Estimating HIV Incidence Using Complex Survey Data

I. Flores Cervantes[1], J.D. Opsomer[1], G. Kalton[1]
R. Bain[2†], A. De[2†], P. Stupp[3]
[1]Westat, 1600 Research Blvd, Rockville, MD 20850
[2]U.S. Centers for Disease Control and Prevention, Atlanta GA 30333
[3]U.S. Centers for Disease Control and Prevention, retiree

Abstract
Accurate estimation of HIV incidence in at-risk countries is an important public health challenge. The Population-based HIV Impact Assessment (PHIA) Project has been conducting large-scale national population surveys using complex survey designs. HIV incidence estimation for these surveys is performed using a survey-weighted version of the incidence estimator originally proposed in Kassanjee, McWalter, Bärnighausen & Welte (2012). We describe a comprehensive variance estimation approach for this estimator that fully accounts for the variance components due to the survey and those due to the estimation of the biomarker assay parameters. The approach is readily integrated into a large-scale survey context, including that of the PHIA Project surveys. We illustrate the approach on data from three African countries and evaluate the sensitivity of the estimates to the values provided for the biomarker assay parameters and their measures of uncertainty.

Keywords: HIV incidence, variance estimation, replication, Taylor series linearization, HIV population-based survey.

Introduction

The incidence of human immunodeficiency virus (HIV) is the rate at which new HIV infections occur in susceptible populations. Accurate, practical, and cost-effective approaches for estimating incidence are a priority among researchers in HIV surveillance. The PHIA Project (Population-based HIV Impact Assessments; see http://phia.icap.columbia.edu/) is a multi-country initiative to measure the reach and impact of HIV programs in 13 African countries and Haiti. As a component of the PHIA Project, the incidence of HIV in participating countries is estimated based on data collected through cross-sectional general-population representative surveys and biomarker-based testing on blood samples of consenting participants.

Starting with Brookmeyer & Quinn (1995), HIV incidence in a population has been frequently estimated from cross-sectional data by testing individuals for biomarkers of recent infection. This avoids the substantial expense and time delay of a longitudinal survey, but it is subject to error because the biomarker tests are themselves imperfect and the recent infection status exhibits population-specific and subject-specific variability. Various incidence estimators have been proposed that account for these sources of error,

[†] Disclaimer: The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention/the Agency for Toxic Substance and Disease Registry.

often incorporating parameters that need to be specified in order to be able to apply the estimator. Kassanjee et al. (2012) proposed an incidence estimator that is based on a weighted mean recent incidence rate over a fixed time interval (often a year), as an approximation to the unobservable instantaneous incidence rate. This estimator has two external parameters, the "mean duration of recent infection" (MDRI) and the "false-recent rate" (FRR). The PHIA Project HIV incidence estimator is based on the Kassanjee et al. (2012) approach and will be more fully described in the next section.

Subsequent work has validated and further refined the biomarker-based approach to incidence estimation. Rehle et al. (2015) compared the biomarker-based approach of incidence estimation with two modeling-based alternatives on data from South Africa and found that the results are similar. Murphy et al. (2017) reviewed the current state of knowledge of HIV incidence testing methodologies and discuss future developments. A continuing challenge in using biomarker assays to determine whether an individual is recently infected is the inherent imprecision of the test. Kassanjee et al. (2014) evaluated the performance of five different incidence assays and stressed the need for further research to improve their accuracy. Duong et al. (2015) analyzed previously collected biospecimens to provide an estimate of the (MDRI) for the LAg-Avidity EIA assay, a key parameter in the Kassanjee et al. (2012) estimator adopted for the PHIA Project surveys. Recently, Kim and Rehle (2018) considered a combination of biomarker testing and information on antiretroviral therapy use to improve the determination of recent HIV infection status.

Many large-scale health surveys, including the PHIA Project surveys, use sampling designs with one or more stages of selection and unequal probabilities to select individuals whose HIV infection status will be assessed. The sampling design needs to be incorporated to obtain valid estimates of the population incidence rate. The standard approach to implement this is by using survey weights in the calculation of the estimates, and to account for the design features in variance estimation (see, for example, Lohr, 2009). The estimation and inference methods described in Kassanjee et al. (2012) cannot be directly applied in this setting, because their methods only apply when individuals are directly sampled from a population using simple random sampling. This has been recognized by the research community, and a survey-weighted version of the estimator is implemented in recent specialized software tools, e.g., the "inctools" package in R (Welte et al., 2018).

