from each coefficient in the final fitted submodel to inform the contribution of each predictor to the final model.

All analyses were done in R (v4.2.3).[27]

## Sensitivity analysis

In a sensitivity analysis, we examined differences in behaviour between incident and existing STBBI cases and compared results obtained with imputed data to those obtained with complete case data (n=323). We also compared the final CRPM with CRPMs developed with HIV/syphilis, HIV/HCV, and HCV as outcomes (*Supplementary methods and sensitivity analysis*).

# Results

## Participant recruitment and characteristics

In the original study, we recruited 400 from RÉZO ($n=n=$164), excluding the removed individual. Most participants (63.75%) self-identified as male, with 18.25% as female and 1.75% as transgender. Nearly half of all participants (49.25%) were 35-year-old and above. We identified 73 STBBIs, corresponding to 16 HIV, 5 syphilis and 60 HCV infections. Of these, 8 were co-infections (5 HCV/HIV and 3 HIV/syphilis) and 18 were new infections. Additional details about the study, visit procedures, multiplexed test results, and treatments are described elsewhere.[9]

Due to key population differences in density, risk factors, and needs, RECAP had a greater proportion of HCV infections (78.33%) while RÉZO had a greater proportion of HIV (100%) and syphilis (80%) infections. RECAP also had a greater proportion of individuals reporting using injection drugs in the past (29.75%) than RÉZO (5.5%). We also observed differences in gender, descent, education, income, and work status by clinic*.* The distribution of each characteristic by clinic is outlined in [Table S 1](#cor-1) for the development and validation datasets.

The outcome (i.e. *STBBI status*) was similarly distributed between the development (17% positive cases) and validation datasets (20% positive cases). The distributions of variables in the pooled imputed datasets ( 5) are shown in [Table S 2](#cor-2).

## The submodels reached the predictive performance of the full model with nine ranked variables

To develop a simple CRPM with a reduced number of predictors of STBBI infection, a sample size of 300 was sufficient to use predictive projection feature selection on a full model with all predictors ().[25] We found that most of the predictor’s posteriors in the full model had CrIs spanning the null due to horseshoe prior effects ([Figure S 3](#lem-3)).

Based on predictor ranking agreement between cross-validation folds ([Figure S 4](#lem-4)), we identified the 9 sparsest submodels with ELPD closest to the reference model. As expected, when we fitted each submodel with the development dataset, the AUC increased as the number of predictors increased, until reaching the value of the full model (AUC = 0.84 ;89%CrI: 0.81-0.86) ([Figure 1](#fig-1)). Submodels 2 to 9 had an acceptable AUC (>0.80). The sensitivity remained very similar, between 0.83 (89%CrI: 0.79-0.46) and 0.85 (89%CrI: 0.81-0.9)), and so did the specificity, between 0.33 (89%CrI: 0.22-0.46) and 0.41 (89%CrI: 0.29-0.53), across all submodels ([Figure S 5](#lem-5)).

## The submodel with two predictors had the best performance with the validation dataset

When we fit these submodels with the internal validation dataset ( 100), we found that the submodel with *Past drug injection* and *Type of past STI test* only had the best performance with an AUC of 0.82 (89%CrI: 0.74-0.82), a sensitivity of 0.83 (89%CrI: 0.76-0.88) and a specificity of 0.36 (89%CrI: 0.21-0.57) ([Figure 1](#fig-1) & [Figure S 5](#lem-5)). All other submodels had AUC values of 0.77 (89%CrI: 0.74-0.57) to 0.81 (89%CrI: 0.77-0.85) and similar values for sensitivity (0.81 (89%CrI: 0.74-0.57) to 0.83 (89%CrI: 0.76-0.88)) and specificity (0.36 (89%CrI: 0.14-0.57) to 0.43 (89%CrI: 0.21-0.64)).

The predictor with the strongest effect on STBBI risk (i.e. indicating a positive diagnosis) was *Past drug injection* (OR=7.62 (89%CrI: 4.41-13.07)) ([Figure 2](#fig-2)).

## Sensitivity analysis

We found little to no difference in behavioural variables’ distributions between incident and existing STBBI cases ([Table S 3](#cor-3)), that the imputation was successful, and some differences in prediction depending on the STBBI of interest (*Supplementary methods and sensitivity analysis*).

## Using the final model

To use the final model, physicians should input patient *Past drug injection* (i.e. Yes or No) and *Type of past STI test* (i.e. Specified, Unspecified, or None) information. The CRPM will then output a risk probability and they can decide whether or not to offer testing ([Table 2](#thm-2)).

**Table 2** Risk of sexually transmitted or blood-borne infection (STBBI) by predictor in the final model (p=2).

| **Past drug injection** | **Type of Past STI test** | **STBBI risk (89%CrI)** |
| --- | --- | --- |
| No | None | 0.06 (0.04-0.10) |
| Yes | None | 0.31 (0.15-0.59) |
| No | Unspecified | 0.19 (0.04-0.55) |
| No | Specified | 0.09 (0.04-0.26) |
| Yes | Unspecified | 0.64 (0.16-0.94) |
| Yes | Specified | 0.43 (0.15-0.82) |

# Discussion

## Interpretation

Overall, the results obtained with the validation dataset had a slightly lower performance and greater uncertainty than those obtained with the development dataset. A final CRPM with a nearly acceptable AUC (i.e. 0.79) is encouraging considering that it would be utilized as a pre-screening tool to identify high-risk individuals and offer them testing.[11] We obtained higher AUC measures than with a HIV CRPM that we previously developed using South African data, albeit greater uncertainty due to a smaller sample size.[20] In addition, these validation values are on the higher end of existing STBBI CRPMs and include uncertainty measures which are often lacking.[11–13,16,17] In addition, unlike with most existing ML models, sensitivity was always higher than specificity, resulting in fewer false positives than false negatives.[16]

PWID are known to be at higher risk of acquiring HCV.[1] Thus, due to numerous HCV positive results in our dataset (n=60), a CRPM with *Past drug injection* as the most important predictor performed as expected. It was also identified as an important predictor of STI risk in another existing Canadian CPRMs but its predicting ability was not quantified.[21] Predictor ranking only differed when we removed HCV cases, which may indicate that the final CRPM will perform better when a HCV infection is present.

## Strengths and limitations

To our knowledge, we developed and validated the first Bayesian statistics-based HCV, HIV, and syphilis CRPM. Using this approach allowed us to incorporate sparsity assumptions into our model with horseshoe priors and to output easily interpretable uncertainty probability measures (i.e. CrIs).[19] Using predictive projection, we obtained a simple model containing two predictors, which will be useful in rapidly identifying high-risk individuals who would benefit from STBBI testing. Unlike other previously developed CRPMs, we validated our CRPM internally using cross-validation with the development dataset and through random-splitting of our initial dataset.[10–14,16] In addition, the performance of candidate submodels based on draws of the posterior distribution allowed us to use a threshold-independent calculation of the sensitivity, specificity and AUC. Threshold-based measures of sensitivity and specificity can often be assigned arbitrarily and are decided based on very specific clinical context that can impact generalizability.[28]

Although a single joint model that predicts risk in two key populations with overlapping risk factors and infections has merit, the limited data source size obtained from two heterogeneous populations could have impacted its validity. To address the small sample size, we employed a method that generates robust models from scarce datasets.[25] To address the heterogeneity, we considered developing a separate model for each population. However, the limited sample size and numerous missing data in the RECAP dataset prevented this, because MICE is not robust with high proportion of missing values. Instead, we accounted for key populations differences with the *Clinic.Location* variable in each step of our analysis, including in the reference model, where it accounts for the inter-population heterogeneity. Grouping incident and existing cases as positive cases may have led to confounding bias from difference in behaviours between those with and without awareness of their status. However, the distribution of most behavioural variables did not differ between these two groups. In practice, a low-specificity CRPM could lead to labour, psychological, and monetary costs associated with low-risk individuals encouraged to test.[29] However, low specificity is preferable to low sensitivity by minimizing the number of missed high-risk individuals for targeted testing.[29] By sourcing data from a cross-sectional primary study, using convenience sampling conducted in urban centres, any bias resulting from the selection of participants that frequently visit these clinics may have impacted the generalizability of our results.[9] Despite using rigorous imputation methods, we are not guaranteed that missing data followed the same pattern as observed data. Additionally, residual confounding may have been introduced by the grouping of variables and from uncollected variables (e.g. partner STBBI status). The absence of methods to incorporate co-infection mechanisms in our final model may limit its predictive accuracy for co-infected individuals. We could consider a multivariable regression to jointly model the correlated risks of co-infections in the future. Finally, apart from testing history, for which we had clinic records available, app-based predictor data was self-reported from an interview with a nurse which may have been impacted by social-desirability bias. However, in practice, CRPM users will have to declare their information to the healthcare provider to receive their risk probability.

## Implications

Our results are aligned with Canadian elimination targets for three STBBIs and calls to improve current screening strategies.[1] By providing a validated CRPM specific to two key populations in two Canadian provinces, with acceptable performance, we are providing decision-makers with a novel pre-screening tool that can be used in combination with targeted testing strategies to improve multiplexed testing initiatives for these populations. Currently, any individual belonging to a key population is encouraged to test frequently, with less emphasis on their individual risk, and the monetary and time costs of all administered tests constitutes a major deterrent to upscaling initiatives. A simple CRPM that can identify those at truly greater HIV, HCV and/or syphilis risk within those populations while minimizing case missing can assist healthcare providers determine when to test on an individual basis and decrease overall costs. In the future, we plan to conduct additional validation and updating in other global populations that would benefit from targeted STBBI screening (e.g. rural populations in India), which we will source from larger homogenous study populations using a sampling method less prone to bias, with incident cases, more syphilis cases, and including mechanisms of co-infections, to confirm our model's performance in key populations and improve on the current limitations of our CRPM.

In addition, costs required for successful implementation and maintenance of a good performing CRPM do not guarantee an improved cost-effectiveness compared to current screening strategies.[30] Thus, our CRPM would benefit from a future cost-effectiveness analysis.

Finally, assigning risk to an individual can be stigmatizing and depends on the CRPM user’s willingness to share their information with healthcare providers. Therefore, our CRPM should be evaluated by members of the targeted community to ensure use of inclusive and appropriate language before implementation.

In conclusion, we have developed and validated a promising novel tool to pre-screen individuals at high-risk for three STBBIs with the potential to improve screening strategies. After additional validation and evaluation, the CRPM could be used to identify and encourage high-risk individuals to use multiplexed screening interventions for HCV and related co-infections.

