Published in final edited form as:

Biometrics. 2015 December ; 71(4): 1121–1129. doi:10.1111/biom.12336.

Sample Size Methods for Estimating HIV Incidence from Cross-Sectional Surveys

Jacob Konikoff and Ron Brookmeyer

Department of Biostatistics, Fielding School of Public Health, University of California, Los Angeles, Los Angeles, California 90095-1772 U.S.A

Jacob Konikoff: konikoff@ucla.edu

Summary

Understanding HIV incidence, the rate at which new infections occur in populations, is critical for tracking and surveillance of the epidemic. In this paper we derive methods for determining sample sizes for cross-sectional surveys to estimate incidence with sufficient precision. We further show how to specify sample sizes for two successive cross-sectional surveys to detect changes in incidence with adequate power. In these surveys biomarkers such as CD4 cell count, viral load, and recently developed serological assays are used to determine which individuals are in an early disease stage of infection. The total number of individuals in this stage, divided by the number of people who are uninfected, is used to approximate the incidence rate. Our methods account for uncertainty in the durations of time spent in the biomarker defined early disease stage. We find that failure to account for this uncertainty when designing surveys can lead to imprecise estimates of incidence and underpowered studies. We evaluated our sample size methods in simulations and found that they performed well in a variety of underlying epidemics. Code for implementing our methods in R is available with this paper at the Biometrics website on Wiley Online Library.

Keywords

cross-sectional; HIV; incidence; sample size; trends

1 Introduction

Of the over 30 million people living with HIV in the world, approximately two million were infected this past year (Joint United Nations Programme on HIV/AIDS, 2014). There exists considerable uncertainty in these estimates. Accurate estimation of the rate at which new infections occur is crucial for tracking and surveillance of the epidemic. Knowledge of this hazard rate, known as the incidence, is also critical for effectively designing, targeting, and evaluating prevention efforts.

Correspondence to: Jacob Konikoff, konikoff@ucla.edu.

Supplementary Materials: Code for implementing the methodology described in Section 4 and Section 5 is available with this paper at the *Biometrics* website on Wiley Online Library. Additionally a table of selected results from the methodology in Section 4 may be found on the website.

Cohort studies are traditionally used to estimate hazard rates. HIV incidence has been measured through longitudinal follow up of cohorts in various subpopulations (Karon et al., 2001). In these studies ethical considerations may necessitate counseling against risky sexual behavior. This counseling can change the incidence – the very quantity the study is designed to estimate. Additional complications arise from selection bias and loss to follow up (Brookmeyer, 2010). These challenges, combined with cost, have led public health agencies to look for alternative methodologies for estimating HIV incidence.

Cross-sectional surveys offer an increasingly promising alternative for estimating HIV incidence (Laeyendecker et al., 2013; Mastro, 2013). The cross-sectional surveys utilize biomarkers, such as HIV viral load and CD4 cell count, to define and mark people in an early disease stage. The total number of individuals found in the early disease stage, divided by the number of people who are uninfected, can be used to approximate the rate at which new infections occur in the susceptible part of the population. Previous work has already examined how best to define the early disease stage using biomarkers (Laeyendecker et al., 2013; Brookmeyer et al., 2013; Konikoff et al., 2013). This paper focuses on the practical question of how researchers are to determine the sample sizes needed for conducting these cross-sectional surveys. We develop sample size methods for a single survey to estimate HIV incidence at a point in time, as well as sample size methods for surveys conducted at two points in time to detect changes in HIV incidence. This methodology is a novel extension of earlier work regarding the properties of the cross-sectional incidence estimator (Brookmeyer, 1997; Brookmeyer and Quinn, 1995; Kaplan and Brookmeyer, 1999). In Section 2, we discuss the statistical framework and notation. Some key distributional results are established in Section 3. Section 4 develops sample size methods for estimating incidence from a single survey. Section 5 discusses sample size methods for assessing changes in HIV incidence from two surveys. In Section 6 we evaluate the performance of these sample size methods by simulating cross-sectional surveys in a variety of epidemics. Further extensions are considered in Section 7 including applications beyond the context of HIV incidence estimation. The methods and results are discussed in Section 8.

2 Framework and Notation

Suppose we have a random sample of n persons from a target population. Blood or other biological specimens are collected from each person. Those specimens are analyzed for various biomarkers in order to classify each person into one of three groups. A standard HIV antibody test divides the n collected samples into N_u samples from uninfected individuals and N_i samples from infected individuals. The infected samples are then further divided into X samples from people in the early disease stage and $N_i - X$ samples from individuals who have progressed out of this early disease stage. While n is fixed before the survey, N_u , N_i , and X are all random variables whose realizations we denote by n_u , n_i , and x .

A biomarker-based testing algorithm is used to separate the X samples in the early disease stage from the rest of the infected population. The testing algorithm classifies each person based on the results of assays on the person's blood or biological specimens. Figure 1 illustrates one such biomarker-based testing algorithm which uses four component assays and was designed for use in Clade B epidemics (Konikoff et al., 2013). In this paper we will

use this specific biomarker-based testing algorithm, and the early disease stage it defines, but the methodology would be applicable to other biomarker-based testing algorithms and associated definitions of the early disease stage.