Unlike the incidence estimator itself, the variance estimation implementations to date do not account directly for the survey design. Instead, they rely on separately computed design effects to adjust the variance estimator of Kassanjee et al. (2012), which is both cumbersome and somewhat prone to error. In this paper, we describe a survey-based variance estimation approach that directly accounts for the design as well as the variability of the externally provided parameters. This facilitates implementation in a large-scale survey environment and provides access to the full complement of standard survey-based variance estimation methods.

As will be made clearer below, these variance estimators still need to account for the uncertainty of the externally provided parameters. Variance estimates for the latter are typically obtained from laboratory experiments done independently from the survey and not part of the PHIA Project study. It is of high practical interest, therefore, to determine the relative importance of the sampling variance and the externally obtained parameter variances and in particular, what the effect is of misspecification of the latter on overall measures of variability. We will evaluate this on data from recent PHIA Project surveys in three African countries.

The rest of the paper is structured as follows. In Section 2, we briefly review the incidence estimator of Kassanjee et al. (2012) and its survey-weighted generalization. In Section 3, we obtain variance estimators that fully account for the sampling design and the biomarker assay uncertainty. Section 4 describes an application of the variance estimation method to data from the PHIA Project and evaluate the sensitivity of the results to the externally provided biomarker assay parameter values and uncertainty measures.

Incidence estimation

Starting with the weighted incidence approximation proposed by Kassanjee et al. (2012), we define the population-level quantity

$$I_T = \frac{N_R - \beta_T N_+}{N_s\left(\Omega_T - \beta_T T\right)} \tag{1}$$

with N_S, N_R, and N_{NR} the number of uninfected, recently infected and non-recently infected individuals, N the total population (so that $N = N_S + N_R + N_{NR}$), and N_+ the number of infected individuals (so that $N_+ = N_R + N_{NR}$). The external parameters are Ω_T, the mean duration of recent infection (MDRI), defined as the average time alive and returning a recent result while infected for times less than T, and β_T, the false-recent rate (FRR) of the test defined as the proportion of recently infected individuals (as determined by the assay) among infected for times greater than T.

As noted, the quantity in (1) is an approximation of the true instantaneous incident in the population of interest, with the error of approximation derived in Kassanjee et al. (2012). The population-level approximation to the incidence is a valid survey estimation target, which is sometimes referred to as a "descriptive population quantity" following Pfeffermann (1993). In equation (1), T is a target duration of the period preceding the time of the survey, and in most cases the duration is $T = 365$ days. In addition to the instantaneous incidence, results are also often reported in terms of cumulative incidence after one year, usually referred to as *annual incidence*, defined as

$$I_a = 1 - \exp\left(-365\, I_T\right)$$

For now, we will focus the discussion on the estimation of I_T.

We are estimating the target quantity (1) based on data from a sample survey that follows a complex design. Let $U = \{1,\ldots,k,\ldots,N\}$ represent the finite population of interest, with k indexing the individuals in that population. A sample s is drawn from U according to a sampling design that can include stratification, clustering, and unequal probabilities of selection. Additionally, the sample can also have been affected by nonresponse. The final sample consists of the selected and responding individual, and each individual $k \in s$ is associated with a survey weight w_k, which is constructed so that it provides statistically valid estimation for population quantities. These weights reflect the sample design as well as post-sampling adjustments for nonresponse and calibration, as is customary in large-scale complex surveys. By construction, they ensure that weighted estimators are approximately unbiased and consistent for their target population quantities (see, for

example, Lohr, 2009). In particular, $\hat{N} = \sum_{k \in s} w_k$ is an estimate of the possibly unknown total population size.

