HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2021 October 30.

Published in final edited form as:

Stat Med. 2020 October 30; 39(24): 3255–3271. doi:10.1002/sim.8661.

Efficient estimation of HIV incidence rate using a pooled cross-sectional cohort study design

Kesaobaka Molebatsi[1,4], **Lesego Gabaitiri**[1], **Lucky Mokgatlhe**[1], **Sikhulile Moyo**[4], **Simani Gaseitsiwe**[4], **Kathleen E. Wirth**[3], **Victor DeGruttola**[3], **Eric Tchetgen Tchetgen**[*,2]

[1]Department of Statistics, University of Botswana, Gaborone, Botswana

[2]The Wharton School, University of Pennsylvania, Philadelphia, USA

[3]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

[4]Botswana Harvard AIDS Institute Partnership, Gaborone, Botswana

Summary

Development of methods to accurately estimate HIV incidence rate remains a challenge. Ideally, one would follow a random sample of HIV-negative individuals under a longitudinal study design and identify incident cases as they arise. Such designs can be prohibitively resource intensive and therefore alternative designs may be preferable. We propose such a simple, less resource-intensive study design and develop a weighted log likelihood approach which simultaneously accounts for selection bias and outcome misclassification error. The design is based on a cross-sectional survey which queries individuals' time since last HIV-negative test, validates their test results with formal documentation whenever possible, and tests all persons who do not have documentation of being HIV-positive. To gain efficiency, we update the weighted log likelihood function with potentially misclassified self-reports from individuals who could not produce documentation of a prior HIV-negative test and investigate large sample properties of validated sub-sample only versus pooled sample estimators through extensive Monte Carlo simulations. We illustrate our method by estimating incidence rate for individuals who tested HIV-negative within 1.5 and 5 years prior to Botswana Combination Prevention Project enrolment. This paper establishes that accurate estimates of HIV incidence rate can be obtained from individuals' history of testing in a cross-sectional cohort study design by appropriately accounting for selection bias and misclassification error. Moreover, this approach is notably less resource-intensive compared to longitudinal and laboratory-based methods.

Keywords

Cross-sectional Cohort; Incidence rate; Misclassification error; Selection bias; Weighted log Likelihood

[*]**Correspondence** Eric Tchetgen Tchetgen, etchetgen@gmail.com.

Conflict of interest

The authors declare no potential conflict of interests.

Data availability

The BCPP enrolment dataset is available upon request from BCPP team. Contact Person: Molly Pretorius Holme, mpretori@hsph.harvard.edu.

1 | INTRODUCTION

Incidence, the rate of new cases of a disease within a specified period of time,[1] plays a key role in understanding dynamics of an epidemic and in evaluating the impact of public health interventions to control its spread. In the case of Human Immunodeficiency Virus (HIV) in Sub-Saharan Africa, obtaining reliable incidence estimates poses several challenges including a lengthy incubation time resulting in slow accumulation of data, lack of precise estimates for incubation period and false recency rates among others.[23] A direct approach is to estimate HIV incidence rate from longitudinal studies where a representative sample of disease free individuals is followed over time and new cases recorded. Incidence rate is then computed as the number of persons newly infected with HIV during a specified time period to the cumulative person-time at risk of infection.[4] However, cohort studies are resource intensive, time consuming, and also subject to selection bias if retention is low.[5] To avoid such problems, certain methods have been developed which estimate incidence rate from cross-sectional data. Among these methods are assay-based techniques which look at levels and proportions of certain antibodies in the blood sample to show whether the infection was recent or has been present for some time. These methods have been justified by maximum likelihood criteria and improved by incorporation of past prevalence and false recency rates on incidence estimates.[6782] Other approaches are based on mathematical models that decompose observed changes in prevalence between two sero-surveys into contributions of new infections and mortality assuming incidence remains constant between surveys.[910] However, these methods are sensitive to the use of anti retro-viral therapy (ART), hence with a global commitment of up-scaling ART to include every HIV-positive individual,[11] they are bound to misclassifying some incident cases as established ones and vice versa. An alternative method is to estimate HIV incidence rate from a cross-sectional cohort study design, where selection of subjects is done presently and assessment covers both individuals' present and past experiences.[12] This design is less resource intensive as it allows one to calculate rates in a one-time survey and according to Hudson et al (2005), it requires fundamentally the same assumptions for its validity as other approaches, even though in practice certain of these assumptions require particular attention in a cross-sectional cohort study. However, there are common threats to the validity of a cross-sectional cohort design, these include selection bias and misclassification error which are a result of retrospective assessment of study subjects.[12] To estimate HIV incidence, we propose a weighted log likelihood approach based on a cross-sectional study design, that queries individuals' time since they last tested for HIV, validates the test results with formal documentation whenever possible, and tests all persons who do not have documentation of an HIV-positive status. The weights correct for differences in key HIV risk factors between persons with and without documentation of most recent HIV test. To gain efficiency, we incorporate into the weighted log-likelihood available information on error-prone self-reports from individuals who could not produce documentation. Our approach addresses two potential problems that arise in the cross-sectional cohort study design;

1. **Selection bias due to missing documentation:** Not all individuals are able to produce documentation, and simply assuming that persons with and without documentation are unconditionally exchangeable might not be valid, hence

inducing selection bias into estimates. We address this by incorporating inverse probability of selection weights conditional on available covariates into individuals' log likelihood function. In the absence of model misspecification for inverse probability weights (IPW), such estimators are consistent and appropriately correct for any bias associated with selection into samples with or without HIV test documentation.[13,14] Properties of such weighted estimators may be deduced by viewing them as solutions to a set of estimating equations and appealing to the well-established theory of M-estimation.[15,16] To gain more statistical efficiency, we incorporate into the weighted log-likelihood available information on error-prone self-reports from individuals who could not produce documentation.

2. **Misclassification error:** There is a notable mismatch between formal documentation and self-reported times since last HIV negative test among individuals who produced both, suggesting that the latter may be reported with error. In the presence of such misclassification error, leveraging information available in self-reported times since last HIV negative test without proper adjustment for misclassification error will induce bias.[17] We account for misclassification error by incorporating into our proposed weighted log likelihood function an explicit probabilistic model relating self-report records to documented dates of last HIV test estimated in the validated sub-sample where both are available. We refer to this new approach as a "pooled cross-sectional cohort study design", where the additional term refers to the fact that we are augmenting documented dates with self-reports.

To investigate finite sample properties of our weighted estimator, we conduct extensive Monte Carlo simulations, and compare our new estimator with other available methods. We then use the "Ya Tsie" data (also known as Botswana Combination Prevention Project, or BCPP) to estimate HIV incidence rates for 1.5 and 5 years prior to the survey simultaneously accounting for selection bias and misclassification error. The rest of the paper is organised as follows; In section 2.1, we outline the study design. We introduce notations in section 2.2, then list and discuss relevant assumptions in section 2.3. We describe the proposed weighted pooled log likelihood estimator in section 2.4 and evaluate its performance through extensive simulation studies in section 3. Results of application to BCPP data, corresponding conclusions and possible limitations are discussed in sections 3.1 and 4 respectively.