In order to measure incidence from the results of the cross-sectional survey we define the “cross-sectional incidence”, I_c , by borrowing a well-known concept from epidemiology. In a steady-state the prevalence of a condition is equal to the incidence of the condition multiplied by the average duration of the condition (Freeman and Hutchison 1980). In our situation the epidemiological relationship is equivalent to saying that the number of individuals in the early disease stage is equal to the number of new individuals in this stage multiplied by the average amount of time people spend in this stage. This leads to the

definition $I_c = \frac{1}{1-p} \frac{\pi}{\mu}$ (Brookmeyer and Quinn, 1995; Kaplan and Brookmeyer, 1999) where p is the proportion of the population infected with HIV (the HIV prevalence), π is the probability of being in the early disease stage at the time of the cross-sectional survey, and μ

is the average amount of time spent in the early disease stage. We then estimate $I_c = \frac{1}{1-p} \frac{\pi}{\mu}$

from the survey as $\hat{I}_c = \frac{x}{\mu n_u}$ since the sample estimates are

$$\frac{1}{1-\hat{p}} = \frac{n}{n_u} \quad \text{and} \quad \hat{\pi} = \frac{x}{n}.$$

However, unless the historical incidence remains unchanged so that we are in a steady-state, I_c is not a measure of the current instantaneous hazard rate but rather an approximate measure of incidence at a time point ψ years before the survey was conducted (Kaplan and Brookmeyer, 1999). This follows from the fact that historical infection rates, not the instantaneous infection rate, determine the composition of the infected and uninfected samples in the survey (see appendix for additional detail). In previous work we estimated $\phi(t)$, the probability that persons infected t years ago will be in the early disease stage at the time of the survey (Konikoff et al., 2013). This function can be used to estimate ψ (Brookmeyer and Quinn, 1995). For our example biomarker-based testing algorithm $\psi \approx 0.48$ years.

In what follows we will assume that a preliminary estimate of the HIV prevalence, p_0 , is available immediately before the cross-sectional survey. Estimating HIV prevalence is a simpler undertaking than estimating HIV incidence (Brookmeyer, 2010). We will structure the problem so that researchers can solve for $u = n(1-p)$ instead of directly solving for the sample size n . We call u the required number of uninfected samples since $n(1-p)$ will approximately equal n_u for any given sample. Researchers may solve for u and wait to solve for n until p_0 becomes available. This is particularly helpful when setting the power for consecutive surveys as it allows researchers to wait until immediately before the second survey to supply an estimate of the HIV prevalence at the second time point.

3 Probabilistic and Distributional Considerations

If we conceptualize the survey as splitting the n samples into those which are, and are not, in the early disease stage, then $X \sim \text{binomial}(n, \pi)$ where, as above, π is the probability of

being in the early disease stage at the time of the cross-sectional survey. Since $I_c = \frac{1}{1-p} \frac{\pi}{\mu}$ we have that $\pi = (1-p)\mu I_c$. Therefore $X \sim \text{binomial}(n, (1-p)\mu I_c)$. The “success” probability will be small since the vast majority of people in a population will be uninfected and most of the infected individuals will not appear in the early disease stage. We therefore apply a Poisson approximation to the binomial distribution and let $X \sim \text{Poisson}(n(1-p)\mu I_c)$ so that $X \sim \text{Poisson}(u\mu I_c)$.

In our estimate $\hat{I}_c = \frac{x}{n_u \mu}$ we are implicitly assuming that μ is known. However, μ is never known exactly and must be estimated for each specific testing algorithm. We can incorporate our uncertainty by placing a distribution, called $h(\mu)$ on μ . One suggestion is to

let $h(\mu) = \frac{\mu^{\gamma-1} \exp^{-\frac{\mu}{\beta}}}{\Gamma(\gamma)\beta^\gamma}$ which is a Gamma distribution with parameters γ and β (Brookmeyer 1997). Under this assumption we have $P(X=x) = \frac{\Gamma(\gamma+x)}{\Gamma(\gamma)x!} \left(\frac{\beta u I_c}{\beta u I_c + 1}\right)^x \left(\frac{1}{\beta u I_c + 1}\right)^\gamma$. This is the result of the fact that a Gamma-Poisson mixture follows a negative binomial distribution (Agresti, 2013). The Gamma distribution thus simplifies the remaining calculations but the methodology can be generalized to any distribution on μ . In this case the maximum

likelihood estimate for I_c is given by $\frac{x}{u\gamma\beta}$. Noting that $E(\mu) = \gamma\beta$ in the Gamma distribution

above and that $u \approx n_u$, we see the consistency with the standard estimate $\hat{I}_c = \frac{x}{n_u \mu}$. We will shortly examine different values of γ and β .

4 Sample Size Methods for a Single Cross-Sectional Survey

We now derive the sample size needed to achieve an estimate of incidence with a desired amount of precision. The calculations will depend on a preliminary estimate of incidence, I_0 , because the underlying incidence dictates the number of samples which will be classified in the early disease stage. The smaller the underlying incidence the larger the sample size we will need to capture a sufficient number of people in the early disease stage. To make this explicit in our calculations we will focus on controlling the width of the confidence interval divided by the true underlying incidence.