The sample-based estimator of I_T in (1) is defined as

$$\hat{I}_T = \frac{\hat{N}_R - \hat{\beta}_T \left(\hat{N}_R + \hat{N}_{NR} \right)}{\left(\hat{N} - \hat{N}_R - \hat{N}_{NR} \right) \left(\hat{\Omega}_T - \hat{\beta}_T T \right)} \tag{2}$$

where $\hat{N} = \sum_{k \in s} w_k$ as above, $\hat{N}_R = \sum_{k \in s} w_k \delta_k \left(R \right)$ is the estimate of the total number of recent infected individuals in the population, and $\hat{N}_{NR} = \sum_{k \in s} w_k \delta_k \left(NR \right)$ is the estimate of the total number of non-recent infected persons in the population, where $\delta_k \left(C \right)$ is the indicator function defined as $\delta_k \left(C \right) = 1$ if the sampled case k has characteristic C and $\delta_k \left(C \right) = 0$ otherwise. We note that letting the weight $w_k = N / n$ for an equal-probability sample design (e.g., simple random sample design) in equation (2) produces the estimator of Kassanjee et al. (2012).

Before discussing the variance of the estimator $\hat{I}_T$ in (2) in the next section, we comment on the types of random variables in the estimator. The estimators $\hat{N}$, $\hat{N}_R$, and $\hat{N}_{NR}$ are random variables that depend on the sample design and are estimated using the survey data, as noted above. They are approximately unbiased and their variances denoted as $\mathrm{V}\left(\hat{N} \right)$, $\mathrm{V}\left(\hat{N}_R \right)$, and $\mathrm{V}\left(\hat{N}_{NR} \right)$, become smaller as the survey sample size increases. In the extreme case of a census (all cases are tested), the variance is zero.

In contrast to these sample-based estimates, the remaining two estimates $\hat{\Omega}_T$ and $\hat{\beta}_T$ are not computed from the survey data. Instead, they are provided by external studies such as the analysis described in Duong et al. (2015). Incorporating the variability of $\hat{\Omega}_T$ and $\hat{\beta}_T$ into an overall variance estimator for $\hat{I}_T$ requires externally provided variance estimates as well. Because they are not estimated as part of the survey, they are independent of the estimators $\hat{N}$, $\hat{N}_R$, and $\hat{N}_{NR}$. Additionally, because no covariance is provided for $\hat{\Omega}_T$ and $\hat{\beta}_T$, they are also treated as if they are independent of each other. In what follows, we will denote the externally provided variance estimates for $\hat{\Omega}_T$ and $\hat{\beta}_T$ as $\hat{\sigma}^2_{\hat{\Omega}_T}$ and $\hat{\sigma}^2_{\hat{\beta}_T}$, respectively. Note that because this variability is external to the survey, it does *not* become smaller as the survey sample increases.

Variance Estimation for the Incidence Estimator Under a Complex Design

Because the estimator of the incidence $\hat{I}_T$ in equation (2) is a nonlinear function of random variables, variance estimation targets an approximate variance instead of the exact one. The approximation is based on Taylor series linearization and is considered reliable for moderate to large samples, and is commonly used in survey practice (see, e.g., Lohr, 2009). To derive the approximate variance, let $\hat{\boldsymbol{\theta}}$ be a vector of five random variables

$\hat{\boldsymbol{\theta}} = \left(\hat{N}, \hat{N}_R, \hat{N}_{NR}, \hat{\beta}_T, \hat{\Omega}_T\right)' = \left(\hat{\mathbf{N}}', \hat{\beta}_T, \hat{\Omega}_T\right)'$ and let incidence be a nonlinear function of $\hat{\boldsymbol{\theta}}$, defined as $\hat{I}_T = \mathrm{g}(\hat{\boldsymbol{\theta}})$. Then the variance $\mathrm{V}\left(\hat{I}_T\right)$ is approximated by

$$\mathrm{V}(\hat{I}_T) = \mathrm{V}(g(\hat{\boldsymbol{\theta}})) \approx \nabla_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\, \mathrm{g}(\boldsymbol{\theta})' \, \mathbf{Cov}(\hat{\boldsymbol{\theta}}) \, \nabla_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\, \mathrm{g}(\boldsymbol{\theta}) \tag{3}$$

where $\nabla_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\, \mathrm{g}(\boldsymbol{\theta})$ is the gradient of $\mathrm{g}(\hat{\boldsymbol{\theta}})$ with respect to its component elements evaluated at $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$, and $\mathbf{Cov}(\hat{\boldsymbol{\theta}})$ is the variance-covariance matrix of $\hat{\boldsymbol{\theta}}$. Note that the covariance between $\hat{\mathbf{N}} = \left(\hat{N}, \hat{N}_R, \hat{N}_{NR}\right)$, $\hat{\Omega}_{\boldsymbol{R}}$ and $\hat{\beta}_T$ are zero by independence. As a result, the variance-covariance matrix $\mathbf{Cov}(\hat{\boldsymbol{\theta}})$ can be written as