2 | METHODS

2.1 | Study Design

The methods developed in this paper are largely motivated by the BCPP study, which is a pair-matched cluster-randomized trial, funded by the United States of America President's Emergency Plan for AIDS Relief, designed to test whether a package of combination prevention interventions reduces population-level cumulative 30-month HIV incidence. The trial is being conducted in 30 communities in Botswana (15 matched-pairs) with a total population of about 180,000 people, representing nearly 10 % of Botswana's estimated

population. Fifteen communities were randomized to a combination prevention arm and 15 to a non-intervention arm. Interventions in the combination prevention group include home-based and mobile HIV testing, and counselling; point-of-care CD4 testing; linkage to care support; expanded ART; and enhanced male circumcision services. Detailed BCPP study procedures were previously published.[18] As part of this study, a random sample of 12,610 adults in 30 communities throughout Botswana, representing approximately 20 % of their respective households was enrolled. HIV status was obtained for 99.7% trial participants at enrolment (either through a documented positive HIV status or in-home rapid testing). Additionally, self-reported information on prior HIV testing and when available, corresponding documentation of self-reported result was also obtained at enrolment. In our analysis, these two sources of information; (1) self-reported and (2) documented dates of most recent HIV-negative tests were combined to retrospectively construct a cohort of HIV-negative persons, all of whom underwent HIV testing at enrolment. Participants who reported dates of their last HIV negative test in the last 1.5 years (6570 days) prior to BCPP enrolment were included in the primary analysis. In secondary analysis, we expanded the study population to include all subjects reporting dates of last HIV negative test within the prior 5 years (21900 days). We defined incident cases of HIV positivity if a person with a previous HIV-negative test result subsequently tested HIV-positive on the date of BCPP enrolment. Person-time at-risk of infection was calculated from date of the most recent HIV negative test to the date of testing during the BCPP enrolment. Through this, we identified 6,542 and 6,942 individuals for primary and secondary analyses respectively.

2.2 | Notations

Let T_i denote person i's time since last HIV negative test until HIV sero-conversion. Also, let F_i denote documented time from the last HIV negative test to BCPP enrolment and HIV test, hereafter referred to as retrospective follow-up time. We define a constant $K = 6570, 21900$ such that the at-risk group is $I(F_i \leq K)$ and zero otherwise. Let $\Delta_i = I(T_i \leq F_i)$, i.e. if person i in the at-risk cohort is found to be HIV positive at enrolment date and $\Delta_i = 0$ otherwise. Let Y_i^* be self-reported time since last HIV negative test at enrolment and HIV test, available for all subjects who self-report as negative whether or not documentation is available. This information is obtained prior to a request for formal documentation of the negative HIV test and therefore purely reflects subjects' recollection. Note also that in an attempt to improve accuracy, time of self-report of last HIV negative test was reported in terms of time windows spanning at most 1 month ($Y_i^* = 0$), between 1 to 5 months ($Y_i^* = 1$), between 6 to 12 months ($Y_i^* = 2$) and more than 12 months ($Y_i^* = 3$) prior to BCPP enrolment. Let Y_i be corresponding discretized version of F_i with $J = 0, 1, 2, 3$ categories, spanning the same time windows as Y_i^*. Because Y_i^* may not be equal to Y_i, the former may be viewed as a misclassified version of the latter. Let $R_i = 1$ if individual i produced both self-report and documentation of last HIV-negative test during the retrospective follow-up period of interest, and $R_i = 0$ denotes an individual with self-reported HIV negative test during the at-risk follow-up period without formal documentation. We refer to these persons as validated (v) and non-validated (nv) sub-samples throughout. Throughout our analysis, we assume that persons who self-report to be HIV positive are in fact positive. We also

assume that as in BCPP study, a large set of covariates $\underline{X}_i$ is measured on all participants at enrolment, which are key for explaining selection mechanism into validated sample. Throughout, we assume that we observe samples of n independent and identically distributed realizations, $\underline{Z}_i = (R_i, R_i F_i, \Delta_i, R_i Y_i, Y_i^*, \underline{X}_i)$. Welet P_T denote population density of T_i, P_F be population density of F_i, $P_{Y_i^* \mid F_i, \Delta_i, R_i, \underline{X}_i}$ be probability mass function of Y_i^* conditional on F_i, Δ_i, R_i and $\underline{X}_i$, $\pi_i = P_{R_i \mid F_i, \Delta_i, Y_i^*, \underline{X}_i}$ is the population density of R_i given $\overline{F}_i, \Delta_i, Y_i^*$, and available covariates $\underline{X}_i$.

2.3 | Assumptions

Throughout our analysis, we make the following assumptions;

1. Non-di erential misclassification, i.e., $P_{Y_i^* \mid F_l, \Delta_i, R_i, \underline{X}_i} = P_{Y_i^* \mid F_i}$.
2. Coarsened misclassification, i.e., $P_{Y_i^* \mid F_i} = P_{Y_i^* \mid Y_i}$.
3. Constant hazard rate of infection, i.e., $T_i \sim$ exponential (λ) where λ is incidence parameter of primary interest.
4. Constant hazard of testing times, i.e., $F_i \sim$ exponential(θ).
5. Missing at random (MAR), i.e., $R_i \perp (F_i, T_i) \mid \underline{X}_i$.

Assumption 1 implies that the probability mass function of self-reported information given documented date of last HIV negative test, HIV status, selection into validated or non-validated sub-samples and a set of covariates measured at BCPP enrolment only depends on documented date of last HIV negative test. Furthermore, from Assumption 2, this function depends on true retrospective follow-up time only through the time interval it belongs to. We encode this model as polytomous logistic regression for misclassification error with unknown parameters $\underline{\beta}_{j^* j}, (j^*, j) = 0, 1, 2, 3$, given by;

$$c_{j_i^* j_i} = Pr(Y_i^* = j^* \mid Y = j) = \frac{\exp(\beta_{j^* j})}{1 + \sum_{j'} \exp(\beta_{j^* j'})} \tag{1}$$

Assumption 3 is reasonable for short enough retrospective follow-up time such as 1.5 years and could be relaxed by assuming a piece-wise constant hazard if necessary. It is also important to note that Assumption 4 can be replaced by an alternative choice of parametric model. Under Assumption 5, we specify a parametric model for the selection process of the form;

$$\text{logit}\ [Pr(R_i = 1 \mid \underline{X}_i)] = \gamma^T \underline{X}_i \tag{2}$$

We propose to estimate γ with the standard maximum likelihood estimator $\hat{\gamma}$ which maximizes the corresponding logistic log-likelihood function based only on data $R_i, \underline{X}_i, i = 1, \ldots n$.

2.4 | Weighted log-likelihood function among the at-risk group

Let $\underline{\upsilon} = (\lambda, \theta, \underline{\beta}_{j^*j})$ and $\log L_i(\underline{z}_i; \underline{\upsilon})$ denote the log likelihood function for the at-risk group, i.e., $I(F_i \leq K)$ under assumptions 1 to 4 and the stronger assumption than Assumption 5, that selection into the validated sample is completely at random, i.e., $\pi_i = Pr(R_i = 1 \mid \underline{X}_i) = \pi$, a constant in (0,1). The corresponding expression $L_i(\underline{z}_i; \underline{\upsilon})$ is derived in Appendix 1. A unit's contribution to the corresponding score function for $\underline{\upsilon}$ has two components, $S_i^{\upsilon}(\underline{z}_i^{\upsilon}; \lambda), \underline{z}_i^{\upsilon} = (f_i, \Delta_i)$, and $S_i^{n\upsilon}(\underline{z}_i^{n\upsilon}; \underline{\upsilon}), \underline{z}_i^{n\upsilon} = (f_i, y_i^*, \Delta_i)$, for the validated and non-validated sub-samples respectively, given by;

$$S_i^{\upsilon}(\underline{z}_i^{\upsilon}; \lambda) = f_i\left(\frac{e^{-\lambda f_i} + \Delta_i - 1}{1 - e^{-\lambda f_i}}\right) \tag{3}$$

and

$$S_i^{n\upsilon}(\underline{z}_i^{n\upsilon}; \underline{\upsilon}) = \left[\frac{\sum_{j=0}^{J} c_{j_i^* j_i} \int_{s_j}^{s_{j+1}} f_i e^{-(\lambda+\theta) f_i} df_i}{\sum_{j=0}^{J} c_{j_i^* j_i} \int_{s_j}^{s_{j+1}} e^{-\theta f_i}\left(1 - e^{-\lambda f_i}\right) df_i}\right]^{\Delta_i} \left[-\frac{\sum_{j=0}^{J} c_{j_i^* j_i} \int_{s_j}^{s_{j+1}} f_i e^{-(\lambda+\theta) f_i} df_i}{\sum_{j=0}^{J} c_{j_i^* j_i} \int_{s_j}^{s_{j+1}} e^{-(\lambda+\theta) f_i} df_i}\right]^{1-\Delta_i} \tag{4}$$

$c_{j^* ji}$ is given in (1).