We will use the exact $100(1-\alpha)\%$ confidence interval for the “success” probability of a negative binomial experiment (Casella and Berger, 2002). In the negative binomial

distribution above the “success” probability is $\frac{1}{\beta u I_c + 1}$ and the corresponding $100(1-\alpha)\%$ confidence interval is given by $\{\text{Beta}(\alpha/2, \gamma, x+1), \text{Beta}(1-\alpha/2, \gamma, x)\}$, where $\text{Beta}(q, a, b)$ is the q^{th} quantile of a Beta distribution with parameters a and b . We may transform the

endpoints of the confidence interval for $\frac{1}{\beta u I_c + 1}$ to get that

$\left[\left\{ \frac{1}{\text{Beta}(1-\alpha/2, \gamma, x)} - 1 \right\} / \beta u, \left\{ \frac{1}{\text{Beta}(\alpha/2, \gamma, x+1)} - 1 \right\} / \beta u \right]$ is a $100(1-\alpha)\%$ confidence interval for I_c . Since x is unknown before sampling we replace x with $E[X] \approx$

$uI_0\gamma\beta$. Then to achieve a $100(1 - \alpha)\%$ confidence interval of width W we solve for u in the equation $W = \left[\left\{ \frac{1}{\text{Beta}(\alpha/2, \gamma, uI_0\gamma\beta+1)} - 1 \right\} - \left\{ \frac{1}{\text{Beta}(1 - \alpha/2, \gamma, uI_0\gamma\beta)} - 1 \right\} \right] / \beta u$.

We will assume that for any given algorithm γ and β are known constants estimated from previous work. For the biomarker-based testing algorithm described above we estimated that $\gamma = 120.77$ and $\beta = 0.003$ by fitting a Gamma distribution to a sampling distribution of μ generated by bootstrapping (Konikoff et al., 2013). Code for determining the required sample size for any W and I_c is available in the Supplementary Material. Alternatively, given a preliminary estimate of incidence, I_0 , and a fixed sample size the anticipated confidence interval can be determined. For any given I_0 a practitioner can set the precision, determine the corresponding sample size, and then solve for the anticipated confidence interval implied by the original choice of W . Repeating this process may provide intuition to a researcher who is comfortable specifying I_0 but unsure how to specify the precision and would rather control the margin of error for the asymmetric confidence interval.

The higher the desired precision the larger the required number of uninfected samples we will need. For example, if the underlying incidence is 0.25% per year 2,060 uninfected samples are needed to achieve $W/I_c = 3.5$, 20,768 uninfected samples are needed to achieve $W/I_c = 1$, and 141,895 uninfected samples are needed to achieve $W/I_c = 0.5$. Since it is much easier to measure larger incidences only 7,095 uninfected samples are needed to achieve $W/I_c = 0.5$ if the underlying incidence were 5% per year instead of 0.25% per year. In fact, we may note that in the equation

$W/I_0 = \left[\left\{ \frac{1}{\text{Beta}(\alpha/2, \gamma, uI_0\gamma\beta+1)} - 1 \right\} - \left\{ \frac{1}{\text{Beta}(1 - \alpha/2, \gamma, uI_0\gamma\beta)} - 1 \right\} \right] / \beta u I_0$ everywhere the sample size u appears it is multiplied by the incidence I_0 . Thus the sample size is inversely proportional to the underlying incidence and, for example, if incidence is halved the sample size will need to be doubled. It is clear that the number of samples needed can become prohibitive when a high degree of precision is required and the incidence is low. In the Supplementary Material we present a table of the required sample sizes to achieve a set W/I_c for a 95% confidence intervals for selected values of W and I_c .

Since this methodology is applicable to other biomarker-based testing algorithms we examine three other hypothetical situations illustrated in Panel A of Figure 2. In the first situation (black line) we ignore uncertainty in μ , in the second (green curve) we increase the variability in μ , and in the third we change the Gamma distribution so that μ is centered at a larger mean (blue curve). In the first two alternative scenarios μ is centered at the same mean as our example biomarker-based testing algorithm. In the third scenario we extend the amount of time, on average, people spend in the early disease stage. Since the Gamma distribution is determined by its first two moments we can examine how differences in the distribution of μ affect the required sample sizes by changing either the mean or the variance and leaving the other fixed.

Panel B of Figure 2 plots W/I_c for 95% confidence intervals against the required number of uninfected samples for the scenarios illustrated in Panel A. The underlying incidence in Panel B was set at 1% per year which might occur in a high risk subpopulation of the United

States. All else equal, smaller sample sizes are needed when individuals spend more time, on average, in the early disease stage since it will be easier to sample individuals in this state. This rule can be seen when comparing the blue and red curves. When we instead fix the mean of the Gamma distributions we can see how additional samples are needed to account for the increased uncertainty by comparing the black, red, and green curves. The difference between the black curve and the red curve shows the extent of under sampling that would occur if we ignored our uncertainty in μ while using our example biomarker-based testing algorithm. While these curves are relatively close we may notice how they separate as the desired precision increases. In general differences in the distributions are magnified when the situation demands larger sample sizes such as when higher precision is required or when the underlying incidence is small. When we compare the green curve, with additional variability, to the red and black curves we see that no sample size can make up for too much uncertainty in μ . It is clear that a $W/I_c < 1.5$, something achievable with under 5,000 uninfected samples in the other cases, is prohibitive in this case. This emphasizes the importance of accurately estimating μ and highlights the significance of properly accounting for our uncertainty.