# Declarations

Funding: This study was funded by the Canadian Institutes of Health Research (grant PJT153149). The study sponsor had no role in study design, collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the paper for publication. NPP acknowledges support from the Fonds de recherche du Québec - Santé (Senior Scientist award 252845, Distinguished Research Scholar award 324154) and the India-Canada Centre for Innovative Multidisciplinary Partnerships to Accelerate Community Transformation and Sustainability (IC-IMPACTS; CNO 3072).

FV is funded by a Fonds de recherche du Québec - Santé Master’s Training scholarship.

Conflicts: We have no conflict of interest to declare.

Ethics/Consent: We obtained ethics approvals from the Research Institute of the McGill University Health Centre (protocol number: 2020-6048) and from the Horizon Health Network Research Ethics Board at the Saint John Regional Hospital (protocol number: 2020-2963).

Data and materials: To protect participants privacy, individual-level data is not available at this time. Descriptive statistics of all relevant variables before and after imputation are available as supplementary material.

Code availability: All scripts that were generated for data wrangling, imputation, and modeling have been uploaded to GitHub and are publicly available (https://github.com/fiovialard/multiplex\_canada\_crpm).

Authors’ contributions: NPP conceived and designed the project. Data was obtained from a cross-sectional study evaluating multiplexed testing in Canada. DW, SM, and ADB were responsible for participant recruitment and data collection. FV cleaned, prepared, and imputed the data with guidance from QZ and NPP. FV developed and validated the reference model and submodels with guidance from QZ. QZ verified the codes and data. NPP, QZ, SB, and SN contributed to results and provided extensive feedback to contextualize findings. FV wrote the first manuscript draft and all authors read and approved the final manuscript.

Acknowledgements: The authors would like to thank Melisa Eraslan for her help with the manuscript uploading process and Cindy Leung Soo for her guidance on coding for risk prediction modeling. The authors also extend their gratitude to the participants of the AideSmart! study, who donated their time and information and the study staff, for meticulously collecting this data.

# References

# Figures

|  |
| --- |
| **Figure** 1: Area under the receiving operating characteristic curve (AUC) of nine sub-models and the full reference model (p=20) fitted with the development (blue; n=300) and validation (purple; n=100) datasets. We measured AUC based on the values of expected STBBI status based on 2000 samples drawn from the posterior predictive distributions. On the x-axis, each predictor named represents the submodel containing it and all predictors to its left. The reference model contains all available predictors. The AUC of the submodels fitted with the development (i.e. training) dataset increased by submodel size whereas the submodel containing two predictors had the highest performance when fitted with the validation (i.e. testing) dataset. Each point represents the median and each line represent 89% Credible Intervals of the variation in expected STBBI status values obtained from each predictor present in each model. |

|  |
| --- |
| **Figure** 2: Median, 89% (thick segment) and 95% Credible Intervals (thin segment) of estimated odds ratio (OR) for the intercept and each predictor of STBBI status in the final submodel (p=2) fitted with the validation (i.e. testing) set (n=100) based on 2000 draws of the predictive distribution. The vertical line represents null effect (OR=1.0) |

# Supplementary materials

## Supplementary TRIPOD checklist

[Will insert the supplemental checklist here.]

## Supplementary methods and sensitivity analysis

### Model specification details

We transformed all categorical (i.e. 2-6) variables into binary factors except for *Gender* and *Type of past STI test.* The gender of seven participants was misclassified as bisexual during the data gathering process which we re-classified as missing.

### Multiple imputation by chained equation details

To assess the legitimacy of the method, we examined chain convergence with 20 datasets and density plots of observed compared to imputed data, which we found to be acceptable ([Figure S 2](#lem-2)).

### Predictive projection details

Predictive projection is a Bayesian method that projects the posterior of a full model containing all variables onto multiple candidate submodels, to identify those with the highest-ranked predictors in each cross-validation and select the sparsest submodel with a performance equivalent to the full model. It was implemented within the *brms* and *projpred* packages.[31,32]

We refer to datasets used for training and testing our model as development and validation datasets to remain consistent with TRIPOD recommendations.[10]

We evaluated the reliability of MC estimates using Pareto-k diagnostic using leave-one out (LOO) cross-validation within the *loo* package.[33] We found acceptable Pareto-k values (i.e. >0.5) and convergence ( 1) for all predictors. We ran the model with four Monte Carlo (MC) chains of 2000 iterations each (1000 warm-up iterations).

We chose horseshoe priors because they are shrinking priors that can be used under the assumptions that most parameters will have small effects while others will have larger non-zero effect set by the scale parameter defined as: with as the number of non-zero predictor coefficient, as the sample size, and as the number of coefficients. Following recommendations from the authors, we specified .[25]

### Sensitivity analysis

To determine whether participants with existing cases of HCV, syphilis, and/or HIV behaved differently to those without knowledge of their infection status, we stratified the positive participants by incident or existing case and conducted a t-test of proportions for behavioural variables. We found no difference in behavioural parameters proportions between incident and existing HIV, HCV, or syphilis cases except for *Sexual activity* in incident HIV cases but this may have been due to our limited sample size ([Table S 3](#cor-3)).

The complete case analysis (n=323) had a similar distribution of predictors and the same number of infections as the imputed data. After model development, we observed differences in predictor ranking, but the two most important predictors were unchanged, suggesting a successful imputation. In addition, the highest performing submodel after validation was the same (i.e. submodel 2) with slightly higher AUC values (0.86 (89%CrI: 0.76-0.86)).

To understand the predictive relationship between the predictors and each STBBI of interest, we compared the final model obtained with STBBI as outcome to models developed using HIV/syphilis, HIV/HCV, or HCV as outcomes. In this analysis and the complete case analysis, we used a 5-fold instead of LOO cross-validation to decrease computational cost in the predictive projection step of model development. The results were similar when we restricted the outcome to HCV status or removed syphilis cases from the analysis, with the same highest performing submodel (AUC = 0.83 (89%CrI: 0.79-0.83) & AUC = 0.79 (89%CrI: 0.76-0.82)). With HIV and/or syphilis as the outcome, predictor ranking was completely different with *PastSyphilis* and *PastSyphilisTest* as the two most important predictors. The highest performing submodel was different and included seven variables (i.e. *PastSyphilis*, *PastSyphilisTest, Gender*, *TypePastSTI*, *Clinic.Location*, *Age*, *MonthlyIncomeStatus*).

### Supplementary references

Burkner, Piironen, Vehtari

## Supplementary tables

**Table S 1**

Participant characteristics distribution by community clinic in the development and validation dataset before imputation.