$$\mathbf{Cov}(\hat{\boldsymbol{\theta}}) = \begin{pmatrix} \mathrm{V}\left(\hat{N}\right) & \mathrm{Cov}\left(\hat{N}, \hat{N}_{NR}\right) & \mathrm{Cov}\left(\hat{N}, \hat{N}_R\right) & 0 & 0 \\ \mathrm{Cov}\left(\hat{N}, \hat{N}_{NR}\right) & \mathrm{V}\left(\hat{N}_{NR}\right) & \mathrm{Cov}\left(\hat{N}_R, \hat{N}_{NR}\right) & 0 & 0 \\ \mathrm{Cov}\left(\hat{N}, \hat{N}_R\right) & \mathrm{Cov}\left(\hat{N}_R, \hat{N}_{NR}\right) & \mathrm{V}\left(\hat{N}_R\right) & 0 & 0 \\ 0 & 0 & 0 & \sigma^2_{\hat{\Omega}_T} & 0 \\ 0 & 0 & 0 & 0 & \sigma^2_{\hat{\beta}_T} \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{Cov}\left(\hat{\mathbf{N}}\right) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma^2_{\hat{\Omega}_T} & 0 \\ \mathbf{0} & 0 & \sigma^2_{\hat{\beta}_T} \end{pmatrix}$$

where $\mathbf{Cov}\left(\hat{\mathbf{N}}\right)$ is the variance-covariance matrix of $\hat{\mathbf{N}} = \left(\hat{N}, \hat{N}_R, \hat{N}_{NR}\right)'$ and $\sigma^2_{\hat{\Omega}_T}$, $\sigma^2_{\hat{\beta}_T}$ represent the variances of $\hat{\Omega}_T$ and $\hat{\beta}_T$, respectively. From this, we can write the approximate variance of $\hat{I}_T$ as the sum of three components

$$\mathrm{V}\left(\hat{I}_T\right) \approx V_1 + V_2 + V_3 \tag{4}$$

where V_1 is the variance due to estimating the totals N, N_R, and N_{NR} from the survey, V_2 is the variance due to estimating the MDRI Ω_T and V_3 is the variance due to estimating the FRR β_T. The explicit expressions for these components are based on the partial derivatives in $\nabla_{\boldsymbol{\theta}}\, \mathrm{g}(\boldsymbol{\theta})$ and are omitted here.

Following standard survey estimation practice again, a variance estimator for $\hat{I}_T$ is obtained by replacing all unknown quantities in (4) by corresponding "plug-in" estimators. Unlike in the standard case, some of these plug-in estimators are provided by external sources. Finally, we obtain the variance estimator $\hat{\mathrm{V}}\left(\hat{I}_T\right)$,

$$\hat{\mathrm{V}}\left(\hat{I}_T\right) = \hat{V}_1 + \hat{V}_2 + \hat{V}_3, \tag{5}$$

where

$$\begin{aligned}\hat{V}_1 = \hat{K}_1 \Bigg(& \left(\left(1-\hat{\beta}_T\right)\hat{N}_R - \hat{\beta}_T\hat{N}_{NR}\right)^2 \hat{\mathrm{V}}\left(\hat{N}\right) \\ & + \left(\left(1-\hat{\beta}_T\right)\hat{N} - \hat{N}_{NR}\right)^2 \hat{\mathrm{V}}\left(\hat{N}_R\right) + \left(\hat{\beta}_T\hat{N} - \hat{N}_R\right)^2 \hat{\mathrm{V}}\left(\hat{N}_{NR}\right) \\ & + 2\left(\left(1-\hat{\beta}_T\right)\hat{N}_R - \hat{\beta}_T\hat{N}_{NR}\right)\left(\left(1-\hat{\beta}_T\right)\hat{N} - \hat{N}_{NR}\right)\widehat{\mathrm{Cov}}\left(\hat{N}, \hat{N}_R\right), \\ & - 2\left(\left(1-\hat{\beta}_T\right)\hat{N}_R - \hat{\beta}_T\hat{N}_{NR}\right)\left(\hat{\beta}_T\hat{N} - \hat{N}_R\right)\widehat{\mathrm{Cov}}\left(\hat{N}, \hat{N}_{NR}\right) \\ & - 2\left(\left(1-\hat{\beta}_T\right)\hat{N} - \hat{N}_{NR}\right)\left(\hat{\beta}_T\hat{N} - \hat{N}_R\right)\widehat{\mathrm{Cov}}\left(\hat{N}_R, \hat{N}_{NR}\right)\Bigg)\end{aligned} \tag{6}$$