5 Sample Size Methods for Detecting Changes in Incidence

We now move from a single cross-sectional survey to two consecutive cross-sectional surveys where the aim is to determine if incidence has changed. The surveys are conducted at calendar times t_1 and t_2 and estimate incidences I_1 and I_2 . We drop the subscript c for convenience. We will concern ourselves with testing the null hypothesis of no difference in incidence against the alternative hypothesis that incidence has increased by some amount. More formally, we wish to test $H_0: I_2/I_1 = 1$ against an alternative that $H_A: I_2/I_1 = r > 1$. For any specific alternative, r , we will derive the needed sample sizes, n_1 and n_2 , to have a high power of detecting this change if it occurred. As above, the sample sizes will be based on the number of uninfected samples, u_1 and u_2 , needed at each survey but the sample sizes, n_1 and n_2 , should be fixed by using preliminary estimates of prevalence at the two time points. While the calculations below are for a one-sided test, the power calculation for a 2-sided test can be conducted by halving the α -level, as under a given alternative hypothesis the probability of rejecting the null hypothesis in the opposite direction will be negligible. If we wish to look for a decrease in incidence the roles of I_1 and I_2 can be interchanged.

In order to calculate the sample size needed for each of the two cross-sectional surveys we extend previous work which derived sample size formulas when X follows a Poisson distribution with known mean (Gail, 1974; Brown and Green 1982). We will do this by

rewriting $P(\text{Reject } H_0 | H_A)$ as $\sum_{t=0}^{\infty} P(\text{Reject } H_0 | H_A, T=t) \times P(T=t | H_A)$ where T is the total number of individuals found in the early disease stage from both surveys.

Specifically, as in the one sample case, we assume that $X_l | \mu \sim \text{Poisson}(u_l \mu I_l)$ for $l = 1, 2$. Then since the surveys are independent we have that $T = X_1 + X_2 \sim \text{Poisson}[(u_1 I_1 + u_2 I_2) \mu]$ conditionally on knowledge of μ . This implies that if we further condition on the observed

$T = t$, we get $X_2 | \mu, t \sim \text{Binomial}\left(t, \frac{u_2 I_2 \mu}{(u_1 I_1 + u_2 I_2) \mu}\right)$. We may notice that μ drops out when

we condition on t because it affects the two samples equally. Thus,

$$X_2|t \sim \text{Binomial}\left(t, \frac{u_2 I_2}{u_1 I_1 + u_2 I_2}\right).$$

We may now calculate the power for any value $T = t$ given a specific alternative hypothesis r . We do not require the sample sizes at the two time points to be equal. However, one must set the ratio of the number of uninfected samples required at the two time periods, $s = u_2/u_1$, before carrying out the power calculation. Note that under the null hypothesis we have

$$X_2|t \sim \text{Binomial}\left(t, \frac{u_2}{u_1 + u_2}\right) \text{ or } X_2|t \sim \text{Binomial}\left(t, \frac{s}{1+s}\right) \text{ and under the alternative}$$

hypothesis $X_2|t \sim \text{Binomial}\left(t, \frac{rs}{1+rs}\right)$. The rejection rule under the null hypothesis is to reject H_0 if x_2 is greater than or equal to the smallest cutoff, C , such that

$$\sum_{k=C}^t \binom{t}{k} \left(\frac{s}{1+s}\right)^k \left(\frac{1}{1+s}\right)^{t-k} \leq \alpha, \text{ where } \alpha \text{ is the desired size of the one sided test. This cutoff, } C, \text{ can be found in statistical packages such as R. Next we calculate}$$

$$\sum_{k=C}^t \binom{t}{k} \left(\frac{rs}{1+rs}\right)^k \left(\frac{1}{1+rs}\right)^{t-k} \text{ which is } P(\text{Reject } H_0 | H_A, T=t), \text{ the probability of rejecting the null hypothesis when the alternative } r \text{ is true for a fixed } t.$$

However before the sample is conducted, T is not fixed but is random. Therefore we must also calculate $P(T=t|H_A)$. Our uncertainty in μ implies that T is a Gamma-Poisson mixture which is, as above, negative binomial. To see this, notice that T follows a Poisson with mean $u_1(I_1 + sI_2)\mu \sim \text{Gamma}(\gamma, \beta u_1(I_1 + sI_2))$. Therefore we have that

$$P(T=t|H_A) = \frac{\Gamma(\gamma+t)}{\Gamma(\gamma)t!} \left(\frac{\beta u_1 I_1 (1+rs)}{\beta u_1 I_1 (1+rs)+1}\right)^t \left(\frac{1}{\beta u_1 I_1 (1+rs)+1}\right)^\gamma.$$

$$\text{Then } P(\text{Reject } H_0 | H_A) = \sum_{k=0}^{\infty} P(\text{Reject } H_0 | H_A, T=t) \times P(T=t | H_A) \text{ which we may write as}$$

$$\sum_{t=0}^{\infty} \left[\left\{ \sum_{k=C}^t \binom{t}{k} \left(\frac{rs}{1+rs}\right)^k \left(\frac{1}{1+rs}\right)^{t-k} \right\} \frac{\Gamma(\gamma+t)}{\Gamma(\gamma)t!} \left(\frac{\beta u_1 I_1 (1+rs)}{\beta u_1 I_1 (1+rs)+1}\right)^t \left(\frac{1}{\beta u_1 I_1 (1+rs)+1}\right)^\gamma \right]$$

