Published in final edited form as:

Stat Commun Infect Dis. 2018 December ; 10(1): . doi:10.1515/scid-2017-0003.

Cross-Sectional HIV Incidence Estimation with Missing Biomarkers

Doug Morrison¹, Oliver Laeyendecker², Jacob Konikoff³, and Ron Brookmeyer¹

¹Department of Biostatistics, Fielding School of Public Health, University of California, Los Angeles, Los Angeles, CA, USA

²Laboratory of Immunoregulation, NIAID, NIH, Baltimore, MD, USA and The Division of Infectious Diseases, Department of Medicine, School of Medicine, Johns Hopkins University, Baltimore, Maryland, USA

³Consultant; work completed as postdoctoral fellow, Johns Hopkins University, Baltimore, Maryland, USA

Abstract

Considerable progress has been made in the development of approaches for HIV incidence estimation based on a cross-sectional survey for biomarkers of recent infection. Multiple biomarkers when used in combination can increase the precision of cross-sectional HIV incidence estimates. Multi-assay algorithms (MAAs) for cross-sectional HIV incidence estimation are hierarchical stepwise algorithms for testing the biological samples with multiple biomarkers. The objective of this paper is to consider some of the statistical challenges for addressing the problem of missing biomarkers in such testing algorithms. We consider several methods for handling missing biomarkers for (1) estimating the mean window period, and (2) estimating HIV incidence from a cross sectional survey once the mean window period has been determined. We develop a conditional estimation approach for addressing the missing data challenges and compare that method with two naïve approaches. Using MAAs developed for HIV subtype B, we evaluate the methods by simulation. We show that the two naïve estimation methods lead to biased results in most of the missing data scenarios considered. The proposed conditional approach protects against bias in all of the scenarios.

Keywords

biomarkers; cross-sectional studies; HIV; incidence; missing data

1. Introduction

Methods for estimating HIV incidence are critical for tracking the spread of the HIV/AIDS epidemic and for evaluating prevention interventions (Mastro 2013). The cohort study is the traditional approach for estimating HIV incidence, but that approach is expensive, time-consuming, and presents important methodological and logistical challenges for following key populations at high risk of HIV infection and for documenting HIV acquisition (Brookmeyer 2010a). Considerable progress has been made in the development of an alternative approach for HIV incidence estimation which is based on a cross-sectional survey

for biomarkers of recent infection (Busch et al. 2010). The approach does not require follow-up of cohorts and instead is based on a single representative cross-sectional survey of the target population. The key idea of this approach is to collect biological samples from survey participants which are then tested for biomarkers indicating recent infection (Brookmeyer and Quinn 1995). Early applications of the approach were based on a single serological biomarker of HIV seropositive individuals. However, those efforts were largely unsuccessful because they identified an appreciable number of individuals as “recently” infected who in fact had become infected many years prior. The challenge has been that antibody responses to HIV are also modulated by viral suppression, either natural or through the treatment of antiretroviral drugs and advanced HIV disease which can lead to those with long term infection appearing recently infected (Longosz et al. 2014; Wendel et al. 2017). In recent years, these hurdles have been largely addressed through the use of multiple serological and non-serological biomarkers (see for example, Laeyendecker et al. 2013; Eshleman et al. 2013; Konikoff et al. 2013). Multiple biomarkers when used in combination can increase the precision of cross-sectional HIV incidence estimates. Cross-sectional estimation is now widely used around the world including for Population-Based HIV Impact Assessments.

Multi-assay algorithms (MAAs) for cross-sectional HIV incidence estimation are hierarchical stepwise algorithms for testing the biological samples with biomarkers. These algorithms classify HIV infected individuals into one of two states: either MAA positive or MAA negative. The objective of the MAA classification is for individuals classified as MAA positive to have shorter durations of infections than those individuals classified as MAA negative. At each step of the algorithms, the biological samples are assayed for one of the biomarkers. The biological sample is then provisionally classified as MAA positive and then proceeds to the next step to be assayed for another biomarker. The average duration of time that a newly HIV-infected person would be classified as MAA positive over the entire course of infection is called the mean window period and is denoted by μ . The mean window period depends on the specific MAA under consideration.

Once the mean window period (μ) of an MAA has been estimated, the HIV incidence rate can be estimated from a cross-sectional survey of the target population. An estimator of the HIV incidence is $\hat{I} = X / \mu N^-$, where X is the number of HIV infected individuals in the cross-sectional survey who are classified as MAA positive, and N^- is the number of individuals who are HIV uninfected and thus at risk of HIV infection (Brookmeyer and Quinn 1995). The parameter μ must be estimated in a separate sample from a population similar to the one in which incidence will be estimated.

One key advantage of MAAs is that they can significantly reduce assay costs compared to testing all biological samples with all biomarkers. The cost savings results because only samples provisionally classified as MAA positive need to be tested in the next step of the algorithm with another biomarker in order to determine the MAA classification. Furthermore, the order of the steps of the algorithm can be arranged to help minimize costs: the less expensive or less labor intensive assays can be performed in the initial steps, and the most expensive assays performed in the final step. By so doing, the number of biological samples that need to be tested in the final step is relatively small. For example, viral load testing is often placed at the final step of MAAs. MAAs have been developed for subtype B

infection based on various combinations of the following biomarkers: BED capture enzyme immunoassay (BED-CEIA), CD4 cell count, BioRad Avidity assay, Limiting Antigen Avidity assay (LAG-Avidity assay) and viral load (Laeyendecker et al. 2013; Konikoff et al. 2013).