Under assumptions 1 to 4 and the weaker Assumption 5, i.e., selection into the validated sub-sample depends only on observed data $\underline{X}_i$, we propose to formally account for selection bias by incorporating inverse probability weights for selection among the at-risk group to obtain the following estimating equation for $\hat{\underline{\upsilon}}$, the estimator of $\underline{\upsilon}$.

$$\sum_{i=1}^{n} \widehat{\psi}(\underline{z}_i; \hat{\underline{\upsilon}}) = 0, \tag{5}$$

where,

$$\psi(\underline{z}_i; \underline{\upsilon}) = \frac{R_i}{\pi_i} S_i^{\upsilon}(\underline{z}_i^{\upsilon}; \lambda) + \frac{1 - R_i}{1 - \pi_i} S_i^{n\upsilon}(\underline{z}_i^{n\upsilon}; \underline{\upsilon}), \tag{6}$$

$\widehat{\psi}$ is equal to ψ evaluated at $\hat{\pi}_i$ the mle of $\pi_i = Pr(R_i = 1 \mid \underline{X}_i)$ under model (2).

Note that the un-weighted estimating equation corresponds to (6) under $\pi_i = \frac{1}{2}$. Under that scenario, probability of selection into validated or non-validated sub-samples is constant and equal for everyone. The variance-covariance matrix of $\hat{\underline{\upsilon}}$ is then estimated by the standard inverse of the observed information matrix. Furthermore, inference based on the Wald, score or likelihood ratio statistics may be obtained under standard maximum likelihood theory. Because equation (6) is weighted, the estimator of the asymptotic variance-covariance matrix of $\hat{\underline{\upsilon}}$ can be obtained from the non-parametric bootstrap or the sandwich estimator

given in Appendix 3. We have also established unbiasedness of the estimating equation (6) in Appendix 2.

3 | SIMULATION STUDIES

In order to evaluate the performance of our proposed estimator, we conducted Monte Carlo simulations under conditions motivated by BCPP dataset. For all individuals ($n = 7000$), we first generated retrospective follow-up time F_i and time to sero-conversion T_i from exponential distributions with parameters 0.3 and 0.2 respectively. As defined earlier, HIV statusf at cross-sectional survey was then given as $\Delta_i = I(T_i \leq F_i)$. To have evident selection bias such as in BCPP, we simulated covariates dependent on HIV status and time since last HIV negative test until HIV sero-conversion as follows; D_{1i} and D_{2i} from Normal (0,4.41) and (0,1.44) respectively, then $X_{1i} = \exp(0.1\Delta_i - 0.3T_i)D_{1i} + 3\Delta_i + 0.2T_i - 0.3F_i$, $X_{2i} = \exp(0.1\Delta_i - 0.3T_i)D_{2i} - \Delta_i - 0.2T_i$. To $- 0.3F_i$, $X_{3i} = \exp(0.1\Delta_i - 0.3T_i)D_{2i} + 3\Delta_i + 0.2T_i - 0.3F_i$ construct validated ($R_i = 1$) and non-validated ($R_i = 0$) sub-samples comparable to BCPP, we simulated a binary variable R_i from Bernoulli $(\text{expit}(0.5X_{1i} + 0.2X_{2i} - X_{3i}))$. To match BCPP, we binned F_i into 4 categories to define Y_i, taking values 0 if $F_i \in [0, 0.5]$, 1 if $F_i \in (0.5, 2]$, 2 if $F_i \in (2, 3]$ and 3 if $F_i \in (3, 7]$. We excluded all individuals with $F_i > 7$ in order to have a reasonably short period of retrospective follow-up motivated by BCPP. To simulate self-reported times since individuals' last HIV negative test, we constructed Y_i^* according to a multinomial distribution with conditional probabilities shown in Table 1.

We performed two types of simulations being un-weighted and weighted analysis. For each of them, we compared the validated sample-only versus pooled sample estimators. For un-weighted analyses, validated sample-only estimator exclusively uses documented, error-free individuals' times since last HIV negative test while the pooled estimator incorporates error-prone self-reports accounting for misclassification but not for selection. The weighted analyses involved adjusting individual's log-likelihood functions with inverse probability weights of selection into validated or non-validated sub-samples given X_{1i}, X_{2i} and X_{3i} in the two estimators to account for selection bias. Estimated weights were computed based on the MLE of a correctly specified logistic regression model for π_i, and solving equation (5). We compared our proposed estimator that simultaneously accounts for selection bias and misclassification error with the other three estimators that ignore at least one of these problems. We conducted 1000 simulations and report Monte Carlo bias, Monte Carlo percent bias, Monte Carlo mean square error and relative efficiencies $(\text{RE}) = \frac{MSE_i}{MSE_{wp}}$, where 'wp' refers to our proposed weighted pooled sample estimator. All computations were performed in R version 3.5.1.[19]

Table 2 shows that as expected, failure to formally account for selection bias when formal documentation is missing at random yields biased estimates for HIV incidence rate. This is reflected in the large absolute percent bias of about 26 % for the un-weighted validated sample-only estimator. Under the same conditions, trying to leverage self-reported dates of

individuals' last HIV negative test increases bias (absolute percent bias = 30 %) due to presence of misclassification error. It is evident from the same table that adjusting individual's log likelihood functions by incorporating inverse probability weights of selection in the validated sub-sample significantly reduces bias in both validated sample-only (absolute percent bias = 0.1 %) and pooled sample estimators (absolute percent bias = 1.2 %). Even though both weighted sample estimators are nearly unbiased, the pooled sample estimator is more Efficient as shown by a relative efficiency of 1.5 for the weighted validated sample-only estimator. These compelling large sample simulation results suggest that incorporating error-prone self-reported information into the weighted log-likelihood function and appropriately accounting for misclassification error on the outcome variable as we did can lead to substantial efficiency gains.

3.1 | Application to BCPP enrolment data

We used BCPP enrolment data to obtain both un-weighted and weighted estimates of HIV incidence rates for individuals reporting negative status in the last 1.5 and 5 years prior to this survey using the validated sample-only and pooled sample estimators. Table 4 provides basic demographic descriptions of enrolled subjects, stratified by availability of documented HIV negative result and time since last documented test as reported in Abuelezam et al.[20] Evidence of misclassification error is shown in Table 5. For weighted analyses, we regressed an indicator for presence of documentation on the following covariates (measured at BCPP baseline household survey); age, gender, current relationship status, religious a liation, education, employment status, income, time spent away from the community, livestock owned by the household, number of children in the household, age at first sexual intercourse, number of sexual partners during the past 12 months, number of lifetime sexual partners, inconsistent condom use, transactional sex, frequency of alcohol use, alcohol use by self/partner during sex, and self reported time since most recent HIV test and self-reported result of most recent HIV test. We incorporated in the selection model 74 two-way interaction terms by taking the cross-product of each socio-demographic covariate with each behavioural covariate. To build the multivariate logistic regression models required by inverse probability weighting, we used a stepwise selection procedure to identify covariates from the list of candidate predictors described above. The entry criteria were set to a $P < 0.2$. We also included missing indicators for each selected variable with missing values in the final logistic regression model for the weights to maximize the number of cases included in the final models and to maintain a constant sample size across analyses.