| dataset |  | RECAP (N=164) | REZO (N=236) | Total (N=400) |
| --- | --- | --- | --- | --- |
| Development | **SexualActivity** |  |  |  |
|  | 0-No or no wish to answer | 66 (53.7%) | 17 (9.6%) | 83 (27.7%) |
|  | 1-Yes | 55 (44.7%) | 159 (89.8%) | 214 (71.3%) |
|  | Missing | 2 (1.6%) | 1 (0.6%) | 3 (1.0%) |
|  | **CondomUse** |  |  |  |
|  | 0-Never | 53 (43.1%) | 25 (14.1%) | 78 (26.0%) |
|  | 1-Ever | 48 (39.0%) | 152 (85.9%) | 200 (66.7%) |
|  | Missing | 22 (17.9%) | 0 (0.0%) | 22 (7.3%) |
|  | **SexualPartners** |  |  |  |
|  | 0 to 5 | 113 (91.9%) | 136 (76.8%) | 249 (83.0%) |
|  | 6 or more | 1 (0.8%) | 39 (22.0%) | 40 (13.3%) |
|  | Missing | 9 (7.3%) | 2 (1.1%) | 11 (3.7%) |
|  | **PastSTD** |  |  |  |
|  | 0-No or do not know | 75 (61.0%) | 84 (47.5%) | 159 (53.0%) |
|  | 1-Yes | 41 (33.3%) | 88 (49.7%) | 129 (43.0%) |
|  | Missing | 7 (5.7%) | 5 (2.8%) | 12 (4.0%) |
|  | **TypePastSTD** |  |  |  |
|  | 0-HIV | 1 (0.8%) | 8 (4.5%) | 9 (3.0%) |
|  | 2-Herpes simplex | 3 (2.4%) | 0 (0.0%) | 3 (1.0%) |
|  | 3-Chlamydia | 22 (17.9%) | 28 (15.8%) | 50 (16.7%) |
|  | 4-Gonorrhoeae | 1 (0.8%) | 33 (18.6%) | 34 (11.3%) |
|  | 5-Syphilis | 0 (0.0%) | 9 (5.1%) | 9 (3.0%) |
|  | 6-Other | 13 (10.6%) | 6 (3.4%) | 19 (6.3%) |
|  | None | 76 (61.8%) | 83 (46.9%) | 159 (53.0%) |
|  | Missing | 7 (5.7%) | 10 (5.6%) | 17 (5.7%) |
|  | **PastInjectDrugs** |  |  |  |
|  | 0-No or no wish to answer | 31 (25.2%) | 159 (89.8%) | 190 (63.3%) |
|  | 1-Yes | 91 (74.0%) | 17 (9.6%) | 108 (36.0%) |
|  | Missing | 1 (0.8%) | 1 (0.6%) | 2 (0.7%) |
|  | **ShareNeedles** |  |  |  |
|  | 0-No | 65 (52.8%) | 168 (94.9%) | 233 (77.7%) |
|  | 1-Yes | 56 (45.5%) | 7 (4.0%) | 63 (21.0%) |
|  | Missing | 2 (1.6%) | 2 (1.1%) | 4 (1.3%) |
|  | **Alcohol** |  |  |  |
|  | 0-No | 95 (77.2%) | 72 (40.7%) | 167 (55.7%) |
|  | 1-5 times per week | 25 (20.3%) | 105 (59.3%) | 130 (43.3%) |
|  | Missing | 3 (2.4%) | 0 (0.0%) | 3 (1.0%) |
|  | **PastHIVTest** |  |  |  |
|  | 0-No | 34 (27.6%) | 8 (4.5%) | 42 (14.0%) |
|  | 1-Yes; less than 6 months ago | 29 (23.6%) | 82 (46.3%) | 111 (37.0%) |
|  | 2-Yes; 6 months to 1 year ago | 11 (8.9%) | 24 (13.6%) | 35 (11.7%) |
|  | 3-Yes; more than 1 year ago | 49 (39.8%) | 63 (35.6%) | 112 (37.3%) |
|  | Missing | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
|  | **PastHCVTest** |  |  |  |
|  | 0-No | 24 (19.5%) | 8 (4.5%) | 32 (10.7%) |
|  | 1-Yes; less than 6 months ago | 31 (25.2%) | 79 (44.6%) | 110 (36.7%) |
|  | 2-Yes; 6 months to 1 year ago | 13 (10.6%) | 24 (13.6%) | 37 (12.3%) |
|  | 3-Yes; more than 1 year ago | 55 (44.7%) | 65 (36.7%) | 120 (40.0%) |
|  | Missing | 0 (0.0%) | 1 (0.6%) | 1 (0.3%) |
|  | **PastSyphilisTest** |  |  |  |
|  | 0-No | 4 (3.3%) | 0 (0.0%) | 4 (1.3%) |
|  | 1-Yes; less than 6 months ago | 20 (16.3%) | 79 (44.6%) | 99 (33.0%) |
|  | 2-Yes; 6 months to 1 year ago | 12 (9.8%) | 25 (14.1%) | 37 (12.3%) |
|  | 3-Yes; more than 1 year ago | 86 (69.9%) | 72 (40.7%) | 158 (52.7%) |
|  | Missing | 1 (0.8%) | 1 (0.6%) | 2 (0.7%) |
|  | **PastSyphilis** |  |  |  |
|  | 0-No or do not know | 120 (97.6%) | 140 (79.1%) | 260 (86.7%) |
|  | 1-Yes | 3 (2.4%) | 35 (19.8%) | 38 (12.7%) |
|  | Missing | 0 (0.0%) | 2 (1.1%) | 2 (0.7%) |
|  | **Age** |  |  |  |
|  | 18 to 34 | 24 (19.5%) | 91 (51.4%) | 115 (38.3%) |
|  | 35 and above | 64 (52.0%) | 85 (48.0%) | 149 (49.7%) |
|  | Missing | 35 (28.5%) | 1 (0.6%) | 36 (12.0%) |
|  | **Gender** |  |  |  |
|  | 1-Male | 55 (44.7%) | 140 (79.1%) | 195 (65.0%) |
|  | 2-Female | 33 (26.8%) | 19 (10.7%) | 52 (17.3%) |
|  | 3-Transgender | 0 (0.0%) | 4 (2.3%) | 4 (1.3%) |
|  | 5-No response | 0 (0.0%) | 8 (4.5%) | 8 (2.7%) |
|  | Missing | 35 (28.5%) | 6 (3.4%) | 41 (13.7%) |
|  | **Descent** |  |  |  |
|  | Mediterranean European, African, Latin American, Caribbean and/or Asian | 17 (13.8%) | 66 (37.3%) | 83 (27.7%) |
|  | Non-Mediterranean European/North American | 71 (57.7%) | 110 (62.1%) | 181 (60.3%) |
|  | Missing | 35 (28.5%) | 1 (0.6%) | 36 (12.0%) |
|  | **EducationStatus** |  |  |  |
|  | 4-Post-graduate degree | 1 (0.8%) | 93 (52.5%) | 94 (31.3%) |
|  | No post-graduate degree | 87 (70.7%) | 83 (46.9%) | 170 (56.7%) |
|  | Missing | 35 (28.5%) | 1 (0.6%) | 36 (12.0%) |
|  | **WorkStatus** |  |  |  |
|  | Currently employed | 11 (8.9%) | 128 (72.3%) | 139 (46.3%) |
|  | Not currently employed | 77 (62.6%) | 48 (27.1%) | 125 (41.7%) |
|  | Missing | 35 (28.5%) | 1 (0.6%) | 36 (12.0%) |
|  | **MonthlyIncomeStatus** |  |  |  |
|  | $2000 CAD and above | 15 (12.2%) | 86 (48.6%) | 101 (33.7%) |
|  | 1-<$2;000 CAD | 73 (59.3%) | 90 (50.8%) | 163 (54.3%) |
|  | Missing | 35 (28.5%) | 1 (0.6%) | 36 (12.0%) |
|  | **CurrentPartner** |  |  |  |
|  | 0-No | 61 (49.6%) | 95 (53.7%) | 156 (52.0%) |
|  | 1-Yes | 27 (22.0%) | 81 (45.8%) | 108 (36.0%) |
|  | Missing | 35 (28.5%) | 1 (0.6%) | 36 (12.0%) |
|  | **hcv.pos** |  |  |  |
|  | Negative | 90 (73.2%) | 169 (95.5%) | 259 (86.3%) |
|  | Positive | 33 (26.8%) | 8 (4.5%) | 41 (13.7%) |
|  | **hiv.pos** |  |  |  |
|  | Negative | 123 (100.0%) | 164 (92.7%) | 287 (95.7%) |
|  | Positive | 0 (0.0%) | 13 (7.3%) | 13 (4.3%) |
|  | **syphilis.pos** |  |  |  |
|  | Negative | 122 (99.2%) | 173 (97.7%) | 295 (98.3%) |
|  | Positive | 1 (0.8%) | 4 (2.3%) | 5 (1.7%) |
|  | **new.hcv** |  |  |  |
|  | Negative | 112 (91.1%) | 177 (100.0%) | 289 (96.3%) |
|  | Positive | 11 (8.9%) | 0 (0.0%) | 11 (3.7%) |
|  | **new.syphilis** |  |  |  |
|  | Negative | 122 (99.2%) | 176 (99.4%) | 298 (99.3%) |
|  | Positive | 1 (0.8%) | 1 (0.6%) | 2 (0.7%) |
|  | **new.hiv** |  |  |  |
|  | Negative | 123 (100.0%) | 176 (99.4%) | 299 (99.7%) |
|  | Positive | 0 (0.0%) | 1 (0.6%) | 1 (0.3%) |
|  | **stbbi.pos** |  |  |  |
|  | Negative | 89 (72.4%) | 158 (89.3%) | 247 (82.3%) |
|  | Positive | 34 (27.6%) | 19 (10.7%) | 53 (17.7%) |
|  | **hiv.hcv.pos** |  |  |  |
|  | Negative | 123 (100.0%) | 174 (98.3%) | 297 (99.0%) |
|  | Positive | 0 (0.0%) | 3 (1.7%) | 3 (1.0%) |
|  | **hiv.syph.pos** |  |  |  |
|  | Negative | 123 (100.0%) | 174 (98.3%) | 297 (99.0%) |
|  | Positive | 0 (0.0%) | 3 (1.7%) | 3 (1.0%) |
|  | **hcv.syph.pos** |  |  |  |