Missing biomarker measurements are an important problem with the use of MAAs for cross-sectional HIV incidence estimation. The missing biomarkers measurements arise for several reasons. First and foremost, inherent in the use of MAAs is the idea that some biological samples can be classified as MAA negative without having to perform assays for all the biomarkers. Second, biomarker measurements could be missing because there is not sufficient biological sample material to adequately perform the assays. Measurements may not be performed because of the high costs of some of the assays. In addition, measurements may be missing because the biological samples have been lost and because of other logistical and administrative reasons.

The objective of this paper is to consider some of the statistical challenges for addressing missing biomarker data for cross-sectional HIV incidence estimation. In section 2 we consider several methods for handling missing biomarkers when estimating the mean window period of an MAA based on biological samples with approximately known durations of infection. In section 3 we evaluate the performance of these various methods by simulation. We examine two naïve approaches, one using all samples that can be classified by the MAA and another using all samples with complete biomarker data, and show that each of these approaches can lead to biased estimators of the mean window period. We propose a conditional approach for handling the missing data. The main idea of the conditional approach with two biomarkers is to decompose the likelihood into a product of two factors. The first factor corresponds to the first biomarker result and the second factor corresponds to the second biomarker result conditional on the first biomarker result. This approach accounts for the possibility that the missingness probability for a given biomarker may depend on the values of the other biomarkers in the MAA. We show that this method performs well. Once an MAA has been developed and its mean window period has been estimated, it is ready to be used in a cross-sectional survey to estimate HIV incidence. However, missing biomarker data can also occur in the survey for any of the reasons mentioned previously. In section 4 we consider approaches for handling missing biomarkers in a cross-sectional survey for estimating HIV incidence, and in section 5 we evaluate these approaches by simulation. We again show that several naïve estimators can lead to biased results, whereas a conditional approach performs well. The results are discussed in section 6.

2. Estimation of the Mean Window Period of MAAs with Missing Biomarkers

In this section we consider approaches for handling missing biomarker data when estimating the mean window period of an MAA. Here we consider algorithms consisting of two biomarkers assayed sequentially, where missing biomarker data is only possible for the second biomarker in the algorithm. One example of a two biomarker MAA is a widely used algorithm based on LAG-Avidity assay and viral load for which MAA positivity requires

LAg-Avidity < 1.5 normalized optical density units (OD-n) and viral load > 1000 copies/mL (Rehle et al. 2015). Another two assay MAA developed for HIV subtype B is based on the LAg-Avidity and BioRad avidity assays, for which MAA positivity requires LAg-Avidity < 2.9 OD-n and BioRad $< 40\%$ (Konikoff et al. 2013).

We label the two biomarkers B_1 and B_2 where biomarker B_2 is the one with potentially missing values. We use the notation m_i to indicate whether the value of the biomarker B_2 is missing for the i^{th} sample, that is, $m_i = 1$ if the B_2 measurement is missing and $m_i = 0$ if the measurement is known. We define the indicator random variables b_{1i} and b_{2i} to indicate whether the biomarkers meet the criteria for MAA positivity for biomarkers B_1 and B_2 respectively. The random variable b_{2i} is only observed if $m_i = 0$. For example, in the two-assay MAA mentioned above (LAg and viral load), b_{1i} is set to 1 if LAg < 1.5 and 0 otherwise, and b_{2i} is set to 1 if viral load > 1000 , to 0 if viral load ≤ 1000 and is not observed if $m_i = 1$. We use Z_i to indicate the MAA classification for the i^{th} sample based on the observed data, where $Z_i = +1$ if the MAA is positive (which occurs if $b_{1i} = b_{2i} = 1$), $Z_i = -1$ if the MAA is negative (which occurs if either $b_{1i} = 0$ or $b_{2i} = 0$), and $Z_i = 0$ if the MAA classification is indeterminate (which occurs if $b_{1i} = 1$ and $m_i = 1$).

Figure 1 shows the classification of biological samples according to MAA status (MAA positive, negative and indeterminate) and biomarker B_2 missing status ($m_i = 0$ or 1). As shown in the figure, some biological samples (the red squares) can be classified as MAA negative ($Z_i = -1$) even though biomarker B_2 is missing. The blue squares correspond to samples with $m_i = 0$, that is, samples for which both biomarkers have been measured. The green square corresponds to samples where the MAA status could not be determined (that is, where $b_{1i} = 1$ and $m_i = 1$).

Typically, the data available for window period estimation are biological samples from HIV positive individuals for whom some information is known about their durations of infection. The data from the i^{th} biological sample is $(L_i, R_i, m_i, b_{1i}, b_{2i})$. Here L_i and R_i are the left and right bounds respectively for the duration of HIV infection at the time of biological sample collection for the individual corresponding to that biological sample; for example, L_i is the duration from the last HIV antibody negative test to collection of the biological sample, and, R_i is the duration from the first HIV antibody positive test to collection of the biological sample. Various statistical methods have been proposed for dealing with the complication that the duration of infection may not be known exactly, including methods for interval censored data and multiple imputation. Suppose the infection duration T_i is imputed by generating a random value from the uniform distribution over the interval $[L_i, R_i]$, and further suppose there is no missing biomarker data. In that situation, we can fit a model $\phi(t) = P(b_{1i} = b_{2i} = 1 | T_i = t)$ such as a logistic regression model using all the biological samples (Konikoff 2015). Then the mean window is estimated by $\hat{\mu} = \int_{t=0}^{\infty} \hat{\phi}(t) dt$ provided $\phi(t)$ converges to 0 as t goes to ∞ (Brookmeyer, Laeyendecker, et al. 2013). Here our focus is on addressing the statistical challenges for estimating the mean window period with incomplete biomarker data. The missing data approaches we investigate can be used in combination with various approaches for handling imprecise measurements of the durations of infection.