$$\hat{V}_2 = \hat{K}_2\left(\left(T-\hat{\Omega}_T\right)\hat{N}_R - \hat{\Omega}_T\hat{N}_{NR}\right)^2 \hat{\sigma}^2_{\hat{\beta}_T}, \text{ and} \tag{7}$$

$$\hat{V}_3 = \hat{K}_2\left(\left(1-\hat{\beta}_T\right)\hat{N}_R - \hat{\beta}_T\hat{N}_{NR}\right)^2 \hat{\sigma}^2_{\hat{\Omega}_T}, \tag{8}$$

with $\hat{K}_1 = \left(\hat{\Omega}_T - \hat{\beta}_T T\right)^{-2}\left(\hat{N} - \hat{N}_R - \hat{N}_{NR}\right)^{-4}$ and $\hat{K}_2 = \left(\hat{\Omega}_T - \hat{\beta}_T T\right)^{-4}\left(\hat{N} - \hat{N}_R - \hat{N}_{NR}\right)^{-2}$. Equations (6) to (8) are applicable to any sample design and can be implemented using standard survey statistical software. The term $\hat{V}_1$ reflects the variability with respect to the sampling design, while the other two terms account for the variability of the biomarker test parameters.

As an alternative to the above Taylor series linearization "plug-in" approach to variance estimation in surveys, it is often more convenient to use a replication variance estimation method. While they are not equivalent, both approaches are statistically valid and result in estimates that are very similar in practice. For an overview of replication variance estimation methods in surveys, see Wolter (2007). Commonly used replication methods in large-scale surveys are the jackknife (see, for example, Kott, 2001) and balanced repeated replication (BRR; see Rao and Shao, 1999). Replication methods can correctly account for the stratification, clustering, and sample weighting, including nonresponse and poststratification weighting adjustments.

The PHIA Project surveys implement a paired unit jackknife variance estimation method called JK2, in which jackknife replicates are formed by randomly deleting a PSU from each paired PSUs in the sample (Rust & Rao, 1996). The advantages of the JK2 replication method are the reduced computational effort compared to other methods while taking into account the precision benefits of implicit stratification of PSUs, as sampled PSUs are paired off in the systematic order they were selected within sampling stratum.

To estimate the variance of the incidence, the JK2 replication method needs to be modified to account for the variance components due to the externally provided quantities. Hence, the JK2 variance estimator replaces the linearization-based $\hat{V}_1$ in (5). For the remaining

two variance components, the explicit expressions $\hat{V}_2$ and $\hat{V}_3$ in (7) and (8) continue to be used.

We now return to the estimation of annual incidence $I_a = 1 - \exp(-365 I_T)$. Using the same Taylor series linearization approach as for the variance of $\hat{I}_T$, $\mathrm{V}\left(\hat{I}_a\right)$ is approximated by

$$\begin{aligned}\mathrm{V}\left(\hat{I}_a\right) &\approx 365^2 e^{-730 I_T}\, \mathrm{V}\left(\hat{I}_T\right) \\ &= 365^2 e^{-730 I_T} \left(V_1 + V_2 + V_3\right). \\ &= V_1^* + V_2^* + V_2^*\end{aligned}$$