|  | Negative | 123 (100.0%) | 177 (100.0%) | 300 (100.0%) |
| Validation | **SexualActivity** |  |  |  |
|  | 0-No or no wish to answer | 16 (39.0%) | 7 (11.9%) | 23 (23.0%) |
|  | 1-Yes | 23 (56.1%) | 52 (88.1%) | 75 (75.0%) |
|  | Missing | 2 (4.9%) | 0 (0.0%) | 2 (2.0%) |
|  | **CondomUse** |  |  |  |
|  | 0-Never | 12 (29.3%) | 8 (13.6%) | 20 (20.0%) |
|  | 1-Ever | 21 (51.2%) | 51 (86.4%) | 72 (72.0%) |
|  | Missing | 8 (19.5%) | 0 (0.0%) | 8 (8.0%) |
|  | **SexualPartners** |  |  |  |
|  | 0 to 5 | 33 (80.5%) | 42 (71.2%) | 75 (75.0%) |
|  | 6 or more | 4 (9.8%) | 16 (27.1%) | 20 (20.0%) |
|  | Missing | 4 (9.8%) | 1 (1.7%) | 5 (5.0%) |
|  | **PastSTD** |  |  |  |
|  | 0-No or do not know | 25 (61.0%) | 32 (54.2%) | 57 (57.0%) |
|  | 1-Yes | 13 (31.7%) | 27 (45.8%) | 40 (40.0%) |
|  | Missing | 3 (7.3%) | 0 (0.0%) | 3 (3.0%) |
|  | **TypePastSTD** |  |  |  |
|  | 0-HIV | 0 (0.0%) | 1 (1.7%) | 1 (1.0%) |
|  | 2-Herpes simplex | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
|  | 3-Chlamydia | 8 (19.5%) | 14 (23.7%) | 22 (22.0%) |
|  | 4-Gonorrhoeae | 0 (0.0%) | 9 (15.3%) | 9 (9.0%) |
|  | 5-Syphilis | 0 (0.0%) | 1 (1.7%) | 1 (1.0%) |
|  | 6-Other | 4 (9.8%) | 2 (3.4%) | 6 (6.0%) |
|  | None | 26 (63.4%) | 31 (52.5%) | 57 (57.0%) |
|  | Missing | 3 (7.3%) | 1 (1.7%) | 4 (4.0%) |
|  | **PastInjectDrugs** |  |  |  |
|  | 0-No or no wish to answer | 13 (31.7%) | 52 (88.1%) | 65 (65.0%) |
|  | 1-Yes | 28 (68.3%) | 5 (8.5%) | 33 (33.0%) |
|  | Missing | 0 (0.0%) | 2 (3.4%) | 2 (2.0%) |
|  | **ShareNeedles** |  |  |  |
|  | 0-No | 19 (46.3%) | 55 (93.2%) | 74 (74.0%) |
|  | 1-Yes | 21 (51.2%) | 1 (1.7%) | 22 (22.0%) |
|  | Missing | 1 (2.4%) | 3 (5.1%) | 4 (4.0%) |
|  | **Alcohol** |  |  |  |
|  | 0-No | 30 (73.2%) | 12 (20.3%) | 42 (42.0%) |
|  | 1-5 times per week | 10 (24.4%) | 47 (79.7%) | 57 (57.0%) |
|  | Missing | 1 (2.4%) | 0 (0.0%) | 1 (1.0%) |
|  | **PastHIVTest** |  |  |  |
|  | 0-No | 14 (34.1%) | 1 (1.7%) | 15 (15.0%) |
|  | 1-Yes; less than 6 months ago | 6 (14.6%) | 24 (40.7%) | 30 (30.0%) |
|  | 2-Yes; 6 months to 1 year ago | 3 (7.3%) | 9 (15.3%) | 12 (12.0%) |
|  | 3-Yes; more than 1 year ago | 18 (43.9%) | 24 (40.7%) | 42 (42.0%) |
|  | Missing | 0 (0.0%) | 1 (1.7%) | 1 (1.0%) |
|  | **PastHCVTest** |  |  |  |
|  | 0-No | 10 (24.4%) | 2 (3.4%) | 12 (12.0%) |
|  | 1-Yes; less than 6 months ago | 7 (17.1%) | 23 (39.0%) | 30 (30.0%) |
|  | 2-Yes; 6 months to 1 year ago | 2 (4.9%) | 9 (15.3%) | 11 (11.0%) |
|  | 3-Yes; more than 1 year ago | 22 (53.7%) | 25 (42.4%) | 47 (47.0%) |
|  | Missing | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
|  | **PastSyphilisTest** |  |  |  |
|  | 0-No | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
|  | 1-Yes; less than 6 months ago | 5 (12.2%) | 23 (39.0%) | 28 (28.0%) |
|  | 2-Yes; 6 months to 1 year ago | 2 (4.9%) | 8 (13.6%) | 10 (10.0%) |
|  | 3-Yes; more than 1 year ago | 34 (82.9%) | 28 (47.5%) | 62 (62.0%) |
|  | Missing | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
|  | **PastSyphilis** |  |  |  |
|  | 0-No or do not know | 40 (97.6%) | 49 (83.1%) | 89 (89.0%) |
|  | 1-Yes | 1 (2.4%) | 9 (15.3%) | 10 (10.0%) |
|  | Missing | 0 (0.0%) | 1 (1.7%) | 1 (1.0%) |
|  | **Age** |  |  |  |
|  | 18 to 34 | 12 (29.3%) | 29 (49.2%) | 41 (41.0%) |
|  | 35 and above | 19 (46.3%) | 29 (49.2%) | 48 (48.0%) |
|  | Missing | 10 (24.4%) | 1 (1.7%) | 11 (11.0%) |
|  | **Gender** |  |  |  |
|  | 1-Male | 17 (41.5%) | 43 (72.9%) | 60 (60.0%) |
|  | 2-Female | 14 (34.1%) | 7 (11.9%) | 21 (21.0%) |
|  | 3-Transgender | 0 (0.0%) | 3 (5.1%) | 3 (3.0%) |
|  | 5-No response | 0 (0.0%) | 3 (5.1%) | 3 (3.0%) |
|  | Missing | 10 (24.4%) | 3 (5.1%) | 13 (13.0%) |
|  | **Descent** |  |  |  |
|  | Mediterranean European, African, Latin American, Caribbean and/or Asian | 5 (12.2%) | 11 (18.6%) | 16 (16.0%) |
|  | Non-Mediterranean European/North American | 26 (63.4%) | 47 (79.7%) | 73 (73.0%) |
|  | Missing | 10 (24.4%) | 1 (1.7%) | 11 (11.0%) |
|  | **EducationStatus** |  |  |  |
|  | 4-Post-graduate degree | 0 (0.0%) | 26 (44.1%) | 26 (26.0%) |
|  | No post-graduate degree | 31 (75.6%) | 32 (54.2%) | 63 (63.0%) |
|  | Missing | 10 (24.4%) | 1 (1.7%) | 11 (11.0%) |
|  | **WorkStatus** |  |  |  |
|  | Currently employed | 3 (7.3%) | 46 (78.0%) | 49 (49.0%) |
|  | Not currently employed | 28 (68.3%) | 12 (20.3%) | 40 (40.0%) |
|  | Missing | 10 (24.4%) | 1 (1.7%) | 11 (11.0%) |
|  | **MonthlyIncomeStatus** |  |  |  |
|  | $2000 CAD and above | 4 (9.8%) | 29 (49.2%) | 33 (33.0%) |
|  | 1-<$2;000 CAD | 27 (65.9%) | 29 (49.2%) | 56 (56.0%) |
|  | Missing | 10 (24.4%) | 1 (1.7%) | 11 (11.0%) |
|  | **CurrentPartner** |  |  |  |
|  | 0-No | 24 (58.5%) | 33 (55.9%) | 57 (57.0%) |
|  | 1-Yes | 7 (17.1%) | 25 (42.4%) | 32 (32.0%) |
|  | Missing | 10 (24.4%) | 1 (1.7%) | 11 (11.0%) |
|  | **hcv.pos** |  |  |  |
|  | Negative | 27 (65.9%) | 54 (91.5%) | 81 (81.0%) |
|  | Positive | 14 (34.1%) | 5 (8.5%) | 19 (19.0%) |
|  | **hiv.pos** |  |  |  |
|  | Negative | 41 (100.0%) | 56 (94.9%) | 97 (97.0%) |
|  | Positive | 0 (0.0%) | 3 (5.1%) | 3 (3.0%) |
|  | **syphilis.pos** |  |  |  |
|  | Negative | 41 (100.0%) | 59 (100.0%) | 100 (100.0%) |
|  | Positive | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
|  | **new.hcv** |  |  |  |
|  | Negative | 40 (97.6%) | 57 (96.6%) | 97 (97.0%) |
|  | Positive | 1 (2.4%) | 2 (3.4%) | 3 (3.0%) |
|  | **new.syphilis** |  |  |  |
|  | Negative | 41 (100.0%) | 59 (100.0%) | 100 (100.0%) |
|  | Positive | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
|  | **new.hiv** |  |  |  |
|  | Negative | 41 (100.0%) | 58 (98.3%) | 99 (99.0%) |
|  | Positive | 0 (0.0%) | 1 (1.7%) | 1 (1.0%) |
|  | **stbbi.pos** |  |  |  |
|  | Negative | 27 (65.9%) | 53 (89.8%) | 80 (80.0%) |
|  | Positive | 14 (34.1%) | 6 (10.2%) | 20 (20.0%) |
|  | **hiv.hcv.pos** |  |  |  |
|  | Negative | 41 (100.0%) | 57 (96.6%) | 98 (98.0%) |
|  | Positive | 0 (0.0%) | 2 (3.4%) | 2 (2.0%) |
|  | **hiv.syph.pos** |  |  |  |
|  | Negative | 41 (100.0%) | 59 (100.0%) | 100 (100.0%) |
|  | Positive | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
|  | **hcv.syph.pos** |  |  |  |
|  | Negative | 41 (100.0%) | 59 (100.0%) | 100 (100.0%) |