Table 3 shows estimates of HIV incidence rate per 100 person years at-risk of infection from both un-weighted and weighted analyses for validated sample-only versus pooled sample estimators. For un-weighted analysis, reported standard errors (SE's) are estimated from the inverse of information matrix while we report non-parametric bootstrap SE's for all weighted estimators. We also report 95% confidence intervals for all four scenarios corresponding to primary and secondary analyses. Among individuals who tested HIV negative 1.5 years prior to BCPP enrolment, the estimated weighted incidence rate from validated only sub-sample was 1.10 per 100 person years at-risk of infection, with estimated standard error of 0.33 corresponding to a 95% confidence interval of (0.54,1.82). The proposed, pooled weighted estimator for the same period yielded an estimate of 1.27 per 100

person years at risk of infection, corresponding non-parametric bootstrap standard error of 0.08, 95% confidence interval of (1.12,1.42). For 5 year period of retrospective follow-up, incidence rate for validated sample-only was estimated to be 1.02 per 100 person years at-risk of infection, with corresponding standard error of 0.24 yielding a 95% confidence interval (0.57,1.50). Our proposed, pooled estimator for this period yielded 1.17 as the annual incidence rate per 100 persons exposed to risk, 0.08 as the corresponding non-parametric bootstrap standard error and 95% confidence interval (1.03,1.32). For weighted analyses, we observe that although incorporating self-reports into the analyses and additionally accounting for misclassification error yielded slightly larger estimates of HIV incidence rate for 1.5 and 5 years prior to BCPP enrolment, non-parametric bootstrap standard errors were 76% and 67% smaller than values of validated sample only estimator that exclusively accounted for selection bias over respective periods. In un-weighted analyses, estimates for incidence rate and standard errors from validated sub-sample only (1.35, SE= 0.37) and (1.14, SE=0.24) were not far from their weighted counterparts for the two respective periods, (1.10, SE=0.33) & (1.01, SE=0.24). However, the un-weighted pooled estimator yielded remarkably larger estimates (8.88, SE= 0.55) & (4.97, SE= 0.30) versus their weighted counterparts (1.27, SE=0.08) & (1.17, SE=0.08) over the two periods, suggesting significant selection bias in un-weighted analysis even after accounting for misclassification error.

4 | CONCLUSIONS

We have proposed a resource-Efficient cross-sectional cohort study design, that relies on querying individuals' history of HIV testing and where possible, validating it with formal documentation to estimate HIV incidence rate, and testing all persons not known to be HIV-positive at the cross-sectional visit. We proposed and validated through extensive Monte Carlo simulations a corresponding weighted log likelihood estimator for incidence rate under this study design and model assumptions. This estimator combines individuals' self-reported and documented times since their last HIV-negative test and simultaneously accounts for possible selection bias and misclassification error assuming no model misspecification. Our estimator is therefore robust to both potential sources of bias in cross-sectional cohort studies as shown in simulation studies and an application to BCPP enrolment data. Our best estimate of HIV incidence rate is largely consistent with figures from Botswana Aids Impact Survey, which estimated a crude incidence rate of 1.35 per 100 person years at-risk of infection, using the Recent Infection Algorithm (RITA)[21] and BCPP laboratory based which rely on recency assays (Incidence rate= 1.06, 95 % CI= (0.70, 1.42)), but notably more statistically efficient and less resource intensive. Our confidence intervals were 77 % and 71% narrower than estimates of Abuelezam et al,[20] who obtained inverse probability weighted incidence rates of 0.98 (0.32, 1.65) and 1.01 (0.52, 1.51) per 100 person years at-risk of infection for 1.5 and 5 years retrospective follow-up periods respectively by using validated only samples. Although our estimator is evidently more efficient, it has potential threats and limitations. These include non-differential exiting from study population over time, i.e., the rate of participant exiting depends on HIV status. Common causes may be high mortality and out-migration rates among HIV positive individuals. We acknowledge that this scenario may seriously affect our estimator. However,

due to Botswana being close to 90–90-90 targets of high ART coverage and viral suppression,[18] we hope that mortality and migration are not major problems in the BCPP baseline data set. Moreover, we expect more countries to commit to the 90–90-90 UNAIDS targets in future, this will reduce such HIV related mortality rates, hence reducing the effect of this type of differential exiting from the study population.

Another potential limitation is that HIV status and testing may not be independent. That is, individuals may go for testing because of the presence of a sero-conversion related illness or they have engaged in some form of risk-inducing behaviour, this will result in sampled individuals differing systematically with those excluded. As a sensitivity analysis, we regressed an indicator variable (C_i=1 for individuals who reported to have tested within k years prior to BCPP enrolment, 0 otherwise) on available covariates to account for factors associated with HIV testing assuming correct model specification. We then multiplied resulting inverse probability of $C_i = 1 \mid \underline{X}_i$ with our prior weights based on R_i. This analysis additionally accounts for bias related with dependence between HIV testing and different variables such as risk inducing behaviours. For the two retrospective follow-up times respectively, validated only analysis appears to be robust to these additional potential sources of selection bias, i.e., 1.13 (SE=0.33) and 1.00 (SE=0.22), while pooled analyses results were more sensitive, increasing point estimates by 37 % and 42 %, i.e., 1.74 (SE=0.12) and 1.66 (SE=0.11). Results are reported in Table 6.

Our estimator is also sensitive to misspecification of the selection model used to construct inverse probability weights. That is, if this model is miss-specified, incidence rate under our proposed method will generally be biased.[1422] In the future, we hope to develop a doubly robust inverse probability weighted estimator to additionally account for possible partial model misspecification. We did not formally account for possible clustering effect, however, this effect was negligible and will unlikely affect our results. A notable possibility is that self-reported time since last HIV test could be dependent on person's underlying HIV status (differential misclassification), i.e. Individuals are saying that they tested negative six months ago, but actually know that they are positive. We intend to account for this problem by using instrumental variables in future.

ACKNOWLEDGMENTS

Financial disclosure

This work was supported through the Sub-Saharan African Network for TB/HIV Research Excellence (SANTHE), a DELTAS Africa Initiative [grant # DEL-15-006]. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [grant #107752/Z/15/Z] and the UK government. The views expressed in this publication are those of the author(s) and not necessarily those of AAS, NEPAD Agency, Wellcome Trust or the UK government.

The BCPP study was supported by the US President's Emergency Plan for AIDS Relief (PEPFAR) through the Centers for Disease Control and Prevention (CDC) under the terms of cooperative agreement U01 GH000447 and GH0001911.

Eric Tchetgen Tchegen and Kathleen E. Wirth received support from the National Institute of Health (NIH) R01AI27271.

Eric Tchetgen Tchegen also received a grant from National Cancer Institute (NCI) R01CA222147.

