2011

Statistical Considerations in Determining HIV Incidence from Changes in HIV Prevalence

Ron Brookmeyer, University of California, Los Angeles
Jacob Konikoff

Available at: http://works.bepress.com/rbrookmeyer/34/
Statistical Considerations in Determining HIV Incidence from Changes in HIV Prevalence

Ron Brookmeyer, University of California, Los Angeles
Jacob Konikoff, University of California, Los Angeles
Statistical Considerations in Determining HIV Incidence from Changes in HIV Prevalence

Ron Brookmeyer and Jacob Konikoff

Abstract

The development of methods for estimating HIV incidence is critical for tracking the epidemic and for designing, targeting and evaluating HIV prevention efforts. One method for estimating incidence is based on changes in HIV prevalence. That method is attracting increased attention because national population-based HIV prevalence surveys, such as Demographic and Health Surveys, are being conducted throughout the world. Here, we consider some statistical issues associated with estimating HIV incidence from two population-based HIV prevalence surveys conducted at two different points in time. We show that the incidence estimator depends on the relative survival rate. We evaluate the sensitivity of estimates to incorrect assumptions about the relative survival rate, and show that small errors in the relative survival can, in some situations, create large biases in HIV incidence. We determine sample sizes of prevalence surveys to estimate incidence with precision and show how the sample sizes depend on baseline prevalence, the relative survival rate, and the population HIV incidence rate. We find that even if the relative survival rate were known exactly, there are situations where prohibitively large prevalence surveys would be required to produce reliable incidence estimates. These situations can occur either when the baseline prevalence is large, the relative survival rate is near 1, or the population incidence is small. Because information on the relative survival rate may be limited or not specific to the population under study, we suggest an approach to empirically estimate this critical parameter by augmenting population-based prevalence surveys with a mortality follow-up sub-study. We determine sample sizes of the prevalence surveys and mortality sub-studies for this augmented design and provide the necessary R code (version 2.13.0) for sample size determinations. We conclude that caution should be exercised when solely relying on changes in prevalence as the method for determining HIV incidence because of the method's sensitivity to mortality assumptions and the very large sample size requirements in some settings.

KEYWORDS: HIV epidemiology, incidence, prevalence, sample size
1 Introduction

HIV incidence is the rate that new infections occur in populations. Knowledge of HIV incidence rates is critical for tracking the course of the epidemic, and for designing, targeting and evaluating HIV prevention efforts. HIV prevalence is the proportion of living persons who are HIV infected. While HIV prevalence measures overall disease burden, the HIV incidence rate tracks the leading edge of the epidemic. Reliable HIV incidence estimates are also required to properly determine sample sizes of HIV prevention trials.

Determination of HIV incidence is a more challenging undertaking than HIV prevalence because of a number of practical and methodological issues. The strengths and weakness of various approaches for HIV incidence estimation have been reviewed [1]. One approach is based on the longitudinal follow-up of uninfected cohorts. High follow-up rates and large sample sizes of representative cohorts of uninfected persons are required for accuracy. Furthermore, the process of actively following cohorts, which should include counseling to reduce high-risk behaviors, could change the population HIV incidence rate, the very quantity we are trying to measure. A second approach is based on surveys which utilize biomarkers to identify infections that occurred recently [2]. The biomarker approach circumvents some of the difficulties that occur with the cohort approach because it requires only a cross-sectional sample and not longitudinal follow-up; however, the biomarker approach has been hampered by the accuracy of currently available biomarker assays for distinguishing recent from chronic long-standing infections [3, 4]. A third approach for incidence estimation utilizes changes in HIV prevalence rates as determined by prevalence surveys conducted at two or more points in time. This approach is based on a demographic balancing equation that involves key assumptions about HIV mortality and migration. The objective of this paper is to study some of the statistical issues underlying this last approach. We explore biases from incorrect assumptions, required sample sizes to achieve a specified precision and a suggestion for a study design to strengthen the approach by lessening its sensitivity to assumptions.