**Table S 2** Participant characteristics distribution by community clinic in the development and validation dataset after imputation.

| dataset |  | RECAP (N=164) | REZO (N=236) | Total (N=400) |
| --- | --- | --- | --- | --- |
| Development | **SexualActivity** |  |  |  |
|  | 0-No or no wish to answer | 66 (53.7%) | 17 (9.6%) | 83 (27.7%) |
|  | 1-Yes | 57 (46.3%) | 160 (90.4%) | 217 (72.3%) |
|  | **CondomUse** |  |  |  |
|  | 0-Never | 59 (48.0%) | 25 (14.1%) | 84 (28.0%) |
|  | 1-Ever | 64 (52.0%) | 152 (85.9%) | 216 (72.0%) |
|  | **SexualPartners** |  |  |  |
|  | 0 to 5 | 120 (97.6%) | 137 (77.4%) | 257 (85.7%) |
|  | 6 or more | 3 (2.4%) | 40 (22.6%) | 43 (14.3%) |
|  | **PastSTD** |  |  |  |
|  | 0-No or do not know | 80 (65.0%) | 85 (48.0%) | 165 (55.0%) |
|  | 1-Yes | 43 (35.0%) | 92 (52.0%) | 135 (45.0%) |
|  | **TypePastSTD** |  |  |  |
|  | 0-HIV | 2 (1.6%) | 8 (4.5%) | 10 (3.3%) |
|  | 2-Herpes simplex | 4 (3.3%) | 0 (0.0%) | 4 (1.3%) |
|  | 3-Chlamydia | 23 (18.7%) | 30 (16.9%) | 53 (17.7%) |
|  | 4-Gonorrhoeae | 1 (0.8%) | 37 (20.9%) | 38 (12.7%) |
|  | 5-Syphilis | 0 (0.0%) | 11 (6.2%) | 11 (3.7%) |
|  | 6-Other | 13 (10.6%) | 6 (3.4%) | 19 (6.3%) |
|  | None | 80 (65.0%) | 85 (48.0%) | 165 (55.0%) |
|  | **PastInjectDrugs** |  |  |  |
|  | 0-No or no wish to answer | 31 (25.2%) | 160 (90.4%) | 191 (63.7%) |
|  | 1-Yes | 92 (74.8%) | 17 (9.6%) | 109 (36.3%) |
|  | **ShareNeedles** |  |  |  |
|  | 0-No | 66 (53.7%) | 169 (95.5%) | 235 (78.3%) |
|  | 1-Yes | 57 (46.3%) | 8 (4.5%) | 65 (21.7%) |
|  | **Alcohol** |  |  |  |
|  | 0-No | 98 (79.7%) | 72 (40.7%) | 170 (56.7%) |
|  | 1-5 times per week | 25 (20.3%) | 105 (59.3%) | 130 (43.3%) |
|  | **PastHIVTest** |  |  |  |
|  | 0-No | 34 (27.6%) | 8 (4.5%) | 42 (14.0%) |
|  | 1-Yes; less than 6 months ago | 29 (23.6%) | 82 (46.3%) | 111 (37.0%) |
|  | 2-Yes; 6 months to 1 year ago | 11 (8.9%) | 24 (13.6%) | 35 (11.7%) |
|  | 3-Yes; more than 1 year ago | 49 (39.8%) | 63 (35.6%) | 112 (37.3%) |
|  | **PastHCVTest** |  |  |  |
|  | 0-No | 24 (19.5%) | 8 (4.5%) | 32 (10.7%) |
|  | 1-Yes; less than 6 months ago | 31 (25.2%) | 79 (44.6%) | 110 (36.7%) |
|  | 2-Yes; 6 months to 1 year ago | 13 (10.6%) | 25 (14.1%) | 38 (12.7%) |
|  | 3-Yes; more than 1 year ago | 55 (44.7%) | 65 (36.7%) | 120 (40.0%) |
|  | **PastSyphilisTest** |  |  |  |
|  | 0-No | 4 (3.3%) | 0 (0.0%) | 4 (1.3%) |
|  | 1-Yes; less than 6 months ago | 21 (17.1%) | 79 (44.6%) | 100 (33.3%) |
|  | 2-Yes; 6 months to 1 year ago | 12 (9.8%) | 25 (14.1%) | 37 (12.3%) |
|  | 3-Yes; more than 1 year ago | 86 (69.9%) | 73 (41.2%) | 159 (53.0%) |
|  | **PastSyphilis** |  |  |  |
|  | 0-No or do not know | 120 (97.6%) | 141 (79.7%) | 261 (87.0%) |
|  | 1-Yes | 3 (2.4%) | 36 (20.3%) | 39 (13.0%) |
|  | **Age** |  |  |  |
|  | 18 to 34 | 40 (32.5%) | 91 (51.4%) | 131 (43.7%) |
|  | 35 and above | 83 (67.5%) | 86 (48.6%) | 169 (56.3%) |
|  | **Gender** |  |  |  |
|  | 1-Male | 71 (57.7%) | 145 (81.9%) | 216 (72.0%) |
|  | 2-Female | 44 (35.8%) | 19 (10.7%) | 63 (21.0%) |
|  | 3-Transgender | 3 (2.4%) | 4 (2.3%) | 7 (2.3%) |
|  | 5-No response | 5 (4.1%) | 9 (5.1%) | 14 (4.7%) |
|  | **Descent** |  |  |  |
|  | Mediterranean European, African, Latin American, Caribbean and/or Asian | 22 (17.9%) | 66 (37.3%) | 88 (29.3%) |
|  | Non-Mediterranean European/North American | 101 (82.1%) | 111 (62.7%) | 212 (70.7%) |
|  | **EducationStatus** |  |  |  |
|  | 4-Post-graduate degree | 6 (4.9%) | 94 (53.1%) | 100 (33.3%) |
|  | No post-graduate degree | 117 (95.1%) | 83 (46.9%) | 200 (66.7%) |
|  | **WorkStatus** |  |  |  |
|  | Currently employed | 16 (13.0%) | 128 (72.3%) | 144 (48.0%) |
|  | Not currently employed | 107 (87.0%) | 49 (27.7%) | 156 (52.0%) |
|  | **MonthlyIncomeStatus** |  |  |  |
|  | $2000 CAD and above | 20 (16.3%) | 86 (48.6%) | 106 (35.3%) |
|  | 1-<$2;000 CAD | 103 (83.7%) | 91 (51.4%) | 194 (64.7%) |
|  | **CurrentPartner** |  |  |  |
|  | 0-No | 87 (70.7%) | 95 (53.7%) | 182 (60.7%) |
|  | 1-Yes | 36 (29.3%) | 82 (46.3%) | 118 (39.3%) |
|  | **hcv.pos** |  |  |  |
|  | Negative | 90 (73.2%) | 169 (95.5%) | 259 (86.3%) |
|  | Positive | 33 (26.8%) | 8 (4.5%) | 41 (13.7%) |
|  | **hiv.pos** |  |  |  |
|  | Negative | 123 (100.0%) | 164 (92.7%) | 287 (95.7%) |
|  | Positive | 0 (0.0%) | 13 (7.3%) | 13 (4.3%) |
|  | **syphilis.pos** |  |  |  |
|  | Negative | 122 (99.2%) | 173 (97.7%) | 295 (98.3%) |
|  | Positive | 1 (0.8%) | 4 (2.3%) | 5 (1.7%) |
|  | **new.hcv** |  |  |  |
|  | Negative | 112 (91.1%) | 177 (100.0%) | 289 (96.3%) |
|  | Positive | 11 (8.9%) | 0 (0.0%) | 11 (3.7%) |
|  | **new.syphilis** |  |  |  |
|  | Negative | 122 (99.2%) | 176 (99.4%) | 298 (99.3%) |
|  | Positive | 1 (0.8%) | 1 (0.6%) | 2 (0.7%) |
|  | **new.hiv** |  |  |  |
|  | Negative | 123 (100.0%) | 176 (99.4%) | 299 (99.7%) |
|  | Positive | 0 (0.0%) | 1 (0.6%) | 1 (0.3%) |
|  | **stbbi.pos** |  |  |  |
|  | Negative | 89 (72.4%) | 158 (89.3%) | 247 (82.3%) |
|  | Positive | 34 (27.6%) | 19 (10.7%) | 53 (17.7%) |
|  | **hiv.hcv.pos** |  |  |  |
|  | Negative | 123 (100.0%) | 174 (98.3%) | 297 (99.0%) |
|  | Positive | 0 (0.0%) | 3 (1.7%) | 3 (1.0%) |
|  | **hiv.syph.pos** |  |  |  |
|  | Negative | 123 (100.0%) | 174 (98.3%) | 297 (99.0%) |
|  | Positive | 0 (0.0%) | 3 (1.7%) | 3 (1.0%) |
|  | **hcv.syph.pos** |  |  |  |
|  | Negative | 123 (100.0%) | 177 (100.0%) | 300 (100.0%) |
| Validation | **SexualActivity** |  |  |  |
|  | 0-No or no wish to answer | 17 (41.5%) | 7 (11.9%) | 24 (24.0%) |
|  | 1-Yes | 24 (58.5%) | 52 (88.1%) | 76 (76.0%) |
|  | **CondomUse** |  |  |  |
|  | 0-Never | 14 (34.1%) | 8 (13.6%) | 22 (22.0%) |
|  | 1-Ever | 27 (65.9%) | 51 (86.4%) | 78 (78.0%) |
|  | **SexualPartners** |  |  |  |
|  | 0 to 5 | 35 (85.4%) | 43 (72.9%) | 78 (78.0%) |
|  | 6 or more | 6 (14.6%) | 16 (27.1%) | 22 (22.0%) |
|  | **PastSTD** |  |  |  |
|  | 0-No or do not know | 25 (61.0%) | 32 (54.2%) | 57 (57.0%) |
|  | 1-Yes | 16 (39.0%) | 27 (45.8%) | 43 (43.0%) |
|  | **TypePastSTD** |  |  |  |
|  | 0-HIV | 0 (0.0%) | 1 (1.7%) | 1 (1.0%) |
|  | 2-Herpes simplex | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
|  | 3-Chlamydia | 11 (26.8%) | 14 (23.7%) | 25 (25.0%) |
|  | 4-Gonorrhoeae | 0 (0.0%) | 9 (15.3%) | 9 (9.0%) |
|  | 5-Syphilis | 0 (0.0%) | 1 (1.7%) | 1 (1.0%) |
|  | 6-Other | 4 (9.8%) | 2 (3.4%) | 6 (6.0%) |
|  | None | 26 (63.4%) | 32 (54.2%) | 58 (58.0%) |
|  | **PastInjectDrugs** |  |  |  |
|  | 0-No or no wish to answer | 13 (31.7%) | 54 (91.5%) | 67 (67.0%) |
|  | 1-Yes | 28 (68.3%) | 5 (8.5%) | 33 (33.0%) |
|  | **ShareNeedles** |  |  |  |
|  | 0-No | 19 (46.3%) | 58 (98.3%) | 77 (77.0%) |
|  | 1-Yes | 22 (53.7%) | 1 (1.7%) | 23 (23.0%) |
|  | **Alcohol** |  |  |  |
|  | 0-No | 31 (75.6%) | 12 (20.3%) | 43 (43.0%) |
|  | 1-5 times per week | 10 (24.4%) | 47 (79.7%) | 57 (57.0%) |
|  | **PastHIVTest** |  |  |  |
|  | 0-No | 14 (34.1%) | 1 (1.7%) | 15 (15.0%) |
|  | 1-Yes; less than 6 months ago | 6 (14.6%) | 24 (40.7%) | 30 (30.0%) |
|  | 2-Yes; 6 months to 1 year ago | 3 (7.3%) | 9 (15.3%) | 12 (12.0%) |
|  | 3-Yes; more than 1 year ago | 18 (43.9%) | 25 (42.4%) | 43 (43.0%) |
|  | **PastHCVTest** |  |  |  |
|  | 0-No | 10 (24.4%) | 2 (3.4%) | 12 (12.0%) |
|  | 1-Yes; less than 6 months ago | 7 (17.1%) | 23 (39.0%) | 30 (30.0%) |
|  | 2-Yes; 6 months to 1 year ago | 2 (4.9%) | 9 (15.3%) | 11 (11.0%) |
|  | 3-Yes; more than 1 year ago | 22 (53.7%) | 25 (42.4%) | 47 (47.0%) |
|  | **PastSyphilisTest** |  |  |  |
|  | 0-No | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
|  | 1-Yes; less than 6 months ago | 5 (12.2%) | 23 (39.0%) | 28 (28.0%) |
|  | 2-Yes; 6 months to 1 year ago | 2 (4.9%) | 8 (13.6%) | 10 (10.0%) |
|  | 3-Yes; more than 1 year ago | 34 (82.9%) | 28 (47.5%) | 62 (62.0%) |
|  | **PastSyphilis** |  |  |  |
|  | 0-No or do not know | 40 (97.6%) | 50 (84.7%) | 90 (90.0%) |
|  | 1-Yes | 1 (2.4%) | 9 (15.3%) | 10 (10.0%) |
|  | **Age** |  |  |  |
|  | 18 to 34 | 16 (39.0%) | 30 (50.8%) | 46 (46.0%) |
|  | 35 and above | 25 (61.0%) | 29 (49.2%) | 54 (54.0%) |
|  | **Gender** |  |  |  |
|  | 1-Male | 23 (56.1%) | 45 (76.3%) | 68 (68.0%) |
|  | 2-Female | 18 (43.9%) | 8 (13.6%) | 26 (26.0%) |
|  | 3-Transgender | 0 (0.0%) | 3 (5.1%) | 3 (3.0%) |
|  | 5-No response | 0 (0.0%) | 3 (5.1%) | 3 (3.0%) |
|  | **Descent** |  |  |  |
|  | Mediterranean European, African, Latin American, Caribbean and/or Asian | 7 (17.1%) | 11 (18.6%) | 18 (18.0%) |
|  | Non-Mediterranean European/North American | 34 (82.9%) | 48 (81.4%) | 82 (82.0%) |
|  | **EducationStatus** |  |  |  |
|  | 4-Post-graduate degree | 0 (0.0%) | 27 (45.8%) | 27 (27.0%) |
|  | No post-graduate degree | 41 (100.0%) | 32 (54.2%) | 73 (73.0%) |
|  | **WorkStatus** |  |  |  |
|  | Currently employed | 5 (12.2%) | 46 (78.0%) | 51 (51.0%) |
|  | Not currently employed | 36 (87.8%) | 13 (22.0%) | 49 (49.0%) |
|  | **MonthlyIncomeStatus** |  |  |  |
|  | $2000 CAD and above | 6 (14.6%) | 29 (49.2%) | 35 (35.0%) |
|  | 1-<$2;000 CAD | 35 (85.4%) | 30 (50.8%) | 65 (65.0%) |
|  | **CurrentPartner** |  |  |  |
|  | 0-No | 32 (78.0%) | 33 (55.9%) | 65 (65.0%) |
|  | 1-Yes | 9 (22.0%) | 26 (44.1%) | 35 (35.0%) |
|  | **hcv.pos** |  |  |  |
|  | Negative | 27 (65.9%) | 54 (91.5%) | 81 (81.0%) |
|  | Positive | 14 (34.1%) | 5 (8.5%) | 19 (19.0%) |
|  | **hiv.pos** |  |  |  |
|  | Negative | 41 (100.0%) | 56 (94.9%) | 97 (97.0%) |
|  | Positive | 0 (0.0%) | 3 (5.1%) | 3 (3.0%) |
|  | **syphilis.pos** |  |  |  |
|  | Negative | 41 (100.0%) | 59 (100.0%) | 100 (100.0%) |
|  | Positive | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
|  | **new.hcv** |  |  |  |
|  | Negative | 40 (97.6%) | 57 (96.6%) | 97 (97.0%) |
|  | Positive | 1 (2.4%) | 2 (3.4%) | 3 (3.0%) |
|  | **new.syphilis** |  |  |  |
|  | Negative | 41 (100.0%) | 59 (100.0%) | 100 (100.0%) |
|  | Positive | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
|  | **new.hiv** |  |  |  |
|  | Negative | 41 (100.0%) | 58 (98.3%) | 99 (99.0%) |
|  | Positive | 0 (0.0%) | 1 (1.7%) | 1 (1.0%) |
|  | **stbbi.pos** |  |  |  |
|  | Negative | 27 (65.9%) | 53 (89.8%) | 80 (80.0%) |
|  | Positive | 14 (34.1%) | 6 (10.2%) | 20 (20.0%) |
|  | **hiv.hcv.pos** |  |  |  |
|  | Negative | 41 (100.0%) | 57 (96.6%) | 98 (98.0%) |
|  | Positive | 0 (0.0%) | 2 (3.4%) | 2 (2.0%) |
|  | **hiv.syph.pos** |  |  |  |
|  | Negative | 41 (100.0%) | 59 (100.0%) | 100 (100.0%) |
|  | Positive | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
|  | **hcv.syph.pos** |  |  |  |
|  | Negative | 41 (100.0%) | 59 (100.0%) | 100 (100.0%) |