We will evaluate several methods for estimating the mean window period under a class of missingness mechanisms applicable to MAAs. We suppose that the probability that the biomarker B_2 's measurement is missing, given $b_{1i} = 0$, is a constant, i.e. $P(m_i = 1 | b_{1i} = 0) = \lambda_1$, and similarly the probability the biomarker B_2 's measurement is missing, given $b_{1i} = 1$, is $P(m_i = 1 | b_{1i} = 1) = \lambda_2$. In the practical application of MAAs we would typically expect $\lambda_1 > \lambda_2$, because biological samples that meet the MAA criterion for the first biomarker are more likely to be assayed for the second biomarker. If the biomarker tests were run strictly sequentially, then $\lambda_1 = 1$. However, in practice some biological specimens may have been evaluated for biomarker B_2 even if $b_{1i} = 1$, because of other research and clinical requirements. In that case λ_1 may be less than 1. In this paper, while we are considering primarily sequentially run evaluations of the biomarkers, we also allow for the possibility that some measurements may be available on B_2 even if $b_{1i} = 0$.

It is interesting to note that even if the probability that B_2 is missing does not depend on B_1 (i.e. $\lambda_1 = \lambda_2 = \lambda$), the probability that the MAA classification is indeterminate still depends on the values of the biomarkers. To see that, note that $P(Z_i = 0 | b_{1i} = 1) = P(m_i = 1 | b_{1i} = 1) = \lambda_2$, but on the other hand, we have $P(Z_i = 0 | b_{1i} = 0) = 0$. It is also worth noting that although $Z_i = -1$ whenever $b_{1i} = 0$, regardless of whether b_{2i} is observed, the value of $\lambda_1 = P(m_i = 1 | b_{1i} = 0)$ still affects the simulation results below, because the second approach considered removes all observations with b_{2i} missing, even when the MAA classification is already determined by $b_{1i} = 0$; we will find that this approach leads to biased results specifically when $\lambda_1 > \lambda_2$.

The first approach that we consider for window period estimation (Method WP1) uses all samples whose MAA classification has been determined (i.e., Z takes the values of +1 or -1) and ignores those samples whose MAA status is undetermined. We thus refer to this approach as the “MAA determined analysis”. The data that would go into the analysis are the red and blue squares in Figure 1. We could fit a flexible logistic regression model for $P(Z_i = 1)$, such as a cubic spline as a function of the time since seroconversion to the data (Z_i, T_i) where Z_i is either -1 or +1. The resulting fitted predicted probability curve, denoted $\hat{\phi}(t)$, is integrated from 0 to ∞ to estimate the mean window period; i.e. $\hat{\mu} = \int_{t=0}^{\infty} \hat{\phi}(t) dt$.

Method WP1 appears simple and straightforward, because it uses all the data where the MAA classification can be determined, but it is a naïve analysis that can lead to biased estimates if $\lambda_2 > 0$. We quantify this bias by simulation in section 3.

The second approach we consider (Method WP2) uses all samples where both biomarkers have been measured; that is, sample i will be excluded from the analysis whenever $m_i = 1$, even if $b_{1i} = 0$, in which case the sample can be classified as MAA-negative. Thus the data that would go into this analysis are only the blue squares in Figure 1. We refer to this approach as the “complete biomarker analysis” as it uses only samples where both biomarkers have been measured. The intuition for this approach is that we should remove all samples with incomplete data, rather than only samples with B_2 missing and $B_1 = 1$, because the subset of samples with $B_1 = 1$ has a higher proportion of MAA-positives than the overall population; therefore, removing incomplete observations from only this subset induces bias. In contrast, if the probability of missingness is not associated with the underlying biomarker

values (implying $\lambda_1 = \lambda_2$), then removing all incomplete samples leaves us with an unbiased sample of biomarker values. However, method WP2 is still biased whenever $\lambda_1 \neq \lambda_2$. As we show in section 3, even if $\lambda_2 = 0$, the method will lead to biased results when $\lambda_1 > 0$.

The third approach we consider (Method WP3) uses all the biological samples illustrated in Figure 1 (blue, red, and green squares). We call this approach a conditional analysis. The factor that is contributed to the likelihood function for the i^{th} individual is:

$$\mathcal{L}_i =$$

$$\left\{ \begin{array}{ll} \left[1 - P(b_{1i} = 1 | T_i = t) \right] \left(1 - \lambda_1 \right) P(b_{2i} = 1 | b_{1i} = 0, T_i = t)^{b_{2i}} \left[1 - P(b_{2i} = 1 | b_{1i} = 1 | b_{1i} = 0, T_i = t) \right]^{1-b_{2i}} & \text{if } b_{1i} = 0, m_i = 0 \\ P(b_{1i} = 1 | T_i = t) \left(1 - \lambda_2 \right) P(b_{2i} = 1 | b_{1i} = 1, T_i = t)^{b_{2i}} \left[1 - P(b_{2i} = 1 | b_{1i} = 1, T_i = t) \right]^{1-b_{2i}} & \text{if } b_{1i} = 1, m_i = 0 \\ \left[1 - P(b_{1i} = 1 | T_i = t) \right] \lambda_1 & \text{if } b_{1i} = 0, m_i = 1 \\ P(b_{1i} = 1 | T_i = t) \lambda_2 & \text{if } b_{1i} = 1, m_i = 1 \end{array} \right.$$

(1)