Nationally representative HIV prevalence surveys, such as the Demographic Health and Interview Surveys containing an HIV/AIDS component, have now been conducted in over 30 countries throughout the world including countries in Africa, Asia, Latin America and Eastern Europe. Such surveys were nearly nonexistent prior to 2000. UNAIDS now relies on these surveys to calculate national prevalence statistics for many countries. Population-based HIV prevalence surveys are planned to be repeated in a number of countries. Furthermore, seroprevalence surveys in select at-risk populations are now routinely being conducted over time in many settings. The increasing availability of data from repeated HIV prevalence surveys
over time motivates the need for statistical study of the problem of inference about the HIV incidence rate from changes in HIV prevalence.

Several statistical approaches have been discussed in the literature for estimating HIV incidence rates from HIV prevalence. Some approaches use age-specific HIV prevalence rates obtained from a survey conducted at only one point in calendar time. Podgor and Leske [5] developed an approach for estimating incidence from age-specific prevalence rates that accounted for differential mortality which assumed the underlying incidence is constant. Using this approach, Saidel et al. [6] estimated age specific HIV incidence in Burundi. The methods of Gregson, Donnelly, Parker and Anderson [7] apply to stable endemic situations where the level and age pattern of HIV prevalence rates are constant. Williams, et al. [8] extended the methods to allow for changing HIV incidence rates and differential mortality, but the generalizations come at the price of additional modeling assumptions. Ades [9] has cautioned that distinguishing changes in incidence from background differential changes in the inclusion in samples may be difficult, limiting the value of sentinel surveys for such purposes. Marschner [10] has discussed sample size considerations for estimating age-specific incidence from a single cross-sectional sample assuming non-differential mortality [11]. Recently, Hallett et al. [12] developed methods for estimating HIV incidence from two representative cross-sectional prevalence surveys based on demographic accounting and evaluated the approach with empirical data from southern and eastern Africa. This method depends on assumptions about the cohort mortality rate or survival function following HIV infection. Hallet et al. [12] conclude that cross-sectional prevalence surveys can provide reliable estimates of HIV incidence. However the method relies on having accurate information about local mortality rates. The difficulty in disentangling HIV incidence from changes in HIV prevalence is highlighted by observations which reveal that in some countries HIV prevalence rises despite stable incidence [13, 14], while in other countries an alarmingly high rate of incident infections continue to occur despite stable HIV prevalence [15].

The objective of this paper is to consider some statistical issues and errors that occur in estimating HIV incidence rates from two serial HIV prevalence surveys. We show that the estimate of HIV incidence depends on assumptions about the relative survival of HIV infected persons relative to the general population, although not necessarily on the absolute survival rates. We derive expressions for the required sample sizes of prevalence surveys to estimate HIV incidence with a desired degree of precision. We consider biases in HIV incidence rates induced by incorrect assumptions about the relative survival rates. We propose an approach for estimating the critical input parameter, the relative survival rate, by embedding a follow-up mortality sub-study within a national HIV prevalence survey. We determine sample sizes with this augmented design.
2 Estimation of HIV Incidence From HIV Prevalence Surveys

Suppose two independent HIV prevalence surveys are conducted at times $t_1$ and $t_2$ in a population. The two prevalence surveys are conducted $\delta = t_2 - t_1$ years apart. We define $p_i = 1 - q_i$ to be the HIV prevalence rate in the population at calendar time $t_i$. That is, $p_i$ is the conditional probability that a person who is alive at time $t_i$ is living with HIV infection. The HIV incidence rate, $I$, is the hazard rate of becoming HIV infected in the time interval between the two surveys which we assume is a constant over that interval. We believe this assumption is reasonable for the situation we have in mind where the time interval between surveys, $\delta$, is short, on the order of a couple years. In this section, we consider the situation where only persons who were eligible for sampling at time $t_1$ are eligible for sampling at time $t_2$, as long as they have not died prior to time $t_2$. In section 4.2 we consider the additional complications introduced by migration into or out of the population between times $t_1$ and $t_2$.