We see that power is a function of the alternative r , the desired ratio of the two samples sizes s , and $u_1 I_1$ (or alternatively $u_2 I_2$). Thus for any specific α level, and desired power, we may solve for $u_1 I_1$ for given r and s . Table 1 shows the necessary $u_1 I_1$ and in parenthesis ($u_1 + u_2$) I_1 to achieve 90% power for various values of r and s . For a fixed value of s the rows correspond to the black, red, green, and blue distributions of μ in Panel A of Figure 2.

This table shows the required number of uninfected samples in the first survey multiplied by the incidence at time t_1 as well as the total uninfected samples needed from both surveys multiplied by the incidence at time t_1 . For example, if a researcher using our example biomarker-based testing algorithm wants to have 90% power of detection for a doubling of the incidence at time t_2 compared to the incidence at time t_1 (taking $s = 1$ so that $u_1 = u_2$) the researcher would need to sample so that $u_1 I_1 = 68.1$. If the incidence at time t_1 is 1% per year

this would imply that $u_1 = u_2 = 6,810$ uninfected individuals are required at both time points. These numbers can then be adjusted to fix the total sample sizes n_1 and n_2 based on preliminary estimates of prevalence. Code for determining the sample sizes in other situations is available in the Supplementary Material.

When we compare the rows marked i and ii, we see that as in the one sample case, accounting for the uncertainty in μ when using our example biomarker-based testing algorithm increases the required sample sizes. The effect is most pronounced when the underlying incidence is low. Since Table 1 is explicitly in terms of the incidence this point is more directly understood than in the previous section. If we look at the row marked iii, where μ has a higher variance, we see that much larger sample sizes will be required. We can also see that the smaller the change in incidence we wish to detect (the closer r is to 1) the larger the sample size we will need. Table 1 combines these two effects to illustrate the tremendous cost associated with measuring small changes in incidence when the underlying incidence is low. A further similarity to the single survey case can be seen by looking at the rows marked iv. We see that larger values of μ again imply smaller sample sizes.

6 Simulation

In order to test our methodology in a variety of underlying epidemics we created a simulation. The initial phase of each epidemic is the same for the first 730 days. We let $50e^{(d/300)}$ infections occur each day for $1 \leq d \leq 730$ days. The number of infections is rounded to the nearest whole number whenever necessary. At $d = 730$ we let there be 1,000,000 uninfected individuals in the population. From here infections occur according to four unique epidemic curves shown in Figure 3. We examined an exponentially decreasing epidemic (Epidemic A), a sinuous epidemic where incidence both decreases and increases (Epidemic B), a linearly increasing epidemic (Epidemic C), and an exponentially increasing epidemic (Epidemic D). Epidemic B could represent a population in which HIV first spreads in a subpopulation where it remains isolated until it breaks into a secondary part of the population. If we let $\tau = 0$ represent the beginning of these unique epidemic curves we examined changes in incidence from $\tau = 6 - \psi$ years to $\tau = 10 - \psi$ years and from $\tau = 8 - \psi$ years to $\tau = 10 - \psi$ years. Vertical grey lines denote these time points in Figure 3. The infections before $\tau = 0$ do not affect the number of individuals in the early disease stage at the three time periods we sample the population – at years 6, 8, and 10. In previous work we estimated that the probability an individual is found in the early disease converges to 0 by 5 years after infection (Konikoff et al., 2013). That is we estimated $\phi(t) = 0$ for $t > 5$. The initial phase of the epidemic is relevant to the prevalence of the disease at the sample dates.

Figure 3 was created by plotting the function $g(10 - \tau)$ against τ for $0 \leq \tau \leq 10$, where $g(t)$ gives the expected number of new infections per year, divided by the size of the population, t years before the cross-sectional survey occurs. The function is defined relative to the sample date in keeping with previous work on cross-sectional HIV incidence surveys (Kaplan and Brookmeyer, 1999). However, because it is easier to conceptualize time moving forward, Figure 3 plots $g(10 - \tau)$ after we have followed each of the epidemics for 10 years.

An overview of how the simulations were conducted can be seen in Figure 4. Births occur at a daily rate of 0.004 percent of the uninfected part of the population. Infections occur according to the curves in Figure 3. Upon infection each individual is assigned a death date (specifically caused by the virus) based on a draw from a Weibull distribution with shape parameter 2.5 and scale parameter 14. This sets the median survival at approximately 12 years past infection. All individuals are assigned a death date (independent of infection) based on a random draw from an exponential distribution with rate 0.00002 so that 0.002 percent of the population would die of other causes each day. Infected individuals thus receive two death dates upon infection and the earlier date is used.