We use two different logistic regression models $\phi_1(t; \alpha)$ and $\phi_2(t; \beta)$ for $P(b_{1i} = 1 | T_i = t)$ and $P(b_{2i} = 1 | b_{1i} = 1, T_i = t)$ respectively, where $\alpha = (\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4)$ and $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$, are the regression parameters:

$$\begin{aligned} \phi_1(t; \alpha) &= P(b_{1i} = 1 | T_i = t) = \text{logit}^{-1} \left(\alpha_0 + \alpha_1 t_i + \alpha_2 t_i^2 + \alpha_3 t_i^3 + \alpha_4 \left(I_{t_i > 2} (t_i - 2)^3 \right) \right) \\ \phi_2(t; \beta) &= P(b_{2i} = 1 | b_{1i} = 1, T_i = t) = \text{logit}^{-1} \left(\beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \beta_3 t_i^3 + \beta_4 \left(I_{t_i > 2} (t_i - 2)^3 \right) \right) \end{aligned}$$

We further assume that $P(b_{2i} = 1 | b_{1i} = 0, T_i = t)$, λ_1 , and λ_2 do not involve the parameters α or β . Then it follows, after rearranging and collecting terms, that the likelihood contribution, up to a proportionality constant $K(\lambda_1, \lambda_2, m_i, b_{1i}, b_{2i})$, is the product of two components:

$$\mathcal{L}_i = K \left(\lambda_1, \lambda_2, m_i, b_{1i}, b_{2i} \right) \left[\phi_1(t_i | \alpha)^{b_{1i}} (1 - \phi_1(t_i | \alpha))^{1-b_{1i}} \right] \left[\phi_2(t_i | \beta)^{b_{2i}} (1 - \phi_2(t_i | \beta))^{1-b_{2i}} \right]^{(1-m_i)b_{1i}} \quad (2)$$

Thus, the first component of the likelihood corresponds to the likelihood for the binary outcome b_{1i} . The second component of the likelihood corresponds to the binary outcome b_{2i} conditional on $b_{1i} = 1$ among biological samples where b_{2i} is not missing, that is where $m_i = 0$. Thus, in practice, to find the maximum likelihood estimates of α and β we (1) fit a logistic regression model for the outcome b_1 using all the data, and (2) fit a separate logistic regression model for the outcome b_2 conditional on $b_1 = 1$, using the data where b_2 is observed and $b_1 = 1$. Then, we estimate $P(b_1 = b_2 = 1 | T_i = t)$ by $\hat{\phi}(t) = \hat{\phi}_1(t | \alpha) \hat{\phi}_2(t | \beta)$.

Finally, the mean window period is estimated by $\hat{\mu} = \int_{t=0}^{\infty} \hat{\phi}(t) dt$ provided $\hat{\phi}(t)$ converges.

The intuition for this approach is that we have a complete dataset for estimating $\phi_1(t | \alpha)$, since this model does not involve b_2 . We can also estimate $\phi_2(t | \beta)$ without bias, using the samples where b_2 is observed (and $b_1 = 1$), assuming that conditional on $b_1 = 1$, the probability that b_2 is missing is a constant, i.e. $P(m_i = 1 | b_{1i} = 1) = \lambda_2$. Therefore, the resulting estimator $\hat{\mu}$ will be unbiased in scenarios where WP1 and WP2 produce biased estimates. Approach WP3 does not require the estimation of parameters for λ_1 or λ_2 ; it only requires the assumption that these parameters are constant with respect to b_2 .

3. Simulation Study of Window Period Estimation Methods with Missing Biomarkers

We assessed the abilities of window period methods WP1-WP3 to estimate μ using the following simulation framework. We first selected a dataset which had complete biomarker data for LAg-Avidity and viral load (Konikoff et al. 2013). This dataset contained 1780 observations from 709 participants of 3 longitudinal cohort studies in the US; participants contributed between 1 and 16 observations, at estimated times since seroconversion ranging between 1 month and >8 years. We then generated 1000 simulated datasets, sampling observations with replacement from the original dataset, clustering by subject ID and stratifying by study cohort, so that each simulated dataset had the same number of sampled subjects in each cohort as the original dataset had. We assigned subjects a new ID number each time they were sampled. We then imputed the infection duration T_i for each ID number by a draw from a uniform distribution over the interval $[L_i, R_i]$, where L_i and R_i are the left and right bounds for the duration of HIV corresponding to that biological sample (Brookmeyer, Konikoff, et al. 2013).

We then performed the following analysis on each simulated dataset. First, we produced recency classifications ($Z_i = +1$ or $Z_i = -1$) using the MAA consisting of LAg-Avidity < 1.5 and viral load > 1000. We used these recency classifications, along with the imputed times

since seroconversion, to produce estimates of the mean window period ($(\hat{\mu}_{j,T})$), using the procedure for completely-observed data described above. Thus, in the simulation study, the estimated mean window period of each simulated dataset is used as the “gold standard.” We considered these estimates to be the “gold standards” for each simulated dataset because they were the estimates which would be achieved with completely observed data, and which methods WP1-3 should attempt to reproduce.

We randomly assigned the biomarker B_2 to be missing by simulating Bernoulli trials for each serum sample in the dataset. The Bernoulli trial probability that B_2 was assigned to be missing depended solely on whether B_1 's value met its criterion for MAA positivity, i.e. whether $b_1 = 1$. We set $P(m_i = 1 | b_{1i} = 0) = \lambda_1$ and $P(m_i = 1 | b_{1i} = 1) = \lambda_2$, and varied the values of λ_1 and λ_2 across six different simulation scenarios (Table 1).