The approximate variance of the annual incidence $\mathrm{V}\left(\hat{I}_a\right)$ is the variance of the instantaneous incidence $\mathrm{V}\left(\hat{I}_T\right)$ in (4), multiplied by the factor $365^2 \exp\left(-730 I_T\right)$. As before, V_1^* can be estimated using the plug-in approach as in (6) or a replication method using the full sample and replicate weights. The components of variance $\hat{V}_2^*$ and $\hat{V}_3^*$ are computed by plugging the estimates $\hat{N}$, $\hat{N}_R$, $\hat{N}_{NR}$, $\hat{\beta}_T$, $\hat{\Omega}_T$ $\hat{V}\left(\hat{\beta}_T\right)$ and $\hat{V}\left(\hat{\Omega}_T\right)$ in equations (7) and (8) and multiplying by $365^2 \exp\left(-730 \hat{I}_T\right)$. The estimate of variance for the annual incidence is then computed as the sum of the estimates of variance from the three sources as

$$\hat{\mathrm{V}}\left(\hat{I}_a\right) = \hat{V}_1^* + \hat{V}_2^* + \hat{V}_3^*$$

When $\exp\left(-730 \hat{I}_T\right) \approx 1$, which occurs for several PHIA countries, it is also possible to further simplify these expressions and use

$$\hat{\mathrm{V}}\left(\hat{I}_a\right) = 365^2\, \hat{\mathrm{V}}\left(\hat{I}_T\right).$$

Numerical Example of Estimates of Annual Incidence

When survey results are released, the validity of the results follows from the fact that the sample selection and data collection follow standardized and fully documented survey methodologies. In doing so, the organization responsible for the survey is able to ensure low bias in the estimates and to provide reliable measures of precision in the form of confidence intervals or coefficients of variation (CV). An unusual aspect of PHIA Project surveys is that both the incidence estimates and the associated measures of precision depend on estimated parameters that are external to the survey. Hence, the survey organization is not able to fully justify the quality of their results based only on their internal procedures and instead relies on external studies, over which it has no control. Therefore, it is useful to assess the sensitivity of published estimates to these external parameters. We will do this in this section, using data from three African countries in which the PHIA Project conducted surveys recently. For confidentiality reasons, they will be referred to as Countries 1, 2 and 3.

Table 1 below shows the estimated annual incidence percentages for individuals 15-49 years old, with their estimated standard errors computed using the approach described in the previous section. The following external parameter estimates were used to obtain these results: point estimates $\hat{\beta}_T = 0$ and $\hat{\Omega}_T = 130$ with variance estimates $\hat{\sigma}^2_{\hat{\beta}_T} = 0$ and $\hat{\sigma}^2_{\hat{\Omega}_T} = 37.5$ respectively. These values are standard across the PHIA Project studies, and the value of $\hat{\Omega}_T$ was also recommended by Duong et al. (2015). Note that these values imply that there is no false-recent error in the biomarker assay.

Table 1: Estimates of Annual Incidence and Standard Errors for Individuals Aged 15-49 by Sex for 3 Countries

Country	*Group*	*Incidence (%)*	*St. error (%)*
1	Female	1.10	0.18
	Male	0.31	0.11
	Total	0.70	0.10
2	Female	0.69	0.16
	Male	0.30	0.12
	Total	0.50	0.10
3	Female	0.46	0.14
	Male	0.26	0.09
	Total	0.36	0.08

Table 2 shows the relative importance of the three estimated variance components of the annual incidence, $\hat{V}_1^*$, $\hat{V}_2^*$, and $\hat{V}_3^*$ as a fraction of the total estimated variance. These results show that $\hat{V}_1^*$ dominates the estimated variance, with $\hat{V}_3^*$ contributing a modest fraction and no contribution from $\hat{V}_2^*$. These result imply that, at the current settings for the externally provided parameters, the estimated standard errors in Table 1 depend almost exclusively on the variance of the survey estimates.

Table 2: Relative Contributions of Estimated Variance Components for Annual Incidence Estimates for Individuals Aged 15-49 by Sex for 3 Countries

Country	*Group*	*Survey (%)*	*FRR (%)*	*MDRI (%)*
1	Female	92.4	0.0	7.6
	Male	98.2	0.0	1.8
	Total	90.7	0.0	9.3
2	Female	95.9	0.0	4.1
	Male	98.6	0.0	1.4
	Total	94.6	0.0	5.4
3	Female	97.8	0.0	2.2
	Male	98.1	0.0	1.9
	Total	96.1	0.0	3.9