**Table S 3** Behavioural parameters distributions by incident compared to existing HIV, HCV, and syphilis cases p-values of a 2-way t-test of proportion conducted for each parameter.

HIV positive

|  | Incident (N=2) | Pre-existing (N=14) | Total (N=16) | p value |
| --- | --- | --- | --- | --- |
| **SexualActivity** |  |  |  | 0.006 |
| 0-No or no wish to answer | 1 (50.0%) | 0 (0.0%) | 1 (6.2%) |  |
| 1-Yes | 1 (50.0%) | 14 (100.0%) | 15 (93.8%) |  |
| **CondomUse** |  |  |  | 0.468 |
| 0-Never | 0 (0.0%) | 3 (21.4%) | 3 (18.8%) |  |
| 1-Ever | 2 (100.0%) | 11 (78.6%) | 13 (81.2%) |  |
| **SexualPartners** |  |  |  | 0.255 |
| N-Miss | 0 | 1 | 1 |  |
| 0 to 5 | 1 (50.0%) | 11 (84.6%) | 12 (80.0%) |  |
| 6 or more | 1 (50.0%) | 2 (15.4%) | 3 (20.0%) |  |
| **PastInjectDrugs** |  |  |  | 0.308 |
| 0-No or no wish to answer | 2 (100.0%) | 9 (64.3%) | 11 (68.8%) |  |
| 1-Yes | 0 (0.0%) | 5 (35.7%) | 5 (31.2%) |  |
| **ShareNeedles** |  |  |  | 0.551 |
| N-Miss | 0 | 1 | 1 |  |
| 0-No | 2 (100.0%) | 11 (84.6%) | 13 (86.7%) |  |
| 1-Yes | 0 (0.0%) | 2 (15.4%) | 2 (13.3%) |  |
| **Alcohol** |  |  |  | 0.182 |
| 0-No | 2 (100.0%) | 7 (50.0%) | 9 (56.2%) |  |
| 1-5 times per week | 0 (0.0%) | 7 (50.0%) | 7 (43.8%) |  |

HCV positive

|  | Incident (N=14) | Pre-existing (N=46) | Total (N=60) | p value |
| --- | --- | --- | --- | --- |
| **SexualActivity** |  |  |  | 0.967 |
| N-Miss | 0 | 1 | 1 |  |
| 0-No or no wish to answer | 6 (42.9%) | 19 (42.2%) | 25 (42.4%) |  |
| 1-Yes | 8 (57.1%) | 26 (57.8%) | 34 (57.6%) |  |
| **CondomUse** |  |  |  | 0.144 |
| N-Miss | 4 | 3 | 7 |  |
| 0-Never | 6 (60.0%) | 15 (34.9%) | 21 (39.6%) |  |
| 1-Ever | 4 (40.0%) | 28 (65.1%) | 32 (60.4%) |  |
| **SexualPartners** |  |  |  | 0.592 |
| N-Miss | 2 | 1 | 3 |  |
| 0 to 5 | 11 (91.7%) | 43 (95.6%) | 54 (94.7%) |  |
| 6 or more | 1 (8.3%) | 2 (4.4%) | 3 (5.3%) |  |
| **PastInjectDrugs** |  |  |  | 0.371 |
| N-Miss | 0 | 1 | 1 |  |
| 0-No or no wish to answer | 2 (14.3%) | 3 (6.7%) | 5 (8.5%) |  |
| 1-Yes | 12 (85.7%) | 42 (93.3%) | 54 (91.5%) |  |
| **ShareNeedles** |  |  |  | 0.391 |
| N-Miss | 2 | 2 | 4 |  |
| 0-No | 6 (50.0%) | 16 (36.4%) | 22 (39.3%) |  |
| 1-Yes | 6 (50.0%) | 28 (63.6%) | 34 (60.7%) |  |
| **Alcohol** |  |  |  | 0.817 |
| N-Miss | 0 | 1 | 1 |  |
| 0-No | 11 (78.6%) | 34 (75.6%) | 45 (76.3%) |  |
| 1-5 times per week | 3 (21.4%) | 11 (24.4%) | 14 (23.7%) |  |

Syphilis positive

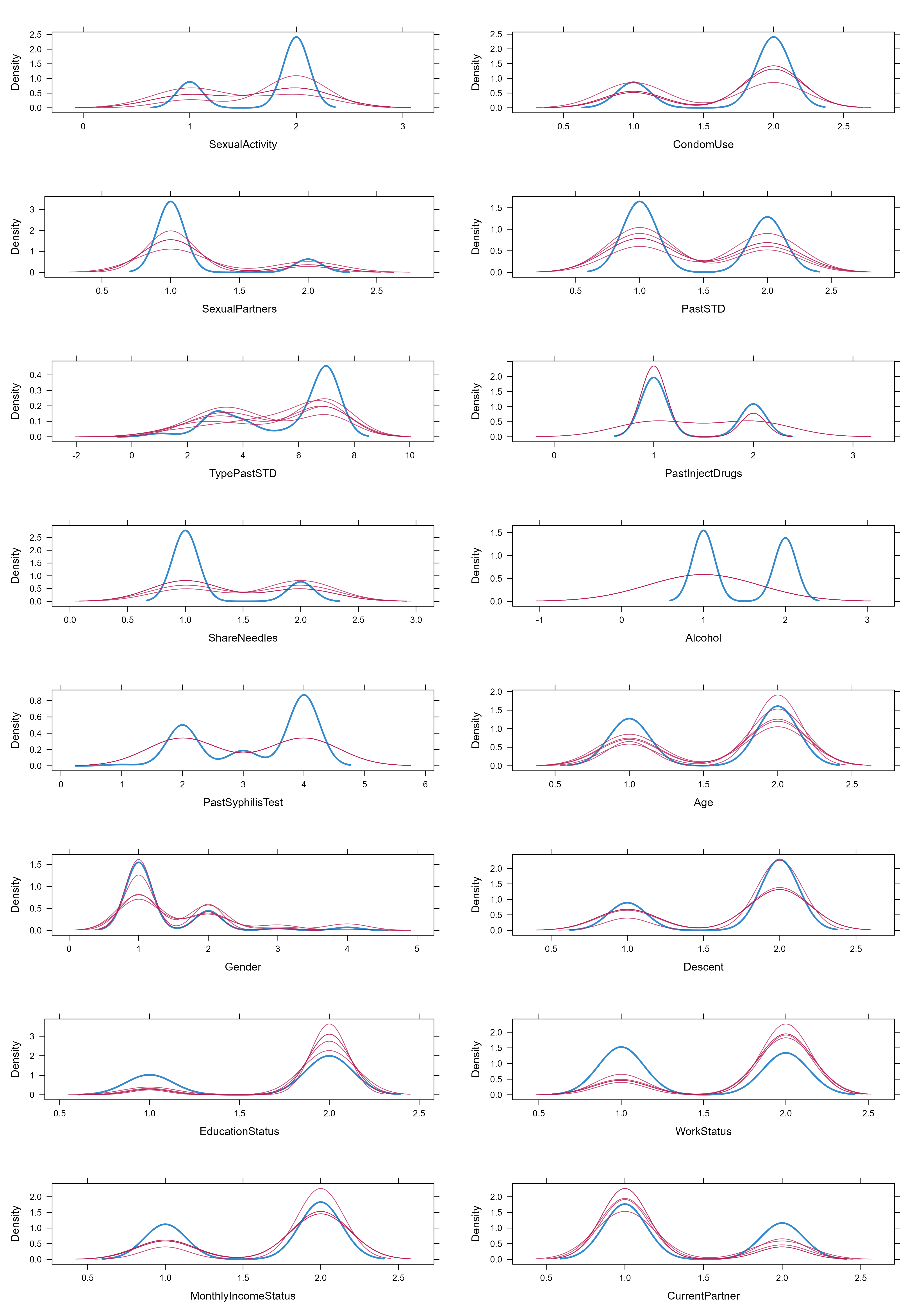
|  | Incident (N=2) | Pre-exisiting (N=3) | Total (N=5) | p value |
| --- | --- | --- | --- | --- |
| **SexualActivity** |  |  |  | 0.655 |
| 1-Yes | 2 (100.0%) | 3 (100.0%) | 5 (100.0%) |  |
| **CondomUse** |  |  |  | 0.655 |
| 1-Ever | 2 (100.0%) | 3 (100.0%) | 5 (100.0%) |  |
| **SexualPartners** |  |  |  | 0.386 |
| N-Miss | 1 | 1 | 2 |  |
| 0 to 5 | 1 (100.0%) | 1 (50.0%) | 2 (66.7%) |  |
| 6 or more | 0 (0.0%) | 1 (50.0%) | 1 (33.3%) |  |
| **PastInjectDrugs** |  |  |  | 0.655 |
| 0-No or no wish to answer | 2 (100.0%) | 3 (100.0%) | 5 (100.0%) |  |
| **ShareNeedles** |  |  |  | 0.655 |
| 0-No | 2 (100.0%) | 3 (100.0%) | 5 (100.0%) |  |
| **Alcohol** |  |  |  | 0.709 |
| 0-No | 1 (50.0%) | 2 (66.7%) | 3 (60.0%) |  |
| 1-5 times per week | 1 (50.0%) | 1 (33.3%) | 2 (40.0%) |  |