4.1 | Appendix 1;: Derivation of the estimating equation for v^¯among the at-risk group

From sections 2.2 to 2.4, we have n independent and identically distributed realizations $\underline{Z_i} = (R_i, R_iF_i, \Delta_i, R_iY_i, Y_i^*, \underline{X_i})$ of individuals in the at-risk group $I(F_i \leq K)$. Using adopted notations, under the Assumptions 1 to 4 and the stronger Assumption 5, that selection into the validated sample is completely at random, i.e., individuals' data likelihood is expressed as;

$$L_i = \left\{ f^{\upsilon}(Y_i^* = j^* \mid Y_i = j) f^{\upsilon}(\Delta_i \mid F_i = f_i) f^{\upsilon}(F_i = f_i) \right\}^{r_i} \tag{7}$$

$$\times \left\{ \int_{\forall f_i} f^{\upsilon}(Y_i^* = j^* \mid Y_i = j) f^{\upsilon}(\Delta_i \mid F_i = f_i) f^{\upsilon}(F_i = f_i) df_i \right\}^{1 - r_i} \tag{8}$$

where;

$$f^{\upsilon}(Y_i^* = j^* \mid Y_i = j) = Pr(Y_i^* = j^* \mid Y = j) = \frac{\exp(\beta_{j^* j})}{1 + \sum_{j'} \exp(\beta_{j^* j'})} = c_{j_i^* j_i}, \tag{9}$$

$j = 0, 1, \ldots, J$, $j^* = 0, 1, \ldots, J^*$ and $\beta_{j*0} = 0, \forall j$.

Among this group, an individual is HIV positive $(\Delta_i = 1)$ if they converted before or at BCPP enrolment, i.e. $I(T_i \leq F_i)$. Therefore, $P^{\upsilon}(\Delta_i = 1 \mid F_i) = P^{\upsilon}(T_i \leq F_i \mid F_i)$, which is the cumulative density function of exponential distribution (λ) according to Assumption 3. Hence,

$$f^{\upsilon}(\Delta_i = 1 \mid F_i = f_i) = \int_0^{f_i} \lambda e^{-\lambda t_i} dt_i \tag{10}$$

$$= 1 - e^{-\lambda f_I}, \lambda > 0; f_i \in [0, \infty) \tag{11}$$

So

$$f^{\upsilon}(\Delta_i = 0 \mid F_i = f_i) = 1 - f^{\upsilon}(\Delta_i = 1 \mid F_i = f_i) \tag{12}$$

$$= e^{-\lambda f_i}, \lambda > 0; f_i \in [0, \infty) \tag{13}$$

Therefore

$$f^{\upsilon}(\Delta_i \mid F_i = f_i) = \left(1 - e^{-\lambda f_i}\right)^{\Delta_i}\left(e^{-\lambda f_i}\right)^{1-\Delta_i}, \lambda > 0; \Delta_i = 0, 1; f_i \in [0, \infty) \tag{14}$$

From Assumption 4 we have,

$$f^{\upsilon}(F_i = f_i) = \theta e^{-\theta f_i}, \theta > 0; f_i \in [0, \infty) \tag{15}$$

and by definition;

$$\begin{aligned} &\int_{\forall f_i} c_{j_i^* j_i} f^{\upsilon}(\Delta_i \mid F_i = f_i) f^{\upsilon}(F_i = f_i) df_i \\ &= \sum_{j=0}^{J} c_{j_i^* j_i} \int_{s_j}^{s_{j+1}} f^{\upsilon}(\Delta_i \mid F_i = f_i) f^{\upsilon}(F_i = f_i) df_i, \end{aligned} \tag{16}$$

Therefore for this group, an individual's data likelihood is formally expressed as;

$$\begin{aligned} L_i(\underline{z}_i; \underline{\upsilon}) = &\left\{c_{j_i^* j_i}\left(1 - e^{-\lambda f_i}\right)^{\Delta_i}\left(e^{-\lambda f_i}\right)^{1-\Delta_i}\theta e^{-\theta f_i}\right\}^{r_i} \\ &\times \left\{\sum_{j=0}^{J} c_{j_i^* j_i} \int_{s_j}^{s_{j+1}} \left(1 - e^{-\lambda f_i}\right)^{\Delta_i}\left(e^{-\lambda f_i}\right)^{1-\Delta_i}\theta e^{-\theta f_i} df_i\right\}^{1-r_i}, \end{aligned} \tag{17}$$

which is proportional to;

$$\left\{\left(1 - e^{-\lambda f_i}\right)^{\Delta_i}\left(e^{-\lambda f_i}\right)^{1-\Delta_i}\right\}^{r_i}\left\{\theta \sum_{j=0}^{J} c_{j_i^* j_i} \int_{s_j}^{s_{j+1}} \left(e^{-\theta f_i} - e^{-(\lambda+\theta) f_i}\right)^{\Delta_i}\left(e^{-(\lambda+\theta) f_i}\right)^{1-\Delta_i} df_i\right\}^{1-r_i}, \tag{18}$$

where $\underline{\upsilon} = \left(\lambda, \theta, \underline{\beta}_{j^* j}\right)$. Therefore the un-weighted log-likelihood function is expressed as;

$$\log L_i(\underline{z}_i; \underline{\upsilon}) = \sum_{i=1}^{n} \left\{r_i\left[\Delta_i \log\left(1 - e^{-\lambda f_i}\right) - \lambda(1 - \Delta_i) f_i\right]\right. \tag{19}$$

$$\left. + (1 - r_i)\left[\log\theta + \log\left(\sum_{j=0}^{J} c_{j_i^* j_i} \int_{s_j}^{s_{j+1}} \left(e^{-\theta f_i} - e^{-(\lambda+\theta) f_i}\right)^{\Delta_i}\left(e^{-(\lambda+\theta) f_i}\right)^{1-\Delta_i} df_i\right)\right]\right\} \tag{20}$$

For $\Delta_i = 1$, the integral in equation (20) becomes;

$$\int_{s_j}^{s_{j+1}} \left(e^{-\theta f_i} - e^{-(\lambda+\theta) f_i}\right) df_i = \left[\frac{e^{-(\lambda+\theta)s_{j+1}} - e^{-(\lambda+\theta)s_j}}{\lambda+\theta} - \frac{e^{-\theta s_{j+1}} - e^{-\theta s_j}}{\theta}\right] \tag{21}$$

For $\Delta_i = 0$, we have;

$$\int_{s_j}^{s_{j+1}} e^{-(\lambda+\theta) f_i} df_i = \left[\frac{e^{-(\lambda+\theta)s_j} - e^{-(\lambda+\theta)s_{j+1}}}{\lambda+\theta}\right] \tag{22}$$

Corresponding score equation for $\underline{\upsilon}$ has two components, $S_i^{\upsilon}(\underline{z}_i^{\upsilon}; \lambda), \underline{z}_i^{\upsilon} = (f_i, \Delta_i)$, and $S_i^{n\upsilon}(\underline{z}_i^{n\upsilon}; \underline{\upsilon}), \underline{z}_j^{n\upsilon} = (f_i, y_i^*, \Delta_i)$, for the validated and non-validated sub-samples respectively, given by;

$$S_i^{\upsilon}(\underline{z}_i^{\upsilon}; \lambda) = f_i\left(\frac{e^{-\lambda f_i} + \Delta_i - 1}{1 - e^{-\lambda f_i}}\right) \tag{23}$$

and

$$S_i^{n\upsilon}(\underline{z}_i^{n\upsilon}; \underline{\upsilon}) = \frac{\sum_{j=0}^{J} c_{j_i^* j_i} \int_{s_j}^{s_{j+1}} f_i e^{-\theta f_i} \left[\Delta_i \left(e^{-\lambda f_i}\right)^{2-\Delta_i} \left(1 - e^{-\lambda f_i}\right)^{\Delta_i - 1} - (1-\Delta_i)\left(e^{-\lambda f_i}\right)^{1-\Delta_i} \left(1 - e^{-\lambda f_i}\right)^{\Delta_i}\right] df_i}{\sum_{j=0}^{J} c_{j_i^* j_i} \int_{s_j}^{s_{j+1}} e^{-\theta f_i} \left(1 - e^{-\lambda f_i}\right)^{\Delta_i} \left(e^{-\lambda f_i}\right)^{1-\Delta_i} df_i}, \tag{24}$$

which simplifies to

$$S_i^{n\upsilon}(\underline{z}_i^{n\upsilon}; \underline{\upsilon}) = \left[\frac{\sum_{j=0}^{J} c_{j_i^* j_i} \int_{s_j}^{s_{j+1}} f_i e^{-(\lambda+\theta) f_i} df_i}{\sum_{j=0}^{J} c_{j_i^* j_i} \int_{s_j}^{s_{j+1}} e^{-\theta f_i} \left(1 - e^{-\lambda f_i}\right) df_i}\right]^{\Delta_i} \left[-\frac{\sum_{j=0}^{J} c_{j_i^* j_i} \int_{s_j}^{s_{j+1}} f_i e^{-(\lambda+\theta) f_i} df_i}{\sum_{j=0}^{J} c_{j_i^* j_i} \int_{s_j}^{s_{j+1}} e^{-(\lambda+\theta) f_i} df_i}\right]^{1-\Delta_i} \tag{25}$$

$c_{j_i^* j_i}$ takes the form from equation (9).