In order to evaluate the dependence of the results to the external parameters, we performed a sensitivity analysis. Considering the estimates themselves first, the value of $\hat{\beta}_T$ was varied between 10^{-6} and 10^{-1} with equal increments on the logarithmic scale, while that of $\hat{\Omega}_T$ ranged from 100 to 250 with increments of 30. The resulting estimates of annual

incidence for Country 1 are shown in Table 3, with the estimates for the current values of the external parameters highlighted in bold in the table (note: the original value of $\hat{\beta}_T$ was 0 but leads to virtually identical estimates). These results show that, as long as the assumption of negligible FRR is reasonable, the annual incidence is quite insensitive to the exact value chosen. However, once the FRR becomes 1% or higher, the estimated incidence increases very rapidly. In contrast, the effect of the MDRI is more gradual, with changes in duration in either direction leading to substantial changes in incidence estimates. The same sensitivity analysis was performed for Countries 2 and 3, with similar results not reported here.

Table 3: Estimated Annual Incidence (in %) for Different Values of False-Recent Rate and MDRI, for Individuals Aged 15-49 in Country 1

	MDRI					
FRR	*100*	*130*	*160*	*190*	*220*	*250*
0.000001	0.91	**0.70**	0.57	0.48	0.42	0.37
0.00001	0.91	0.70	0.57	0.48	0.42	0.37
0.0001	0.92	0.71	0.57	0.48	0.42	0.37
0.001	0.96	0.74	0.60	0.51	0.44	0.38
0.01	1.42	1.09	0.88	0.74	0.64	0.56
0.1	8.43	5.80	4.42	3.66	3.00	2.58

We also evaluated the sensitivity of the variance estimates to the externally provided variances. Table 4 below shows the standard errors of the estimated annual incidence for individuals 15-49 years old in Country 1, for a range of values for the variance parameters. The point estimates were set at $\hat{\beta}_T = 10^{-9}$ and $\hat{\Omega}_T = 130$. The value for the FRR was chosen to be negligible, as for the value used in the current incidence estimates, but larger than zero so that a variance greater than zero would be sensible. For the variances, we considered $\hat{\sigma}^2_{\hat{\beta}_T} = 9.5\times10^{-7},\ 3.8\times10^{-6}$, and 9.5×10^{-6} $\hat{\sigma}^2_{\hat{\Omega}_T} = 37.5,\ 150,\ 375$. The lowest value for the variance of the MDRI is the same as currently used for the current incidence variance estimates, and the other two were chosen arbitrarily to represent 4 and 10 times higher variances. We discuss the choice of values for the variance of the FRR further below.

The results in Table 4 show that the standard errors of the annual incidence estimates are moderately sensitive to the values of the external variance parameters. In the most extreme case considered, if these external variances underestimate the uncertainty of the FFR or the MDRI by a factor of 10 (corresponding to the last row and column of Table 4, respectively), the reported standard errors can be too small by 35% or more.

Table 4: Standard Errors (in %) of the Estimated Annual Incidence for Individuals Aged 15-49 in Country 1, for Different Values of the External Variance Parameters

	MDRI		
FRR	*37.5*	*150*	*375*
9.5×10^{-7}	0.11	0.12	0.15
3.8×10^{-6}	0.13	0.14	0.16
9.5×10^{-6}	0.15	0.16	0.18

As in Table 2 above, Table 5 shows the relative contributions of the variance components $\hat{V}_2^*$ and $\hat{V}_3^*$ to the total estimated variance of the annual incidence. The variance values of the FFR were chosen so that the variance components $\hat{V}_2^*$ and $\hat{V}_3^*$ were the same size for each of the increments in the respective variance values (see the diagonal values in Table 5). Note that, although the variance values $\hat{\sigma}^2_{\hat{\beta}_T}$ are very small in absolute terms, they correspond to much larger coefficients of variation for the FFR than do the variance values for the MDRI. This would indicate that the values chosen for $\hat{\sigma}^2_{\hat{\beta}_T}$ in this evaluation are conservative and that the contribution of $\hat{V}_2^*$ can be expected to be smaller than that of $\hat{V}_3^*$ if the coefficients of variation of the FRR and the MDRI are similar. The results in Table 5 confirm that for the lower variance values considered, the effect of the non-survey variance components on the overall variance is modest. This changes markedly as the external variance values increase so that the overall estimated standard error of the annual incidence becomes highly dependent on the external parameters.