## Supplementary figures

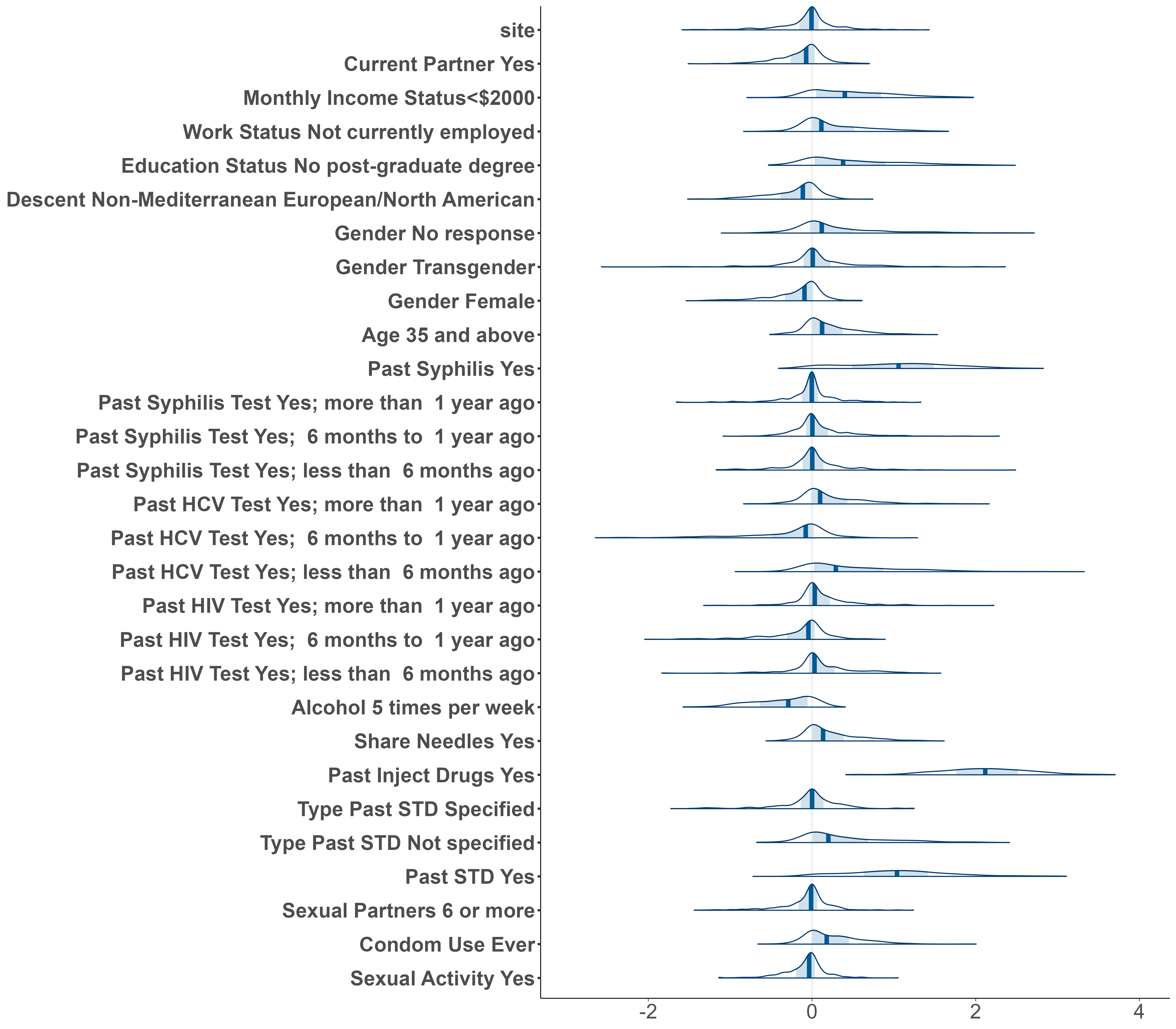
**Figure S 1** Missing data for each model parameter showed a pattern that differed per site. Each panel represents the frequency of missing and non-missing data for the REZO compared to the RECAP site. The labels represent the significance of a 2-way t-test of missingness proportion conducted for each parameter. Significant results (i.e <0.05) are highlighted in gold.



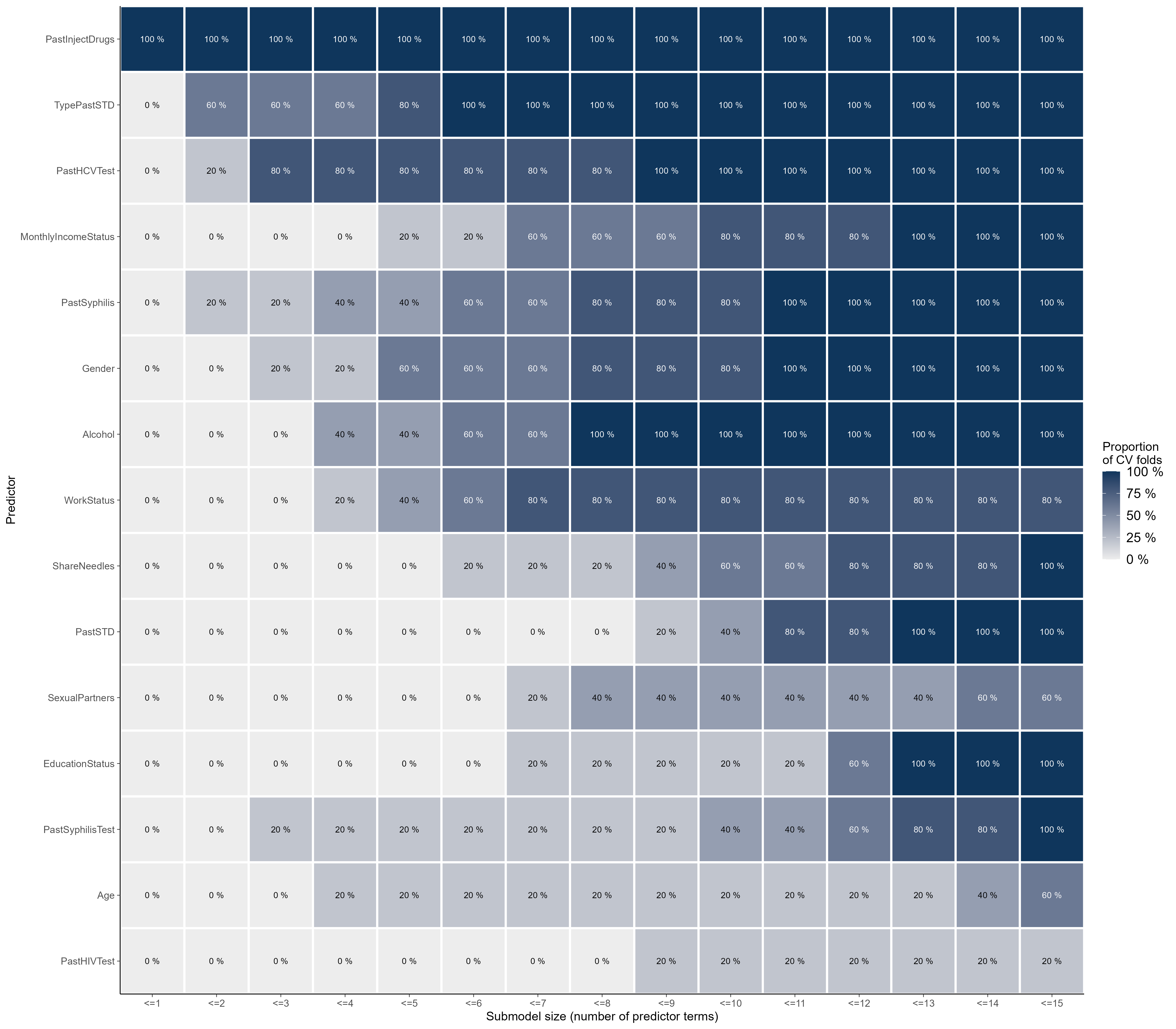
**Figure S 2** Density plot of observed values of model parameters (blue) and their corresponding imputed values (red). Data was imputed into five datasets using Bayesian logistic regression and Bayesian polytomous regression for binary and multi-categorical (>2) parameters, respectively. The imputation successfully reproduced observed data distribution under the assumption that unobserved data would have been the same as observed data.