Under Assumptions 1 to 4 and weaker Assumption 5, i.e., selection into the validated sub-sample depends only on observed data $\underline{X}_i$, we propose to formally account for selection bias in the at-risk group by incorporating inverse probability weights for selection through a parametric model of the form;

$$\text{logit}[Pr(R_i = 1 \mid \underline{X}_i)] = \gamma^T \underline{X}_i \tag{26}$$

We propose to estimate y with the maximum likelihood estimator f which maximizes the corresponding logistic log-likelihood function based only on data $R_i, \underline{X}_i, i = 1, \ldots, n$. The resulting estimating equation for $\hat{\upsilon}$, the estimator of $\underline{\upsilon}$, is given by

$$\sum_{i=1}^{n} \widehat{\psi}(\underline{z}_i; \widehat{\underline{\upsilon}}) = 0, \tag{27}$$

$$\psi(\underline{z}_i; \underline{\upsilon}) = \frac{R_i}{\pi_i} S_i^{\upsilon}(\underline{z}_i^{\upsilon}, \lambda) + \frac{1 - R_i}{1 - \pi_i} S_i^{n\upsilon}(\underline{z}_i^{n\upsilon}, \underline{\upsilon}), \tag{28}$$

$\widehat{\psi}$ is equal to ψ evaluated at $\hat{\pi}_i$ the mle of $\pi_i = Pr(R_i = 1 \mid \underline{X}_i)$ under model (26).

4.2 | Appendix 2;: Proof that the estimating equation for v^ is unbiased

We show that the estimating equation has mean zero (i.e., unbiased) at the true value of $\underline{\upsilon}$.

$$E[\psi(\underline{z}_i; \underline{\upsilon})] = E\left\{ \frac{R_i}{\pi_i} S_i^{\upsilon}(\underline{z}_i^{\upsilon}; \lambda) + \frac{1 - R_i}{1 - \pi_i} S_i^{n\upsilon}(\underline{z}_i^{n\upsilon}; \underline{\upsilon}) \right\} \tag{29}$$

$$= E\left\{ E\left[\frac{R_i}{\pi_i} S_i^{\upsilon}(\underline{z}_i^{\upsilon}; \lambda) \mid \underline{X}_i, F_i, \Delta_i \right] + E\left[\frac{1 - R_i}{1 - \pi_i} S_i^{n\upsilon}(\underline{z}_i^{n\upsilon}; \underline{\upsilon}) \mid \underline{X}_i, F_i, \Delta_i \right] \right\} \tag{30}$$

$$= E\left\{ \frac{S_i^{\upsilon}(\underline{z}_i^{\upsilon}; \lambda)}{\pi_i} E(R_i = 1 \mid \underline{X}_i, F_i, \Delta_i) + \frac{S_i^{n\upsilon}(\underline{z}_i^{n\upsilon}; \underline{\upsilon})}{1 - \pi_i} [1 - E(R_i = 1 \mid \underline{X}_i, F_i, \Delta_i)] \right\} \tag{31}$$

$$= E\left\{ \frac{S_i^{\upsilon}(\underline{z}_i^{\upsilon}; \lambda)}{\pi_i} \pi_i + \frac{S_i^{n\upsilon}(\underline{z}_i^{n\upsilon}; \underline{\upsilon})}{1 - \pi_i} (1 - \pi_i) \right\} \tag{32}$$

$$= E\left[S_i^{\upsilon}(\underline{z}_i^{\upsilon}; \lambda) \right] + E\left[S_i^{n\upsilon}(\underline{z}_i^{n\upsilon}; \underline{\upsilon}) \right] \tag{33}$$

Now,

$$E\left[S_i^{\upsilon}(\underline{z}_i^{\upsilon}; \lambda) \mid F_i = f_i \right] = \sum_{\Delta_i = 0}^{1} S_i^{\upsilon}(\underline{z}_i^{\upsilon}; \lambda) f^{\upsilon}(\Delta_i \mid F_i = f_i) \tag{34}$$

$$= \sum_{\Delta_i = 0}^{1} f_i \left(\frac{e^{-\lambda f_i} + \Delta_i - 1}{1 - e^{-\lambda} f_i} \right) \left(1 - e^{-\lambda f_i} \right)^{\Delta_i} \left(e^{-\lambda f_i} \right)^{1 - \Delta_i} \tag{35}$$

$$= f_i\left(\frac{e^{-\lambda f_i}-1}{1-e^{-\lambda f_i}}\right)\left(e^{-\lambda f_i}\right) + f_i\left(\frac{e^{-\lambda f_i}}{1-e^{-\lambda f_i}}\right)\left(1-e^{-\lambda f_i}\right) \tag{36}$$

$$= \frac{1}{1-e^{\lambda f_i}}\left\{-f_i e^{-\lambda f_i}\left(1-e^{-\lambda f_i}\right) + f_i e^{-\lambda f_i}\left(1-e^{-\lambda f_i}\right)\right\} = 0, \tag{37}$$

$$E\left[S_i^{nv}(\underline{z}_i^{nv};\underline{v}) \mid F_i = f_i\right] = \sum_{\Delta_i=0}^{1} S_i^{nv}(\underline{z}_i^{nv};\underline{v}) f^{nv}(\Delta_i \mid F_i = f_i) \tag{38}$$

$$= \sum_{\Delta_i=0}^{1}\sum_{j=0}^{J} c_{j_i^* j_i}\int_s^{s_j+1} e^{-\theta f_i}\left(1-e^{-\lambda f_i}\right)^{\Delta_i}\left(e^{-\lambda f_i}\right)^{1-\Delta_i} df_i \tag{39}$$

$$\times\left[\frac{\sum_{j=0}^{J} c_{j_i^* j_i}\int_{s_j}^{s_j+1} f_i e^{-(\lambda+\theta)f_i} df_i}{\sum_{j=0}^{J} c_{j_i^* j_i}\int_{s_j}^{s_j+1} e^{-\theta f_i}\left(1-e^{-\lambda f_i}\right) df_i}\right]^{\Delta_i}\left[-\frac{\sum_{j=0}^{J} c_{j_i^* j_i}\int_{s_j}^{s_j+1} f_i e^{-(\lambda+\theta)f_i} df_i}{\sum_{j=0}^{J} c_{j_i^* j_i}\int_{s_j}^{s_j+1} e^{-(\lambda+\theta)f_i} df_i}\right]^{1-\Delta_i} \tag{40}$$