After determining which observations were missing, we recalculated the MAA classifications Z_i , updating the value to $Z_i = 0$ when $b_{1i} = 1$ and $m_i = 1$, and then applied methods WP1, WP2 and WP3 to the resulting incomplete datasets, producing estimates $\hat{\mu}_{j,k}$ for the j^{th} simulated dataset using the k^{th} method (where $j = 1, \dots, 1000$ and $k = 1, 2, 3$). We computed error scores $e_{j,k} = \hat{\mu}_{j,k} - \hat{\mu}_{j,T}$ comparing each gold standard estimate with the corresponding incomplete-data estimates. We defined the bias of estimator $\hat{\mu}_k$ as the expectation of the error score for the k^{th} method, $E[e_{j,k}]$, and we estimated this bias using the sample mean $\bar{e}_k = \frac{1}{n} \sum_{j=1}^n e_{j,k}$. We also calculated sample standard deviations of these error scores, $s_k = \sqrt{\frac{1}{n-1} \sum_{j=1}^n (e_{j,k} - \frac{1}{n} \sum_{j=1}^n e_{j,k})^2}$, as well as the sample standard error of each estimator $\hat{\mu}_k$, $SE(\hat{\mu}_k) = \sqrt{\frac{1}{n-1} \sum_{j=1}^n (\hat{\mu}_{j,k} - \frac{1}{n} \sum_{j=1}^n \hat{\mu}_{j,k})^2}$.

The simulation results are shown in Table 1. As λ_2 increased across the simulation scenarios, Method WP1 produced increasingly negatively biased estimates of μ , and correspondingly positively biased incidence estimates. The value of $\lambda_1 = P(m_i = 1 | b_{1i} = 0)$ did not affect this method's results, because all of the samples with $b_{1i} = 0$ could be classified as $Z_i = -1$, regardless of whether B_2 was missing. Method WP2 had substantial positive bias for μ in scenarios with $\lambda_2 < \lambda_1$. In scenarios where λ_1 approximately equaled λ_2 , this method had less bias. Method WP3 had relatively small biases across all the scenarios studied, but did have a consistent, modestly negative bias for μ . Relative to the naïve methods, the conditional approach reduced bias in the mean window period estimates by over 80% in several scenarios considered; for example, in the scenario with $\lambda_1 = 0.5$ and $\lambda_2 = 0$, method WP3 had 89% smaller magnitude of bias than WP2.

The simulation results indicate that Method WP1 produced biased estimates of μ in the scenarios in which a substantial fraction of the samples could not be classified, even in the idealized scenarios in which the probability that B_2 was missing did not depend on B_1 (i.e. $\lambda_1 = \lambda_2$). Method WP2 was approximately unbiased only if $\lambda_1 \approx \lambda_2$. The bias was small with Method WP3 for all scenarios considered.

4. Methods for Cross-Sectional Incidence Estimation with Missing Biomarkers

The previous section demonstrated that if missing biomarkers are not handled appropriately, estimates of the mean window period can be biased. In this section, we consider estimation of HIV incidence from a cross-sectional sample using an MAA with missing biomarkers. In what follows, it is assumed that the mean window period has been previously estimated appropriately using methods such as Method WP3 described in Section 2.

As in section 2, we consider the situation in which an MAA uses two biomarkers, B_1 and B_2 , and only B_2 has missing values. If both biomarkers were observed for all persons, the incidence estimate would be $\hat{I} = X / (\mu N^-)$, where $X = \sum_{i=1}^{N^+} 1_{(Z_i=1)}$ is the number of persons in the cross-sectional survey who are classified as MAA positive, $1_{(Z_i=1)}$ is an indicator function with value 1 if $Z_i = 1$ and 0 otherwise, N^- is the number of seronegative persons in the survey, N^+ is the number of seropositive persons in the survey, and μ is the (previously estimated) mean window period. We evaluate several methods for estimating the incidence rate under the same class of missingness mechanisms as in section 2.

The first cross-sectional incidence approach (Method CS1), which we call the “MAA determined analysis”, uses all samples whose MAA status can be determined (i.e. for which $Z_i = \pm 1$, the red and blue squares in Figure 1) and ignores those samples whose MAA status is indeterminate ($Z_i = 0$, green square in Figure 1). Method CS1 uses this subset to estimate the proportion of seropositive samples that would have been classified as MAA positive if there were no missing values, using the estimator $\hat{p}_1 = X / n_1$ where $n_1 = \sum_{i=1}^{N^+} |Z_i|$ is the number of seropositive samples determined as MAA positive or MAA negative. Then the incidence is estimated by:

$$\hat{I}_1 = \frac{\hat{p}_1 N^+}{\mu N^-} = \left(\frac{X}{\sum_{i=1}^{N^+} |Z_i|} \right) \left(\frac{N^+}{\mu N^-} \right) \quad (3)$$

As before, Method CS1 appears simple and straightforward, because it uses all the samples for which the MAA classification can be determined. However, it turns out to be a naïve analysis that can lead to severely biased estimates if $\lambda_2 > 0$, and we quantify the bias by simulation in section 5.

The second approach (Method CS2) which we call the complete biomarker analysis, uses all samples where both biomarkers have been measured (blue squares in Figure 1). Method CS2 estimates $P(b_{1i}=b_{2i}=1)$ by $\hat{p}_2 = X / n_2$, where $n_2 = \sum_{i=1}^{N^+} (1 - m_i)$ is the number of samples with no missing biomarker values. Then, the incidence rate estimate is:

$$\hat{I}_2 = \frac{\hat{p}_2 N^+}{\mu N^-} = \left(\frac{X}{\mu N^-} \right) \left(\frac{N^+}{\sum_{i=1}^{N^+} (1 - m_i)} \right) \quad (4)$$

This method is also biased whenever $\lambda_1 \neq \lambda_2$, as we will show in section 5.