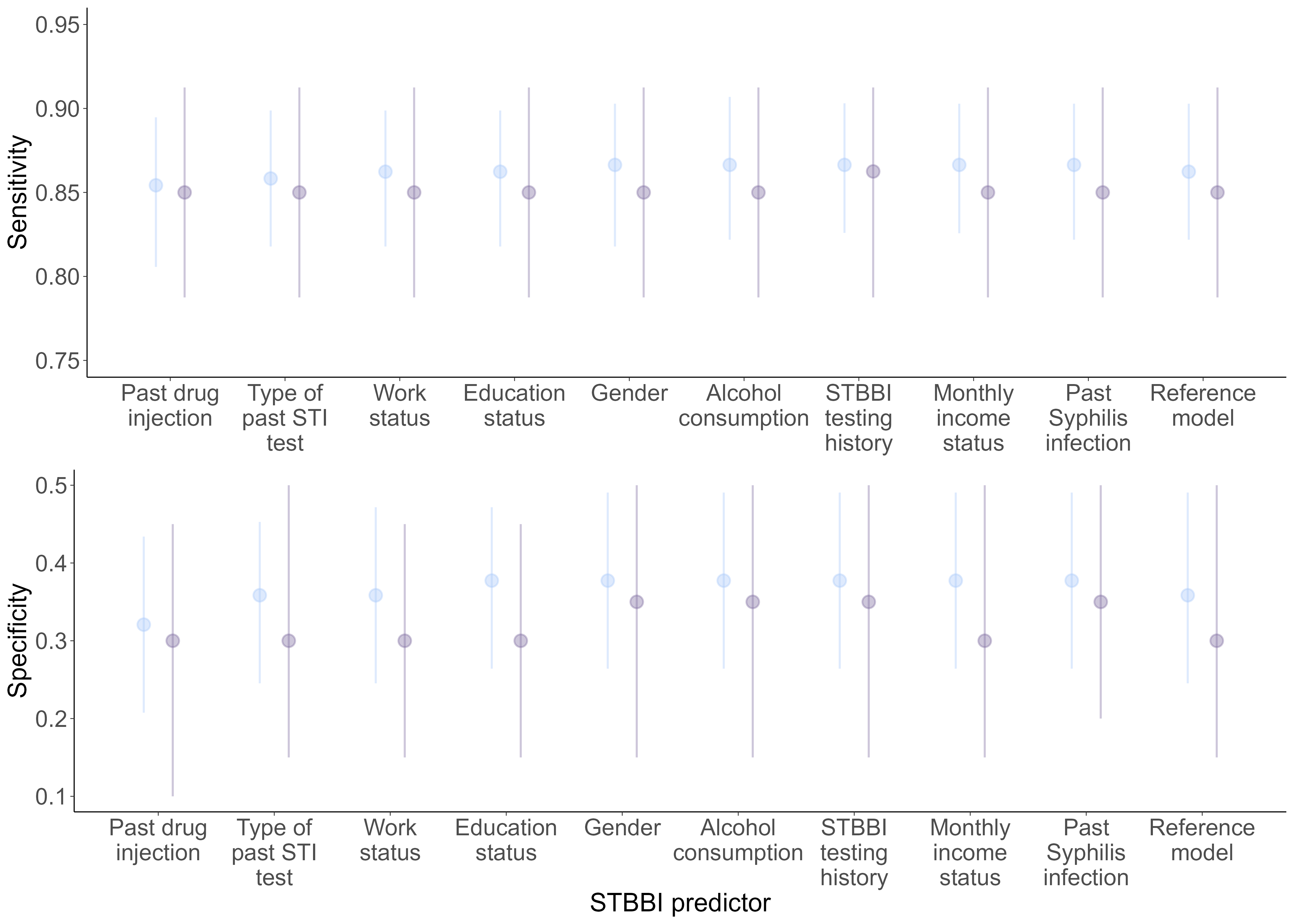
**Figure S 3** Posterior predictive distribution of the full model () fitted with the development (i.e. training) dataset (n=300) for which we assigned horseshoe priors () to the coefficient of each predictor in the model. Horseshoe priors shrunk most unimportant predictors to values close to 0.



**Figure S 4** Cumulative agreement of predictor ranking in each cross-validation fold of the predictive projection using a solution path method. Each cell represents the position of the predictor (y-axis) for a submodel of a specific size (x-axis). *PastInjectDrugs* was identified as the first predictor of all submodels in each leave-one-out cross-validation fold.



**Figure S 5** Sensitivity and specificity of nine sub-models and the full reference model (p=20) fitted with the development (i.e. training; blue; n=300) and validation (i.e. testing; purple; n=100) datasets. We measured sensitivity and specificity based on 2000 samples of STBBI status drawn from the posterior predictive distributions for each model. On the x-axis, each predictor named represents the submodel containing it and all predictors to its left. The reference model contains all available predictors. Sensitivity and specificity are similar across models for both datasets. Each point represents the median and each line represents the 89% Credible Intervals of the predictive distribution



1 Accelerating our response: Government of canada five-year action plan on sexually transmitted and blood-borne infections. 2019. <https://www.canada.ca/en/public-health/services/reports-publications/accelerating-our-response-five-year-action-plan-sexually-transmitted-blood-borne-infections.html>

2 Estimates of HIV incidence, prevalence and canada’s progress on meeting the 90-90-90 HIV targets, 2020. 2022. <https://www.canada.ca/en/public-health/services/publications/diseases-conditions/estimates-hiv-incidence-prevalence-canada-meeting-90-90-90-targets-2020.html>

3 MacClean R. [Resources to address stigma related to sexuality, substance use and sexually transmitted and blood-borne infections](https://doi.org/10.14745/ccdr.v44i02a05). *Canada Communicable Disease Report*. 2018;44:62–7.

4 Aho J, Lybeck C, Tetteh A, *et al.* [Rising syphilis rates in canada, 2011-2020](https://doi.org/10.14745/ccdr.v48i23a01). *Can Commun Dis Rep*. 2022;48:52–60.

5 HIV self-testing. 2023;2024. <https://www.catie.ca/hiv-self-testing-0>

6 Harrigan M. Dried blood spot collection for hepatitis c and HIV testing in canada. 2021;2024. <https://www.catie.ca/prevention-in-focus/dried-blood-spot-collection-for-hepatitis-c-and-hiv-testing-in-canada>

7 Karellis A, Naeem F, Nair S, *et al.* [Multiplexed rapid technologies for sexually transmitted infections: A systematic review](https://doi.org/10.1016/s2666-5247(21)00191-9). *Lancet Microbe*. 2022;3:e303–15.

8 Miller D. Research update: Rapid HIV and syphilis test now approved for clinical use. 2023;2024. <https://www.catie.ca/prevention-in-focus/research-update-rapid-hiv-and-syphilis-test-now-approved-for-clinical-use#>

9 Karellis A, Webster D, Boulanger J, *et al.* [Feasibility, preference, and impact of a rapid multiplexed point-of-care digital innovation (AideSmart!) to expedite screening of sexually transmitted blood-borne infections in at-risk populations in canada: Cross-sectional study](https://doi.org/10.2196/55075). *J Med Internet Res*. 2024;26:e55075.

10 Collins GS, Reitsma JB, Altman DG, *et al.* [Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement](https://doi.org/10.1186/s12916-014-0241-z). *BMC Medicine*. 2015;13:1.

11 Ong JJ, Coulthard K, Quinn C, *et al.* [Risk-based screening tools to optimise HIV testing services: A systematic review](https://doi.org/10.1007/s11904-022-00601-5). *Current HIV/AIDS Reports*. 2022;19:154–65.

12 Jia KM, Eilerts H, Edun O, *et al.* [Risk scores for predicting HIV incidence among adult heterosexual populations in sub-saharan africa: A systematic review and meta-analysis](https://doi.org/10.1002/jia2.25861). *J Int AIDS Soc*. 2022;25:e25861.

13 Luo Q, Luo Y, Cui T, *et al.* [Performance of HIV infection prediction models in men who have sex with men: A systematic review and meta-analysis](https://doi.org/10.1007/s10508-023-02574-x). *Arch Sex Behav*. 2023;52:2011–23.

14 Falasinnu T, Gilbert M, Gustafson P, *et al.* [Deriving and validating a risk estimation tool for screening asymptomatic chlamydia and gonorrhea](https://doi.org/10.1097/olq.0000000000000205). *Sex Transm Dis*. 2014;41:706–12.

15 Cantor A, Dana T, Griffin JC, *et al.* [Screening for chlamydial and gonococcal infections: Updated evidence report and systematic review for the US preventive services task force](https://doi.org/10.1001/jama.2021.10577). *Jama*. 2021;326:957–66.

16 Moulaei K, Sharifi H, Bahaadinbeigy K, *et al.* [Machine learning for prediction of viral hepatitis: A systematic review and meta-analysis](https://doi.org/10.1016/j.ijmedinf.2023.105243). *Int J Med Inform*. 2023;179:105243.

17 Xu X, Yu Z, Ge Z, *et al.* [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](https://doi.org/10.2196/37850). *J Med Internet Res*. 2022;24:e37850.

18 Nichols JA, Chan HWH, Baker MAB. [Machine learning: Applications of artificial intelligence to imaging and diagnosis](https://doi.org/10.1007/s12551-018-0449-9). *Biophys Rev*. 2019;11:111–8.

19 Schoot R van de, Depaoli S, King R, *et al.* [Bayesian statistics and modelling](https://doi.org/10.1038/s43586-020-00001-2). *Nature Reviews Methods Primers*. 2021;1:1.

20 Soo CL, Bhatnagar S, Bartlett SJ, *et al.* [Development and evaluation of a digital HIV risk assessment tool incorporated within an app-based self-testing program](https://doi.org/10.1097/qai.0000000000003210). *J Acquir Immune Defic Syndr*. 2023;93:387–94.

21 O’Byrne P, Musten A, Orser L, *et al.* [Automated STI/HIV risk assessments: Testing an online clinical algorithm in ottawa, canada](https://doi.org/10.1177/09564624211031322). *Int J STD AIDS*. 2021;32:1365–73.

22 Ablona A, Falasinnu T, Irvine M, *et al.* [Validation of a clinical prediction rule to predict asymptomatic chlamydia and gonorrhea infections among internet-based testers](https://doi.org/10.1097/olq.0000000000001340). *Sex Transm Dis*. 2021;48:481–7.

23 Guides for health professionals. Canadian Guidelines for Sexually transmitted and blood-borne infections: 2024. <https://www.canada.ca/en/public-health/services/infectious-diseases/sexual-health-sexually-transmitted-infections/canadian-guidelines.html>

24 Buuren SV. *Flexible imputation of missing data. Second edition*. Chapman & Hall/CRC Press 2018.

25 Piironen J, Paasiniemi M, Aki V. [Projective inference in high-dimensional problems: Prediction and feature selection](https://doi.org/10.1214/20-EJS1711). *Electronic Journal of Statistics*. 2020;14:2155–97.

26 Kruschke J. Doing bayesian data analysis : A tutorial introduction with r. 2014. <https://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=5754481>

27 Team RC. R: A language and environment for statistical computing. 2021. <https://www.R-project.org/>

28 Wynants L, Smeden M van, McLernon DJ, *et al.* [Three myths about risk thresholds for prediction models](https://doi.org/10.1186/s12916-019-1425-3). *BMC Medicine*. 2019;17:192.

29 Screening programmes: A short guide. Increase effectiveness, maximize benefits and minimize harm. 2020. <https://iris.who.int/handle/10665/330829>

30 Giessen A van, Peters J, Wilcher B, *et al.* [Systematic review of health economic impact evaluations of risk prediction models: Stop developing, start evaluating](https://doi.org/10.1016/j.jval.2017.01.001). *Value Health*. 2017;20:718–26.

31 Bürkner P-C. [Bayesian item response modeling in r with brms and stan](https://doi.org/10.18637/jss.v100.i05). *Journal of Statistical Software*. 2021;100:1–54.

32 Piironen J, Paasiniemi M, Catalina A, *et al.* Projpred: Projection predictive feature selection. 2023. <https://mc-stan.org/projpred/>

33 Vehtari A, Gabry J, Paasiniemi M, *et al.* Efficient leave-one-out cross-validation and WAIC for bayesian models. 2023. <https://mc-stan.org/loo/>