The probability that a person who is alive at calendar time $t_1$ is living with HIV infection at time $t_2$ can be expressed as the sum of the probabilities of two events, $A$ and $B$. The event $A$ is the event that an individual who was alive at time $t_1$ was living with HIV infection at that time and then survives to time $t_2$. The probability of event $A$, $P(A)$, is $p_1 S_I$ where $p_1$ is the HIV prevalence rate at calendar time $t_1$ and $S_I$ is the probability that an individual who is living with HIV infection at $t_1$ survives to calendar time $t_2$. Event $B$ is the event that an individual who is alive at time $t_1$ is uninfected at that time, subsequently becomes infected in the interval $(t_1, t_2)$, and is still alive as of time $t_2$. The probability of event $B$, $P(B)$, is approximately equal to $I \delta q_1 S_U$ where $q_1 = 1 - p_1$, and $S_U$ is the probability that a person who is uninfected at time $t_1$ survives to calendar time $t_2$. This expression for $P(B)$ was derived by approximating the probability that a person becomes infected in the interval $(t_1, t_2)$ by $I \delta$, and by approximating the probability that such a person survives to time $t_2$ by $S_U$.

The HIV prevalence rate, $p_2$, can then be expressed as $p_2 = [P(A) + P(B)]/S$ where $S$ is the survival probability that an individual alive at calendar time $t_1$ survives at least to time $t_2$; $S$ is a weighted average of the survival probabilities of HIV infected and uninfected persons, that is, $S = p_1 S_I + (1 - p_1) S_U$. The relative survival rate is defined as $R = S_I / S$ which is the ratio of the survival probability that an HIV infected person who is alive at time $t_1$ survives to $t_2$ divided by the corresponding probability for the entire population, which is a mixture of infected and uninfected persons. Then, we have

$$p_2 \approx \frac{p_1 S_I}{S} + \frac{I \delta q_1 S_U}{S} \approx p_1 R + I \delta q_1,$$
where the last approximation follows from the approximation $S_U \approx S$. Thus we have, $I = (p_2 - p_1R)/q_1\delta$. Therefore the incidence rate depends approximately on the absolute survival probabilities only through the relative survival rate, $R$. Furthermore, the equation shows that $p_2$ is constrained to be $\geq p_1R$ because the only way persons may exit the population is through death. If $\hat{p}_i = 1 - \hat{q}_i$ are the maximum likelihood estimators of the HIV prevalence rates, then the maximum likelihood estimator of the incidence rate is

$$\hat{I} = (\hat{p}_2 - \hat{p}_1R)/\hat{q}_1\delta$$  \hspace{1cm} (1)

if $\hat{p}_2 \geq \hat{p}_1R$, and we set $\hat{I} = 0$ if $\hat{p}_2 < \hat{p}_1R$.

### 3 Considerations When the Relative Survival Rate is Known

In this section, we consider sample size requirements necessary to estimate the HIV incidence rate, $I$, with a desired degree of precision. We assume in this section that the relative survival rate, $R$, is known perfectly. In section 4 we consider the case when $R$ is not known. Suppose the variance (var) of $\hat{p}_i$ is such that $\text{var}(\hat{p}_i) = p_iq_iD_i/N_i$ where $D_i$ is the design effect to account for sampling designs more complex than simple random sampling. Note that $D_i = 1$ corresponds to binomial variation with simple random sampling. We used the delta method to find the asymptotic standard error (se) of $\hat{I}$ from equation (1). We then derived an expression for the coefficient of variation $c = \text{se}(\hat{I})/I$ and obtained,

$$c = (p_2 - p_1R)^{-1} \sqrt{\left(p_1(p_2 - R)^2D_1(N_1q_1)^{-1} + p_2q_2D_2N_2^{-1}\right)}.$$  \hspace{1cm} (2)