We used the simulation to measure the power when we compare incidences at $6 - \psi$ years and $10 - \psi$ years and when we compare incidences at $8 - \psi$ years and $10 - \psi$ years. We fixed the alternative hypothesis r based on the true incidences ψ years before the sample dates and we let $s = 1$ for simplicity. The expected power for these studies was 0.90. We used the true underlying incidence at $6 - \psi$ years (for the first comparison) and the true underlying incidence at $8 - \psi$ years (for the second comparison) to determine the needed uninfected sample sizes based on the formula in Section 5. These numbers are shown in Table 2. We fixed the total number of samples at each time point based on the true prevalence at the time of the survey.

When we sampled the population we separated the uninfected individuals from the infected individuals. The infected individuals were further “tested” with our biomarker-based testing algorithm by taking a binomial draw where the probability of being in the early disease stage was determined by the estimated $\varphi(t)$ function (Konikoff et al., 2013) adjusted by the probability of surviving to the time of the survey once infected with the virus. We then calculated our estimates of the incidences at the two time points as well as whether or not we would have rejected the null hypothesis. We repeated this procedure 1,000 times. To account for uncertainty in μ we bootstrapped the sample we used to estimate the $\varphi(t)$ function 999 times so that each time the true value of μ changed (see appendix). We calculated the power as the number of times we rejected H_0 over the 999,000 total runs. The calculated power for all four epidemics is displayed in Table 2. In each comparison we also calculated the anticipated 95% confidence intervals for both incidences being measured based on the formula in Section 4.

If we look at the first two columns of Table 2 we see that even after 999,000 iterations the average of the estimated incidences do not perfectly match the true underlying incidences. We would expect this based on two assumptions the cross-sectional methodology makes. These assumptions are explained in the appendix. The more important of these assumptions is that $g(t)$ must be approximately linear in the few years preceding sampling. If $g(t)$ is convex then we will tend to overestimate the true incidence and if $g(t)$ is concave we will tend to underestimate the true incidence. The expected 90% power is achieved in the linear, exponential decay, and exponential growth epidemics where the bias in the incidence estimates at the first and second time points is in the same direction. In the epidemic where incidence both decreases and increases the true quantities we are comparing are closer together than the true incidences would lead us to believe. If we look at Table 1 we can see that if r is truly closer to 1 than was assumed we may wind up drastically under sampling.

This can happen even with a small change. Still in the two scenarios given the power is greater than 80% which may be acceptable.

An additional complication at play is that ψ is also defined by $\varphi(t)$ (see appendix). Therefore changing $\varphi(t)$ in each bootstrap changes ψ slightly and thus each bootstrap would need its own sample size calculation. We fix the sample sizes across all the bootstraps to mirror reality where we must estimate ψ as $\hat{\psi}$. Since this curve is never known exactly we are forced to base our alternative hypothesis on the incidence $\hat{\psi}$ years before the survey.

Although the sample sizes were chosen for the power calculations we can still use the simulation to check the validity of the methods developed in Section 4. Using the determined samples sizes and the underlying incidences we calculated the percentage of the 999,000 estimated incidences which were covered by the anticipated 95% confidence interval used to define W . We see that the expected 95% coverage is achieved in most of the scenarios presented in the third column of Table 2. The confidence intervals seem to be anti-conservative when the underlying incidence is 0.3% per year. This appears to be a consequence of the fact that the number of individuals marked in the early disease stage must be a whole number so that the distribution of the simulated incidence values increases

approximately in steps between $0, \frac{1}{\mu u}, \frac{2}{\mu u}, \frac{3}{\mu u}$, etc. There is some variability around each step as our actual estimates of incidence use n_u , the number of uninfected individuals sampled, rather than u , which is the expected value of N_u . When the underlying incidence is particularly close to 0 the mass is centered around a small number of these jumps and it becomes hard to get proper coverage. For example, for Epidemic D in Table 2 when the true incidence is 0.21%, 1.6% of the simulated values are exactly 0, none lie between 0 and 0.0515%, and 6.5% are between 0.0515%, and 0.0520%. Thus even a small shift in the bound of the confidence interval will exclude all 6.5% instead of including this mass.

7 Further Extensions

We note that this methodology could be used to compare the incidence of two different subpopulations if the underlying distribution for μ is the same in both groups. This may be the case in two different states with similar demographics.