$$= \left\{\sum_{j=0}^{J} c_{j_i^* j_i}\int_{s_j}^{s_j+1} e^{-(\lambda+\theta)f_i} df_i\left[-\frac{\sum_{j=0}^{J} c_{j_i^* j_i}\int_{s_j}^{s_j+1} f_i e^{-(\lambda+\theta)f_i} df_i}{\sum_{j=0}^{J} c_{j_i^* j_i}\int_{s_j}^{s_j+1} e^{-(\lambda+\theta)f_i} df_i}\right]\right. \tag{41}$$

$$\left. + \sum_{j=0}^{J} c_{j_i^* j_i}\int_{s_j}^{s_j+1}\left(e^{-\theta f_i} - e^{-(\lambda+\theta)f_i}\right) df_i \left[\frac{\sum_{j=0}^{J} c_{j_i^* j_i}\int_{s_j}^{s_j+1} f_i e^{-(\lambda+\theta)f_i} df_i}{\sum_{j=0}^{J} c_{j_i^* j_i}\int_{s_j}^{s_j+1}\left(e^{-\theta f_i} - e^{-(\lambda+\theta)f_i}\right) df_i}\right]\right\} \tag{42}$$

$$= -\sum_{j=0}^{J} c_{j_i^* j_i}\int_{s_j}^{s_j+1} f_i e^{-(\lambda+\theta)f_i} df_i + \sum_{j=0}^{J} c_{j_i^* j_i}\int_{s_j}^{s_j+1} f_i e^{-(\lambda+\theta)f_i} df_i = 0 \tag{43}$$

This result shows that the estimating equation is unbiased at the true value $\underline{v}$.

4.3 | Appendix 3;: Asymptotic properties of the weighted estimator v¯^

Suppose that $\underline{v}^* = (\underline{v}, \underline{\gamma})$, according to theory of M-estimators,[15]

$$\sqrt{n}(\hat{\underline{v}}^* - \underline{v}^*) \text{ is AMN}(0, \Sigma), n \to \infty \tag{44}$$

where Σ is a sandwich estimator given by;

$$\Sigma = A_n(\underline{\upsilon}^*)^{-1} B_n(\underline{\upsilon}^*) \left\{ A_n(\underline{\upsilon}^*)^{-1} \right\}^T \tag{45}$$

$$A_n(\underline{\upsilon}^*) = E\left[-\frac{\partial}{\partial \underline{\upsilon}^{*T}} \psi(\underline{z}_i; \underline{\upsilon}^*) \right] \tag{46}$$

and

$$B_n(\underline{\upsilon}^*) = E\left[\psi(\underline{z}_i; \underline{\upsilon}^*) \psi(\underline{z}_i; \underline{\upsilon}^*)^T \right] \tag{47}$$

AMN means “asymptotically multivariate normal.” The asymptotic variance of $\hat{\upsilon}^*$ can therefore be estimated consistently by;

$$\Sigma = \left[-\frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \hat{\upsilon}^{*T}} \psi(\underline{z}_i; \hat{\underline{\upsilon}}^*) \right]^{-1} \left[\frac{1}{n} \sum_{i=1}^{n} \psi(\underline{z}_i; \hat{\underline{\upsilon}}^*) \psi(\underline{z}_i; \hat{\upsilon}^*)^T \right] \left\{ \left[-\frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \hat{\underline{\upsilon}}^{*T}} \psi(\underline{z}_i; \hat{\underline{\upsilon}}^*) \right]^{-1} \right\}^T \tag{48}$$

Abbreviations:

0

HIV	Human Immunodeficiency Virus

References

1. Mark Woodward. Epidemiology: study design and data analysis. CRC press; 2013.
2. Lesego Gabaitiri, Mwambi Henry G, Lagakos Stephen W, Pagano Marcello. A likelihood estimation of HIV incidence incorporating information on past prevalence. South African Statistical Journal. 2013;47(1):15–31. [PubMed: 25197147]
3. Brian Williams, Eleanor Gouws, David Wilkinson, Karim Salim Abdool. Estimating HIV incidence rates from age prevalence data in epidemic situations. Statistics in medicine. 2001;20(13):2003–2016. [PubMed: 11427956]
4. Breslow Norman E, Day Nicholas E, Davis Walter. Statistical methods in cancer research. International agency for research on cancer Lyon; 1980.
5. Kaplan Edward H, Brookmeyer Ron. Snapshot estimators of recent HIV incidence rates. Operations Research. 1999;47(1):29–37.
6. Ron Brookmeyer, Quinn Thomas C Estimation of current human immunodeficiency virus incidence rates from a cross-sectional survey using early diagnostic tests. American journal of epidemiology. 1995;141(2):166–172. [PubMed: 7817972]
7. Janssen RS. New testing strategy to detect early HIV-1 infection for use in incidence estimates and for clinical and prevention purposes (vol 280, pg 42, 1998). JAMA-JOURNAL OF THE AMERICAN MEDICAL ASSOCIATION. 1999;281(20):1893–1893.
8. Raji Balasubramanian, Lagakos Stephen W. Estimating HIV incidence based on combined prevalence testing. Biometrics. 2010;66(1):1–10. [PubMed: 19397583]
9. Hallett Timothy B, Zaba Basia, Todd Jim, et al. Estimating incidence from prevalence in generalised HIV epidemics: methods and validation. PLoS medicine. 2008;5(4):e80.

10. Rehle Thomas M, Hallett Timothy B, Shisana Olive, et al. A decline in new HIV infections in South Africa: estimating HIV incidence from three national HIV surveys in 2002, 2005 and 2008. PloS one. 2010;5(6):e11094.
11. HIV/AIDS Joint United Nations Programme, HIV/Aids Joint United Nations Programme, others. 90–90-90: an ambitious treatment target to help end the AIDS epidemic. Geneva: Unaids. 2014;.
12. Hudson James I, Pope Harrison G Jr, Glynn Robert J. The cross-sectional cohort study: an underutilized design. Epidemiology. 2005;16(3):355–359. [PubMed: 15824552]
13. Robins James M. Robust estimation in sequentially ignorable missing data and causal inference models. In: :6–10Indianapolis, IN; 2000.
14. Anastasios Tsiatis. Semiparametric theory and missing data. Springer Science & Business Media; 2007.
15. Stefanski Leonard A, Boos Dennis D. The calculus of M-estimation. The American Statistician. 2002;56(1):29–38.
16. Lunceford Jared K, Davidian Marie. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in medicine. 2004;23(19):2937–2960. [PubMed: 15351954]
17. Carroll Raymond J, Ruppert David, Crainiceanu Ciprian M, Stefanski Leonard A. Measurement error in nonlinear models: a modern perspective. Chapman and Hall/CRC; 2006.
18. Tendani Gaolathe, Wirth Kathleen E, Holme Molly Pretorius, et al. Botswana's progress toward achieving the 2020 UNAIDS 90–90-90 antiretroviral therapy and virological suppression goals: a population-based survey. The lancet HIV. 2016;3(5):e221–e230. [PubMed: 27126489]
19. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing Vienna, Austria2018.
20. Abuelezam NNGaolathe T, Unoda Chakalisa, et al. Estimation of HIV incidence using retrospectively-collected HIV testing history in Botswana. Poster presented at the 21st International AIDS Conference, Durban, South Africa, July 18–22, 2016. Abstract WEPEC137.
21. Botswana Statistics. Botswana AIDS Impact Survey (BAIS) IV. Statistics Botswana. 2013;:26–27.
22. Wirth Kathleen E, Tchetgen Tchetgen Eric J, Murray Megan. Adjustment for missing data in complex surveys using doubly robust estimation: application to commercial sexual contact among Indian men. Epidemiology (Cambridge, Mass.). 2010;21(6):863.
23. Sigal Matthew J, Chalmers R Philip. Play it again: Teaching statistics with Monte Carlo simulation. Journal of Statistics Education. 2016;24(3):136–156.
24. Maurice Kendall, Alan Stuart, Ord Keith J, Arnold Steven. Kendall's Advanced Theory of Statistics: Volume 2A–Classical Inference and and the Linear Model (Kendall's Library of Statistics). A Hodder Arnold Publication,. 1999;.
25. Schouten Rianne Margaretha, Lugtig Peter, Gerko Vink. Generating missing values for simulation purposes: a multivariate amputation procedure. Journal of Statistical Computation and Simulation. 2018;88(15):2909–2930.