We note from equation (2) that the coefficient of variation does not depend on $\delta$, but the standard error of $\hat{I}$ does. We used equation (2) to derive a formula for sample sizes required to achieve a desired degree of precision of the incidence rate. When $N_1 = N_2 = N$ the sample size required to achieve a coefficient of variation equal to $c$ is

$$N = \frac{p_1(p_2 - R)^2D_1 + p_2q_2q_1D_2}{c^2(p_2 - p_1R)^2q_1}.$$  \hspace{1cm} (3)

Figure 1 shows the required sample sizes, $N$, (at each time point) to achieve a coefficient of variation of 0.20 when $R = 0.80$ (with $D_1 = D_2 = 1$). If $I = 0.02$ per year, the figure illustrates that the required sample size increases as $p_1$ increases. This is intuitively explained by the fact that a higher baseline prevalence at time $t_1$ should make it more difficult to distinguish incident infections from long-term prevalent infections at time $t_2$. The figure also illustrates that the sample size increases as $I$ decreases.
Figure 1: Sample sizes, $N = N_1 = N_2$, required to achieve a coefficient of variation of 0.20 of the HIV incidence rate with $R = 0.80$ plotted versus baseline HIV prevalence, $p_1$, for different values of the true incidence rate $I$ (with the design effects $D_1 = D_2 = 1$).

Figure 2 plots the required sample sizes, $N$, needed to achieve a coefficient of variation of 0.20 when $R = 0.50$.

Comparison of Figures 1 and Figure 2 illustrates that the sample sizes are dependent on the relative mortality $R$. The closer $R$ is to 1 the more difficult it is to distinguish a new incident infection from a long-standing prevalent infection. For example, if $I = 0.02$ per year, and $p_1 = 0.10$, the required sample sizes for $R = 0.80$ and 0.50 are 11046 and 6490, respectively. The cautionary warning suggested by these figures is that under certain conditions the sample sizes that are required to
precisely estimate the incidence rate from changes in prevalence are prohibitively large. For example, if the baseline prevalence, $p_1$, at $t_1$ is 0.20 and the HIV incidence rate is 0.01 per year, the sample sizes required to achieve a coefficient of variation of 0.20 are $N = 93606$ and $N = 52638$ when $R = 0.80$ and 0.50 respectively. An R (version 2.13.0) function is provided in our supplementary material to calculate the sample sizes that produced Figures 1 and Figure 2.

4 Issues When the Relative Survival Rate Is Unknown

4.1 Impact of Errors in the Relative Survival Rate

In this section we consider the complications that arise when the relative survival rate, $R$, is either not known or an incorrect value is assumed. First, suppose the true value for the relative survival rate in the population is $R_0$ but an incorrect value $R_1$ is assumed instead, so that the bias in the relative survival is $B_R = R_1 - R_0$. Then, the estimator $\hat{I}$ in equation (1) is not converging to the true incidence $I_0$ but is instead converging to $I_1 = (p_2 - p_1 R_1)/(q_1 \delta)$. The asymptotic bias of the incidence estimator is $B_I = I_1 - I_0 = -(p_1 B_R)/(q_1 \delta)$. Figure 3 plots the bias $B_I$ as a function of baseline prevalence $p_1$ for different values of $B_R$. If the relative survival rate is overestimated ($B_R > 0$), the incidence rate will be underestimated ($B_I < 0$), and if the relative survival is underestimated, the incidence will be overestimated.

![Figure 3: Asymptotic bias, $B_I$, of the estimated incidence rate induced by incorrect assumptions about the relative survival rate as a function of baseline prevalence, $p_1$, for different values of bias in the relative survival rate, $B_R$.](image)

The figure shows that the bias $B_I$ increases with increasing baseline prevalence, $p_1$, for a fixed value of $B_R$. Indeed, if the baseline prevalence is high, small
errors in the relative survival rate can be propagated into very large errors in incidence. For example, if \( p_1 = 0.20 \), and the true relative mortality is 0.90 but is incorrectly assumed to be 0.80 (i.e., \( B_R = -0.10 \)), the incidence would be overestimated by 0.025 per year; that is, if the true incidence rate was 1% per year, the estimated incidence would be (asymptotically) estimating 3.5% per year, and would thus be off by over a factor of 3.