Table 5: Relative Contributions (in %) of Estimated Variance Components for Annual Incidence Estimates for Individuals Aged 15-49 in Country 1, for Different Values Of The External Variance Parameters. Left-hand Numbers in Each Entry Are FRR Contributions, Right-hand Numbers Are MDRI Contributions

	MDRI		
FRR	*37.5*	*150*	*375*
9.5×10^{-7}	8.5 / 8.5	6.8 / 27.0	4.8 / 48.1
3.8×10^{-6}	27.0 / 6.8	22.5 / 22.5	16.8 / 42.0
9.5×10^{-6}	48.1 / 4.8	42.0 / 16.8	33.6 / 33.6

Conclusions

In this paper, we have described the variance estimation approach for a commonly used HIV incidence estimator in large-scale surveys. The approach accounts for the variance contributions due to the survey estimates and those due to the biomarker-based parameters. We show how the variance estimator is derived based on standard survey statistics results, and how to implement it within a survey estimation environment. While we explored the specific case of the incidence estimator of Kassanjee et al. (2012) and the PHIA Project surveys, the general approach will apply to other estimators that contain both survey and non-survey estimated components.

We evaluated the behavior of the estimation approach on data from PHIA Project surveys in three African countries. Overall, it appears that the HIV incidence estimates are sensitive to the values provided for the FRR and the MDRI, with the effect of the latter more likely to be important in the range of values similar to the current estimate. Hence, it appears important that these external parameters be estimated accurately, to avoid bias in the annual incidence estimates.

At least in the countries considered, the standard errors of the annual incidence are dominated by the uncertainty of the survey-estimated inputs to the incidence formula, given the variance values currently provided to quantify the uncertainty of the FRR and the MDRI. This is a useful result because it indicates that focusing on conducting high-quality surveys is the best way to maintain (or improve) the precision of the HIV incidence

estimates. However, this is no longer the case if the variances of the FRR or the MDRI are severely underestimated, in which case the standard errors of the annual incidences become highly dependent on these external values rather than on the survey estimates.

References

Welte A., Grebe E., McIntosh A., Bäumler P., and Ongarello S. (2018). inctools: Incidence Estimation Tools. R package version 1.0.11. https://CRAN.R-project.org/package=inctools.

Brookmeyer, R., and Quinn, T.C. (1995). "Estimation of current human immunodeficiency virus incidence rates from a cross-sectional survey using early diagnostic tests." *Am J Epidemiol;* 141:166–172.

Kassanjee R., McWalter, T.A., Bärnighausen, T., and Welte A. (2012). "A new general biomarker-based incidence estimator." *Epidemiology*; 23:721-728.

Kim A. A., and Rehle T. (2018). "Recent Infection Testing Algorithm That Includes Viral Load Testing and Exposure to Antiretroviral Therapy." *AIDS Res Hum Retroviruses*, Published Online:24 Jul 2018.

Lohr, S. (2009). *Sampling: Design and Analysis*. Nelson Education.

Pfeffermann, D. (1993). "The Role of Sampling Weights When Modeling Survey Data." *International Statistical Review*. 61(2): 317-337.

Rehle T., Johnson L., Hallett T., Mahy M., Kim A., Odido H., Onoya D., Jooste S., Shisana O., Puren A., Parekh B., and Stover J. (2015). "A comparison of South African national HIV incidence estimates: A critical appraisal of different methods." *PLoS One* ;10:e0133255.

Reshma K., Pilcher, C. D., Keating, S. M, et al. (2014). "Independent Assessment Of Candidate HIV Incidence Assays on Specimens in the CEPHIA Repository." *AIDS*, 28(16): 2439-2449.

Rust, K., and Rao, J.N.K. (1996). "Variance Estimation for Complex Surveys Using Replication Techniques." *Statistical Methods in Medical Research*, 5, 283-310.

Wolter, K.M. (2007). Introduction to Variance Estimation (2nd Edition). Springer.

Yen T. Duong, Reshma Kassanjee, Alex Welte, Meade Morgan, et al. (2015). "Recalibration of the Limiting Antigen Avidity EIA to Determine Mean Duration of Recent Infection in Divergent HIV-1 Subtypes." *PLoS One* 10:e0114947.