TABLE 1

Probabilities assumed for simulating self-reported times since last HIV-negative test conditional on documented ones

	Y_i^*			
Y_i	**0**	**1**	**2**	**3**
0	0.71	0.27	0.01	0.01
1	0.27	0.79	0.16	0.12
2	0.1	0.12	0.73	0.29
3	0.01	0.01	0.10	0.58

Source: Monte Carlo Simulation.

TABLE 2

Monte Carlo simulation results for validated sample-only versus pooled sample estimators for un-weighted and weighted log likelihood functions in presence of selection bias and misclassification error, n=7000

	Un-weighted analysis		Weighted analysis	
Estimate	**Validated-only**	**Pooled**	**Validated-only**	**Pooled**
Monte Carlo Bias	0.0525	0.0591	0.0001	0.0025
Monte Carlo Percent Bias	26.2362	29.5718	0.0669	1.2454
Monte Carlo Mean square error	0.0028	0.0035	0.00005	0.00003
Relative E ciency	93.5667	117.7333	1.5333	1

Source: Monte Carlo Simulation Results.

TABLE 3

Un-weighted and weighted estimates of HIV incident rates per 100 person-years at-risk, standard errors and 95 % confidence intervals for HIV-negative individuals in 1.5 and 5 years prior to BCPP enrolment data.

	Un-weighted analysis		Weighted analysis	
Estimate	**Validated only**	**Pooled**	**Validated only**	**Pooled**
1.5 years prior to BCPP enrolment				
Incidence Rate	1.3509	8.8753	1.1021	1.2665
Standard Error	0.3747	0.5526	0.3310	0.0790
95 % lower limit	0.6166	7.7922	0.5400	1.1220
95 % upper limit	2.0853	9.9583	1.8150	1.4220
5 years prior to BCPP enrolment				
Incidence Rate	1.1410	4.9690	1.0189	1.1671
Standard Error	0.2379	0.3014	0.2370	0.0790
95 % lower limit	0.6747	4.3782	0.5690	1.0260
95 % upper limit	1.6073	5.5597	1.5020	1.3170

Source: BCPP enrolment. Standard errors reported for weighted analyses are from bootstrap samples with replacement.

TABLE 4

Baseline characteristics of N=7221 participants with HIV negative test result prior to in-home HIV testing during the BCPP enrolment survey, overall and according to availability of accompanying test documentation and time since last documented test, Botswana, 2013–2015.

Characteristic	Overall (N=7221)	Availability of documented HIV negative result		
		SR (N=4740)[1]	Documented (N=2378)[2]	Documented (N=1967)[3]
Age at the BHS (n=7221)				
16 to 24 years	1945 (27)	1282 (27)	659 (28)	584 (30)
25 to 34 years	2474 (34)	1631 (34)	821 (35)	698 (35)
35 to 44 years	1124 (16)	718 (15)	394 (17)	323 (16)
45 to 54 years	881 (12)	556 (12)	303 (13)	225 (11)
5 to 64 years	797 (11)	553 (12)	201 (8)	139 (7)
Female (n=7221)	4664 (65)	2967 (63)	1626 (68)	1361 (69)
Pregnant at BHS[4] (n=3681)	215 (6)	60 (3)	154 (12)	151 (13)
Education (n=7183)				
Primary or less	1799 (25)	1173 (25)	565 (24)	430 (22)
Junior secondary	2538 (35)	1670 (35)	845 (36)	706 (36)
Senior secondary	1401 (19)	924 (19)	475 (20)	416 (21)
Higher than senior secondary	1445 (20)	947 (20)	482 (20)	408 (21)
Income per month (n=7168)				
None	3709 (52)	2396 (51)	1252 (53)	1030 (53)
Less than $ 96	1183 (17)	776 (17)	386 (16)	336 (17)
$ 96 to $ 477	1645 (23)	1074 (23)	557 (24)	449 (23)
More than $ 477	631 (9)	452 (10)	172 (7)	144 (7)
Nights spent outside the community, past year (n=7204)				
0 nights	3016 (42)	1904 (40)	1059 (45)	870 (44)
1 to 6 weeks	1565 (22)	1053 (22)	489 (21)	385 (20)
1 to 2 weeks	699 (10)	474 (10)	214 (9)	179 (9)
3 weeks to less than 1 month	795 (11)	569 (12)	223 (9)	194 (10)
1 to 3 months	806 (11)	529 (11)	269 (11)	231 (12)
More than 4 months	323 (4)	201 (4)	118 (5)	105 (5)
Self-reported timing of most recent negative HIV test (n=7072)				
In the last month	378 (5)	100 (2)	277 (12)	
1 to 5 months ago	1337 (19)	649 (14)	683 (29)	680 (35)
6 to 12 months ago	1978 (28)	1262 (27)	712 (30)	676 (34)
More than 12 months ago	3379 (48)	2595 (56)	695 (29)	331 (17)
Age at first sexual intercourse[5] (n=6145)				

Characteristic	Overall (N=7221)	Availability of documented HIV negative result		
		SR (N=4740)[1]	Documented (N=2378)[2]	Documented (N=1967)[3]
10 to 14 years	138 (2)	89 (2)	49 (2)	44 (3)
15 to 17 years	1673 (27)	1105 (28)	548 (26)	458 (26)
18 to 21 years	3559 (58)	2278 (57)	1226 (58)	1025 (58)
22 years or older	781 (13)	490 (12)	283 (13)	229 (13)
Inconsistent condom use, past year[6] (n=5653)	3603 (64)	2366 (64)	1197 (63)	1010 (63)
Transactional sex, past year6 (n=5803)	342 (6)	181 (5)	158 (8)	135 (8)

Source: BCPP enrolment data.

[1] Self Reported without documentation.

[2] Documented within 5 years.

[3] Documented within 1.5 years.

[4] Proportions calculated among female participants.

[5] Proportions calculated among persons reporting any lifetime sexual activity.

[6] Proportions calculated among persons reporting one or more sexual partners during the past years.

TABLE 5

Evidence of misclassification error on BCPP dataset in months

	Self-reported times (Y_i^*)			
Documented times (Y_i)	**Less than 1 month**	**1 to 5 months**	**5 to 12 months**	**more than 12 months**
Less than 1 month	194	56	5	4
1 to 5 months	75	535	106	39
5 to 12 months	3	80	487	96
more than 12 months	3	9	78	192

Source: BCPP enrolment data.

TABLE 6

Sensitivity analysis results

	Weighted sensitivity analysis	
Estimate	**Validated only**	**Pooled**
1.5 years prior to BCPP enrolment		
Incidence Rate	1.1322	1.7358
Standard Error	0.3304	0.1234
95 % lower limit	0.4980	0.1234
95 % upper limit	1.8110	1.9890
5 years prior to BCPP enrolment		
Incidence Rate	0.9974	1.6554
Standard Error	0.2226	0.1107
95 % lower limit	0.6029	1.4460
95 % upper limit	1.4535	1.8790

Source: BCPP enrolment. Standard errors reported for weighted sensitivity analyses are from bootstrap samples with replacement.