### 4.2 A Proposed Study Design if the Relative Survival Rate is Unknown

The relative survival rate depends on local epidemic conditions including the stage of the epidemic (e.g. the proportions of infected persons in advanced stages of HIV disease), the proportion of individuals on antiretrovirals, the age distribution, and background local mortality rates. In general, the relative survival rate is not fixed, but changes over calendar time for reasons that include changes in the proportions on antiretrovirals or with advanced HIV disease. In addition, \( R \) also depends on the duration of the time interval \( \delta \) between the two prevalence surveys. Accordingly, it is important that the assumed value of \( R \) refers specifically to the calendar period between the two prevalence surveys and to the population that is being sampled. Here we propose an approach for obtaining just such a value of \( R \). A disadvantage of follow-up studies for HIV incidence determination, where counseling during follow-up visits may reduce high-risk behaviors, is that this can potentially affect that which is being measured, HIV incidence. That complication is not an issue for the type of follow-up study proposed here where mortality is the endpoint, as only vital status need be ascertained at the time of the second prevalence survey.

Our proposal is to embed a mortality follow-up sub-study within the prevalence surveys. The sub-study involves choosing a random sample of \( m_I \) infected persons from among the infected persons who are identified at the first prevalence survey at time \( t_1 \). In addition, a random sample is drawn of \( m_U \) uninfected individuals from among the uninfected persons identified at the first prevalence survey. The persons in the sub-study are followed to determine their vital status at time \( t_2 \).

Suppose \( a_I \) persons among the \( m_I \) infected persons are still alive at time \( t_2 \), and \( a_U \) persons among the \( m_U \) uninfected persons are still alive at time \( t_2 \). Then, an estimate of the relative survival rate from the sub-study is

\[
\hat{R} = \frac{a_I}{\hat{p}_1 \left( \frac{a_I}{m_I} \right) + \hat{q}_1 \left( \frac{a_U}{m_U} \right)}.
\]

The incidence estimator using the relative survival estimate (4) is

\[
\hat{I} = (\hat{p}_2 - \hat{p}_1 \hat{R})/\hat{q}_1 \delta.
\]
The estimated survival proportions \( \frac{a_I}{m_I} \) and \( \frac{a_U}{m_U} \) are assumed independent with binomial variances \( S_I (1 - S_I)/m_I \) and \( S_U (1 - S_U)/m_U \), respectively. More generally, a complex sampling scheme could be used to choose the members of the sub-study. In that case, design effects \( d_I \) and \( d_U \) could be introduced into the binomial variances. In what follows we consider the simpler case of simple random sampling in which those design effects (\( d_i \)) are 1. We used the multivariate delta method to determine the standard error of the incidence estimate that accounts for uncertainty in the relative survival rate estimated from equation (4). We then calculated the standard errors and coefficient of variation of the incidence estimates as a function of the sample sizes of the prevalence surveys (\( N_1 = N_2 = N \)) and the sample sizes of the sub-study (\( m_I = m_U = m \)).

Figure 4 plots the sample sizes required to achieve a coefficient of variation of 0.20 when \( I = 0.05 \) per year and \( R = 0.80 \) (with \( S_U = 0.95 \)), versus baseline prevalence, \( p_1 \), and displays the graphs for different sample sizes, \( m \), of the sub-study. For example, if \( p_1 = 0.10 \), a sample size of \( N = 2732 \) is required with a sub-study of size \( m = 100 \). For \( m = 500, 1000 \) and \( 2000 \), the required sample sizes are \( N = 2118, N = 2060 \) and \( N = 2032 \), respectively. The figure shows that the required sample size, \( N \), increases as baseline prevalence, \( p_1 \), increases. The required sample size also increases as the sample size of the sub-study decreases. If the sample size of the sub-study is only \( m = 100 \), we find that if the baseline prevalence \( p_1 > 0.20 \), then it is not possible to achieve a coefficient of variation of 0.20 or less, no matter how large we choose \( N \). Figure 4 also shows that the sample size, \( N \), corresponding to sub-studies of size \( m = 1000 \) and \( m = 2000 \) are quite similar. Thus, there is a point of diminishing returns as the size of the sub-study increases.
Figure 5 is a graph of the sample sizes needed to achieve a coefficient of variation of 0.50 when $I = 0.01$ per year and $R = 0.80$ (with $S_U = 0.95$). If the baseline prevalence, $p_1$, is 0.10, for a sub-study size of $m = 500$, 1000 and 2000 the required sample sizes are $N = 8770$, $N = 7726$ and $N = 7292$, respectively. The figure illustrates that if $m$ is only 100, a coefficient of variation of 0.50 cannot be achieved regardless how large one chooses $N$. That is, prevalence surveys, even if made exceedingly large, cannot compensate for very limited information on mortality. An R (version 2.13.0) function is given in the supplementary material to calculate the sample sizes, with a mortality follow-up sub-study, needed to estimate incidence precisely.