If the distribution of the average duration spent in the early disease stage is not the same in the two groups then we can generalize the above methodology. This could be the case if we were comparing incidence in two different countries where there are differences in the specific clade of virus infecting people and in the use of antiretrovirals. We will need to specify the joint distribution, $h(\mu_1, \mu_2)$. We assume independence so that $h(\mu_1, \mu_2) = e(\mu_1)f(\mu_2)$. We can then write

$$P(\text{Reject } H_0 | H_A) = \int_0^\infty \int_0^\infty \sum_{t=0}^\infty (\text{Rej. } H_0 | H_A, T=t, \mu_1, \mu_2) P(T=t | H_A, \mu_1, \mu_2) e(\mu_1) f(\mu_2) d\mu_1 d\mu_2.$$

To simplify the calculations we will let $s = 1$ so that $u_1 = u_2$. Then under H_0 we have

$$X_2 | t, \mu_1, \mu_2 \sim \text{Binomial} \left(t, \frac{\mu_2}{\mu_1 + \mu_2} \right) \text{ and under } H_A \text{ we get}$$

$X_2|t, \mu_1, \mu_2 \sim \text{Binomial}\left(t, \frac{\mu_2 r}{\mu_1 + \mu_2 r}\right)$. Therefore the power conditional on t, μ_1 , and μ_2 can be calculated by finding the smallest cutoff C such that

$\sum_{k=C}^t \binom{t}{k} \left(\frac{\mu_2}{\mu_1 + \mu_2}\right)^k \left(\frac{\mu_1}{\mu_1 + \mu_2}\right)^{t-k}$ is less than or equal to the desired α level and then

evaluating $\sum_{k=C}^t \binom{t}{k} \left(\frac{\mu_2 r}{\mu_1 + \mu_2 r}\right)^k \left(\frac{\mu_1}{\mu_1 + \mu_2 r}\right)^{t-k}$. The next term, $P(T = t|H_A, \mu_1, \mu_2)$, is simply the probability density function of a Poisson random variable which has mean $u_1 I_1(\mu_1 + r\mu_2)$. Lastly we assume that, as in the case above, $e(\mu_1)$ and $f(\mu_2)$ are known distributions estimated before the cross-sectional survey. Then the power is

$$P(\text{Reject } H_0 | H_A) = \int_0^\infty \int_0^\infty \sum_{t=0}^\infty \left[\sum_{k=C}^t \binom{t}{k} \left(\frac{\mu_2 r}{\mu_1 + \mu_2 r}\right)^k \left(\frac{\mu_1}{\mu_1 + \mu_2 r}\right)^{t-k} \right] \frac{(u_1 I_1(\mu_1 + r\mu_2))^t}{e^{(u_1 I_1(\mu_1 + r\mu_2))} t!} e(\mu_1) f(\mu_2) d\mu_1 d\mu_2$$

We have solved for $u_1 I_1$ in Mathematica for the specific scenario where $e(\mu_1)$ follows the Gamma distribution shown by the red curve in Panel A of Figure 2 and $f(\mu_2)$ follows the Gamma distribution shown by the green curve in Panel A of Figure 2. We omit those results in the interest of leaving this section as general as possible. The computations require numerical integration and are significantly more intensive than the cases we examined in Section 5 which were solved rapidly in R.

This extension allows for further flexibility if one wishes to use the methods developed in this paper in other contexts. Our methods apply to any situation where $X \sim \text{Poisson}(C \times \lambda)$ where X is observable, uncertainty in C motivates specification of a prior distribution, and inference is focused on λ . This situation arises if λ is the event rate of any condition and X is the observed number of events over an uncertain “length” C . For any disease in steady-state (Freeman and Hutchison 1980) this emerges from the relationship prevalence = (incidence $\times$ average duration) when the average duration of infection is not directly observable. Outside of the arena of incidence estimation we may readily find applications for this methodology. For example, in occupational epidemiology we may be interested in comparing the rate of accidents at two different locations. Traditionally a comparative Poisson trial would be used to compare the rate of accidents. However, uncertainty may exist in the person hours contributed in each location. The methods developed in this paper allow for the traditional comparative Poisson trial to proceed by accounting for this uncertainty by placing a prior distribution on the number of person hours.

8 Discussion

We have derived a methodology for finding the needed sample sizes for both single and consecutive cross-sectional surveys designed to estimate incidence. The methodology gives the required number of uninfected samples to conduct a single survey with a desired amount of precision or successive surveys to detect trends in incidence. These numbers can be adjusted to find the required sample sizes. Since, in practice, cross-sectional surveys occur over a period of several weeks preliminary estimates used for this calculation could be

updated as sampling occurs. Further work could explore how best to make these adjustments.

We have detailed situations where one must account for uncertainty in μ to avoid under sampling. When we account for this uncertainty by placing a Gamma distribution on μ we showed how the shape of the distribution affects the required sample sizes. The Gamma distribution is defined by its mean and variance. Larger sample sizes are required when the variance increases or the mean decreases. In general, larger sample sizes are required the more mass the distribution puts on smaller values of μ . Researchers should keep these concepts in mind when considering a potential biomarker-based testing algorithm to use in a cross-sectional survey. The effects are magnified when the situation already demands large sample sizes such as when the underlying incidence is small, a large amount of precision is required, or we wish to detect a small change in incidence over time. In the United States where incidence is significantly less than 1% ignoring the uncertainty in μ could lead to a sizable loss of precision and reduction in power.