The third approach (Method CS3), which we call the conditional analysis, uses all of the biological samples in Figure 1 (blue, red, and green squares). The idea of the approach is to estimate $P(b_{1i} = b_{2i} = 1)$ as a product of $P(b_{1i} = 1)$ and the conditional probability $P(b_{2i} = 1 \mid b_{1i} = 1)$. First, we estimate $P(b_{1i} = 1)$ by $\hat{p}_{3a} = \sum_{i=1}^{N^+} b_{1i} / N^+$. Second, we estimate $P(b_{2i} = 1 \mid b_{1i} = 1)$ by $\hat{p}_3 = X / n_{3b}$, where $n_{3b} = \sum_{i=1}^{N^+} b_{1i} (1 - m_i)$ is the number of samples where $b_1 = 1$ and the second biomarker B_2 is observed. Then the estimator of $P(b_{1i} = b_{2i} = 1)$ is $\hat{p}_3 = \hat{p}_{3a} \hat{p}_{3b}$. It follows that the estimator of incidence is:

$$\hat{I}_3 = \frac{\hat{p}_3 N^+}{\mu N^-} = \left(\frac{X}{\mu N^-} \right) \left(\frac{\sum_{i=1}^{N^+} b_{1i}}{\sum_{i=1}^{N^+} b_{1i} (1 - m_i)} \right) \quad (5)$$

5. Simulation Study of Cross-Sectional Incidence Estimation Methods with Missing Biomarkers

We assessed the abilities of the three methods to estimate incidence using the following simulation framework. We used the same source dataset of N^+ seropositive samples as in the previous simulation. We then generated 1000 simulated datasets, by sampling N^+ observations from this original dataset, with replacement for each simulation.

We then performed the following analysis on each simulated dataset j . First, we produced recency classifications ($Z_i = +1$ or $Z_i = -1$) using the same MAA as in the previous section.

We determined the (True) count of MAA-positive samples, denoted $X_{j,T} = \sum_{i=1}^{N^+} b_{1i} b_{2i}$.

As above, we introduced missing values in B_2 by generating a Bernoulli trial for each serum sample in the dataset. We considered the same six missingness scenarios as in Table 1.

After determining which observations were missing, we recalculated the MAA classifications Z_i , updating the value to $Z_i = 0$ when $b_{1i} = 1$ and $m_i = 1$, and then we applied the three methods to the resulting incomplete dataset to produce estimates $\hat{p}_{j,k}$ for the j^{th} simulated dataset using the k^{th} method (where $j = 1, \dots, 1000$ and $k = 1, 2, 3$).

We computed relative error scores $r_{j,k} = (\hat{I}_{j,k} - \hat{I}_{j,T}) / \hat{I}_{j,T}$ for the j^{th} simulation using the k^{th} method using the fact that $\hat{I}_{j,k} = \hat{p}_{j,k} N^+ / (\mu N^-)$ and $\hat{I}_{j,T} = X_{j,T} / (\mu N^-)$ share the term $1/(\mu N^-)$; thus the relative error reduces to $r_{j,k} = (\hat{p}_{j,k} N^+ - X_{j,T}) / X_{j,T}$. We defined the relative bias of estimator $\hat{I}_k$ as the expectation of the relative error score for the k^{th} method, $E[r_{j,k}]$, and we estimated this bias using the sample mean $\bar{r}_k = \frac{1}{n} \sum_{j=1}^n r_{j,k}$. We also calculated the sample standard deviations of these error scores,

$$s_k = \sqrt{\frac{1}{n-1} \sum_{j=1}^n \left(r_{j,k} - \frac{1}{n} \sum_{j=1}^n r_{j,k} \right)^2}.$$

The results are shown in Table 2. We find that as λ_2 increased across the simulation scenarios, Method CS1 produced increasingly negatively biased incidence estimates (table 2). The value of λ_1 did not affect this method's results, because all of the samples with $b_1 = 0$ can be classified as $Z = -1$, regardless of whether B_2 is missing. Method CS2 estimated incidence with substantial positive bias in scenarios with $\lambda_2 < \lambda_1$. In scenarios where $\lambda_1 \approx \lambda_2$, this method showed less bias. Method CS3 had minimal biases across all the scenarios simulated.

In conclusion, Method CS1 was substantially biased for incidence estimation in the scenarios in which a substantial fraction of the samples could not be classified, even under the optimistic assumption that B_2 's missingness probability did not depend on B_1 ($\lambda_1 = \lambda_2$). Method CS2 was approximately unbiased only if $\lambda_1 \approx \lambda_2$. Method CS3 was nearly unbiased in all scenarios considered.

6. Discussion

The objective of this paper was to consider some of the statistical challenges posed by incomplete biomarker data for cross-sectional HIV incidence estimation. We examined this problem both in the context of estimating the mean window period of an MAA and in the context of estimating incidence from a cross-sectional survey using an MAA with a previously estimated mean window period. We evaluated three methods for handling missing data in each of these contexts, simulating the methods' performance across a range of six missingness mechanism scenarios.