An important complication may arise from migration. If HIV infected persons are differentially moving in or out of the survey catchment area, the incidence estimators (equation 1 and equation 5) will be biased. Here, we briefly outline an approach to account for migration effects when there is a mortality follow-up sub-study. The sampling frame for the second HIV prevalence survey at time $t_2$ should be restricted to those persons eligible for inclusion in the survey at time $t_1$. This can be accomplished by asking the following screening question to each person who is sampled at time $t_2$: “Were you living in the catchment at time $t_1$?” If the person was not living in the catchment area at time $t_1$, then the person would be deemed ineligible for inclusion in the prevalence sample at $t_2$ and another person would be chosen. To account for persons emigrating out of the catchment area between $t_1$ and $t_2$, the endpoint in the follow-up sub-study should be broadened to include not just death but either death or emigration outside of the catchment area.

Published by De Gruyter, 2011
5 Discussion

The objective of this paper was to consider statistical issues associated with estimating HIV incidence from two population-based HIV prevalence surveys. We found that the incidence estimator depends critically on the relative survival rate associated with HIV. Even if there are only small errors in the assumed value for $R$, we find that such small errors can cause large biases in HIV incidence, and this is particularly true when the baseline prevalence, $p_1$, is large. Furthermore, we find that even if the relative survival rate was known exactly, there are situations when prohibitively large sample sizes would be required to produce precise incidence estimates. These situations may occur when either the baseline prevalence, $p_1$, is large, the relative survival, $R$, is near 1, or the population incidence, $I$, is very small. Such situations might correspond to countries with high background HIV prevalence, with high use of effective antiretrovirals and where current levels of HIV incidence have declined. The cautionary warning is that in some settings, even if the relative survival, $R$, could be assumed to be known perfectly (and of course in practice it can never be assumed to be known perfectly), the method of estimating HIV incidence from changes in prevalence may have limited practical utility because of statistical instability. It is possible that additional information about incidence from changes in prevalence may be gleaned by accounting for age. For example, at young ages, near the time of initiation of sexual activity, changes in HIV prevalence may more accurately reflect incident HIV infection in populations where the main risk of infection is through sexual transmission [16].

To account for uncertainty in the relative survival, $R$, we have suggested augmenting national HIV prevalence surveys with a mortality follow-up sub-study. Such a sub-study could provide critical information about the relative survival at little marginal cost. Nevertheless, the cautionary warning that was discussed above still applies, that is, in some settings prohibitively large sample sizes for the prevalence surveys would be required (e.g., when $p_1$ is large, $R$ near 1, and $I$ is small).

An alternative approach to estimating HIV incidence is based on biomarkers that distinguish recent from long-standing infection. However, the biomarkers that are currently available have limitations and active research is ongoing to improve the biomarker approach [3, 17, 18]. Thus, at this point in time we do not have a single widely accepted method for estimating population HIV incidence levels that is accurate, cost-effective, practical and easily implementable. As there are important sources of error with each of the existing approaches to estimating incidence, corroborating incidence estimates using multiple methods is prudent.
References


Published by De Gruyter, 2011