Our main findings were that the "MAA determined" and "complete biomarker" methods (WP1/CS1 and WP2/CS2, respectively) produced substantially biased results in a range of plausible missingness conditions. The "MAA determined" methods had bias magnitudes associated with the fraction of samples that could not be classified by the MAA, whereas the "complete biomarker" methods had more bias when the probability of missing B_2 values depended strongly on the value of B_1 . In contrast, the conditional methods which we proposed (WP3 and CS3) produced accurate results in all of the scenarios that we considered. However, we do note that while the conditional WP3 approach had small bias, it was consistently negative. Further study to identify any situations when the bias would be positive would be useful.

The results in this work highlight the perhaps surprising fact that even if the probability of missing biomarkers is a constant, simply ignoring unclassifiable samples can lead to bias. This result is due to the asymmetry between the positive (recent) and negative classifications; one requires the joint occurrence (intersection) of several events, while the other requires only that at least one event occur (union). Thus the MAA classification can only be missing if all of the observed biomarkers indicate a positive result. Our proposed conditional analysis accounts for this asymmetry by modeling each biomarker separately. It thus avoids the biases that affect the other two methods.

There are several important extensions to this work worth considering in future research. First, we only assessed methods for estimating the mean window period and incidence rate. There are other MAA characteristics whose incomplete-data estimation should be studied, such as the shadow (Kaplan and Brookmeyer 1999; Brookmeyer 2010b), False Recent Rate (FRR), and Mean Duration of Recent Infection (MDRI) (Kassanjee et al. 2012). Since these characteristics are all functions of the probability curve $\phi(t)$, we expect that “MAA determined” and “complete biomarker” analyses for these characteristics would produce biased estimates as well. We expect that our proposed conditional approach would produce approximately unbiased estimates since it estimates $\phi(t)$ without bias.

Another complication to consider is that the probability that B_2 is missing may depend on a different dichotomization of B_1 from the one used by the MAA. For example, in a dataset encountered in recent work, viral load was assayed with probability ≈ 1 if LAg-Avidity was below 3.0 units, and was assayed with a smaller probability otherwise. With such a dataset, when using Method WP3 and the LAg-Avidity < 1.5, viral load > 1000 MAA to estimate μ or incidence, we could condition on $B_1 < c$, where c is any value between the missingness cutoff (3.0) and the MAA cutoff (1.5), rather than conditioning on $b_{1j} = 1$. It seems that the optimal choice of c would be the largest value, in order to maximize the number of observations used to estimate the conditional model $P(b_2 = 1 \mid B_1 < c)$; smaller cutoffs might lead to instability in the estimation of the conditional model.

Finally, more complex scenarios, including MAAs using more than two biomarkers and MAAs with missing data in more than one biomarker, may require more sophisticated analyses. We only considered missingness in viral load, which is a relatively expensive biomarker. As new generations of potentially expensive assays for detecting incident HIV infections are developed, it seems likely that situations will arise in which multiple biomarkers have missing values, due to scientific or resource constraints. It will be important to further develop and refine methods for addressing these challenges.

Acknowledgement

The authors thank the following investigators for providing or generating data analyzed in this study: Caroline E. Mullis, Matthew M. Cousins, Thomas C. Quinn, Deborah Donnell, Connie Celum, Susan P. Buchbinder, George R. Seage III, Lisa P. Jacobson, Joseph B. Margolick, Joelle Brown, Gregory D. Kirk, Shruti H. Mehta, Richard D. Moore, and Jeanne C. Keruly. The authors gratefully acknowledge the data provided by: The HIVNET 001 Study funded by the HIV Network for Prevention Trials (HIVNET) and sponsored by the NIAID; the ALIVE Study funded by the NIDA; the MACS Study funded by the NIAID with additional supplemental funding from the National Cancer Institute (NCI) and National Heart, Lung, and Blood Institute (NHLBI); the Johns Hopkins HIV Clinical Practice Cohort was funded by NIDA, the National Institute of Alcohol Abuse and Alcoholism (NIAAA) and NIAID.

This work was supported by R01-AI095068 (DM, JK, RB sponsored by NIAID of the National Institutes of Health (NIH), and the Division of Intramural Research, NIAID (OL).

References

- Brookmeyer R Measuring the HIV/AIDS Epidemic: Approaches and Challenges *Epidemiologic Reviews* 2010a 32 1 26–37 [PubMed: 20203104]
- Brookmeyer R On the Statistical Accuracy of Biomarker Assays for HIV Incidence *Journal of Acquired Immune Deficiency Syndrome* 2010b 54 4 406–14
- Brookmeyer R Konikoff J Laeyendecker O et al. Estimation of HIV Incidence Using Multiple Biomarkers *American Journal of Epidemiology* 2013 177 3 264–72 [PubMed: 23302151]
- Brookmeyer R Laeyendecker O Donnell D et al. Cross-Sectional HIV Incidence Estimation in HIV Prevention Research *Journal of Acquired Immune Deficiency Syndrome* 2013 63 0 2 1–13
- Brookmeyer R Quinn T Estimation of Current Human Immunodeficiency Virus Incidence Rates from a Cross-Sectional Survey Using Early Diagnostic Tests *American Journal of Epidemiology* 1995 141 2 166–72 [PubMed: 7817972]
- Busch MP Pilcher CD Mastro TD et al. Beyond Detuning: 10 Years of Progress and New Challenges in the Development and Application of Assays for HIV Incidence Estimation *AIDS* 2010 24 18 2763–71 [PubMed: 20975514]
- Eshleman SH Hughes JP Laeyendecker O et al. Use of a Multifaceted Approach to Analyze HIV Incidence in a Cohort Study of Women in the United States: HIV Prevention Trials Network 064 Study *Journal of Infectious Diseases* 2013 207 2 223–31 [PubMed: 23129758]
- Kaplan EH Brookmeyer R Snapshot Estimators of Recent HIV Incidence Rates *Operations Research* 1999 47 1 29–37
- Kassanjee R McWalter TA Bärnighausen T et al. A New General Biomarker-Based Incidence Estimator *Epidemiology* 2012 23 5 721–28 [PubMed: 22627902]
- Konikoff, J, Ph.D. Dissertation. University of California at Los Angeles; 2015. Cross-Sectional HIV Incidence Estimation: Techniques and Challenges.
- Konikoff J Brookmeyer R Longosz AF et al. Performance of a Limiting-Antigen Avidity Enzyme Immunoassay for Cross-Sectional Estimation of HIV Incidence in the United States *PLoS ONE* 2013 8 12 1–9
- Laeyendecker O Brookmeyer R Cousins MM et al. HIV Incidence Determination in the United States: A Multiassay Approach *Journal of Infectious Diseases* 2013 207 2 232–39 [PubMed: 23129760]
- Longosz AF Mehta SH Kirk GD et al. Incorrect Identification of Recent HIV Infection in Adults in the United States Using a Limiting-Antigen Avidity Assay *AIDS* 2014 28 8 1227–32 [PubMed: 24513567]
- Mastro TD Determining HIV Incidence in Populations: Moving in the Right Direction *Journal of Infectious Diseases* 2013 207 2 204–6 [PubMed: 23129757]
- Rehle T, Johnson L, Hallett T. et al. A Comparison of South African National HIV Incidence Estimates: A Critical Appraisal of Different Methods. *PLoS ONE*. 2015; 10(7)
- Wendel SK Longosz AF Eshleman SH et al. Short Communication: The Impact of Viral Suppression and Viral Breakthrough on Limited-Antigen Avidity Assay Results in Individuals with Clade B HIV Infection *AIDS Research and Human Retroviruses* 2017 33 4 325–27 [PubMed: 27875908]

Figure 1:

Illustration of a Two-Biomarker Multi-Assay Algorithm (MAA) with missing data on B_2 . MAA positive, negative, and indeterminate are indicated by $Z = +1, -1$, and 0 respectively. Methods WP1 and CS1 use all samples where Z is determined (red and blue squares); methods WP2 and CS2 use samples only if B_2 is observed (blue squares); method 3 uses all samples (red; blue; and green squares where MAA is indeterminate).

Table 1:
Performance of 3 methods for handling missing biomarkers for estimation of the Mean Window Period

λ_1	λ_2	Method	Estimated bias (mean error, days)*	$SD(error)$ (days)**	$SE(\hat{\mu}_k)$ (days)***
0.25	0.25	WP1	-24	8	17
		WP2	0	9	20
		WP3	-8	8	18
0.50	0.50	WP1	-53	12	16
		WP2	-1	17	25
		WP3	-9	13	20
0.50	0.25	WP1	-24	8	17
		WP2	40	15	27
		WP3	-8	8	18
0.50	0	WP1	0	0	18
		WP2	74	13	28
		WP3	-8	5	17
0.20	0.10	WP1	-9	5	17
		WP2	11	6	20
		WP3	-8	6	17
0.10	0.05	WP1	-4	3	18
		WP2	5	4	19
		WP3	-8	6	17

* Error is defined as $e_{j,k} = \hat{\mu}_{j,k} - \hat{\mu}_{j,T}$ in the j^{th} simulation using method k , where $j = 1, \dots, 1000$ and $k = 1, 2, 3$, and T indicates the result if there were no missing values. Bias is defined as the expectation $E[e_k]$ and is estimated by $\bar{e}_k = 1 / 1000 \sum_{j=1}^{1000} e_{j,k}$.

$$^{**} SD(error) = \sqrt{\frac{1}{n-1} \sum_{j=1}^n \left(e_{j,k} - \frac{1}{n} \sum_{j=1}^n e_{j,k} \right)^2}.$$

$$^{***} SD(\hat{\mu}_k) = \sqrt{\frac{1}{n-1} \sum_{j=1}^n \left(\hat{\mu}_{j,k} - \frac{1}{n} \sum_{j=1}^n \hat{\mu}_{j,k} \right)^2}.$$

Table 2:
Performance of 3 methods for handling missing biomarker data for estimation of HIV incidence from cross-sectional surveys

λ_1	λ_2	Method	Estimated bias (mean relative error, %)*	SD(relative error) (%)**
0.25	0.25	CS1	-22	5
		CS2	0.4	6
		CS3	0.1	5
0.50	0.50	CS1	-47	6
		CS2	0.007	11
		CS3	-0.3	9
0.50	0.25	CS1	-22	5
		CS2	42	9
		CS3	0.1	5
0.50	0	CS1	0	0
		CS2	78	4
		CS3	0	0
0.20	0.10	CS1	-9	3
		CS2	11	4
		CS3	-0.001	3
0.10	0.05	CS1	-4	2
		CS2	5	3
		CS3	-0.05	2

* relative error is defined as $r_{j,k} = \left(\hat{I}_{j,k} - \hat{I}_{j,T} \right) / \hat{I}_{j,T}$, where $j = 1, \dots, 1000$ and $k = 1, 2, 3$. Bias is defined as the expected relative error, $E[r_k]$, and is estimated by the sample mean $\bar{r}_k = \frac{1}{n} \sum_{j=1}^n r_{j,k}$.

** $SD(\text{relative error}) = \sqrt{\frac{1}{n-1} \sum_{j=1}^n \left(r_{j,k} - \frac{1}{n} \sum_{j=1}^n r_{j,k} \right)^2}$.