We tested our methodology in four different underlying epidemics. We found that the formula developed for the single survey generally had good coverage. When we examined successive surveys we found that the bias in the cross-sectional estimator can decrease the power when the underlying epidemic is fluctuating. However, we achieved the desired power in a linearly increasing, an exponentially increasing, and an exponentially decreasing epidemic. The samples size calculations therefore seem to be accurate in settings where the underlying methodology behind cross-sectional HIV incidence estimation is not biased. One solution to the possibility of under powering the studies would be to place a prior on the initial estimate of incidence at ψ years before the first time point. This would allow for the reality that the initial estimate will have uncertainty both because we are trying to estimate current incidence and because we will not know the underlying shape of the epidemic.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by National Institutes of Health grant R01-AI095068 as well as the National Institutes of Health Training Grant T32 AI 007370.

Appendix

We asserted that $I_c = \frac{1}{1-p} \frac{\pi}{\mu}$ is approximately measuring the incidence ψ years before the cross-sectional survey. We note that we can define the incidence at any time point before the survey based on the function $g(t)$. For example, the incidence ψ years before the survey can

be written as $I(\psi) = \frac{g(\psi)}{1-p(\psi)}$ where $p(\psi)$ is the prevalence ψ years before the survey.

To see why $I_c \approx I(\psi)$ we write $\pi = \int_0^\infty g(t)\phi(t)dt$, where $\phi(t)$ is the probability that persons infected t years ago will be in the early disease stage at the time of the survey. Noting that $\mu = \int_0^\infty \phi(t)dt$ we define a random variable S whose probability density is given by

$$f_s(t) = \frac{\phi(t)}{\mu} \quad \text{for} \quad t \geq 0. \text{ Thus, } \pi = \mu \cdot E[g(S)] \text{ and } I_c = \frac{1}{1-p} \frac{\pi}{\mu} = \frac{E[g(S)]}{1-p}.$$

If we are willing to assert that $E[g(S)] \approx g(E[S])$. Then $I_c \approx \frac{g(E[S])}{1-p}$. Thus, letting

$$\psi = E[S] = \int_0^\infty t \frac{\phi(t)}{\mu} dt, \text{ and assuming the prevalence at the time of the survey is similar to the}$$

prevalence ψ years ago, we get that $I_c \approx \frac{g(\psi)}{1-p(\psi)} = I(\psi)$. The first assumption that $E[g(S)] \approx g(E[S])$ will hold as long as $g(t)$ is approximately linear where $\phi(t)$ is appreciably greater than 0. Note that when $\phi(t) \approx 0$ the density $f_s(t) \approx 0$ and $E[g(S)] \approx g(E[S])$ by default. The second assumption that $p \approx p(\psi)$ will be reasonable if ψ is a short period of time.

References

- Agresti, A. Categorical data analysis. 3rd. Hoboken, New Jersey: John Wiley & Sons; 2013.
- Brookmeyer R. Accounting for follow-up bias in estimation of human immunodeficiency virus incidence rates. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 1997; 160:127–140.
- Brookmeyer R. Measuring the HIV/AIDS epidemic: Approaches and challenges. *Epidemiologic Reviews*. 2010; 32:26–37. [PubMed: 20203104]
- Brookmeyer R, Konikoff J, Laeyendecker O, Eshleman SH. Estimation of HIV incidence using multiple biomarkers. *American Journal of Epidemiology*. 2013; 177:264–272. [PubMed: 23302151]
- Brookmeyer R, Quinn TC. Estimation of current human immunodeficiency virus incidence rates from a cross-sectional survey using early diagnostic tests. *American Journal of Epidemiology*. 1995; 141:166–172. [PubMed: 7817972]
- Brown CC, Green SB. Additional power computations for designing comparative Poisson trials. *American Journal of Epidemiology*. 1982; 115:752–758. [PubMed: 7081205]
- Casella, G.; Berger, R. Statistical inference. 2nd. Pacific Grove, CA: Duxbury; 2002.
- Duong YT, Qiu M, De AK, Jackson K, Dobbs T, Kim AA, Nkengasong JN, Parekh BS. Detection of recent HIV-1 infection using a new limiting-antigen avidity assay: Potential for HIV-1 incidence estimates and avidity maturation studies. *PLoS ONE*. 2012; 7:e33328. [PubMed: 22479384]
- Freeman J, Hutchison GB. Prevalence, incidence and duration. *American Journal of Epidemiology*. 1980; 112:707–723. [PubMed: 6969024]
- Gail M. Power computations for designing comparative Poisson trials. *Biometrics*. 1974; 30:231–237.
- Joint United Nations Programme on HIV/AIDS. Gap report. Geneva, Switzerland: UNAIDS; 2014.
- Kaplan EH, Brookmeyer R. Snapshot estimators of recent HIV incidence rates. *Operations Research*. 1999; 47:29–37.
- Karon JM, Fleming PL, Steketee RW, De Cock KM. HIV in the United States at the turn of the century: An epidemic in transition. *American Journal of Public Health*. 2001; 91:1060–1068. [PubMed: 11441732]
- Konikoff J, Brookmeyer R, Longosz AF, Cousins MM, Celum C, Buchbinder SP, Seage GR III, Kirk GD, Moore RD, Mehta SH, Margolick JB, Brown J, Mayer KH, Koblin BA, Justman JE, Hodder SL, Quinn TC, Eshleman SH, Laeyendecker O. Performance of a limiting-antigen avidity enzyme immunoassay for cross-sectional estimation of HIV incidence in the United States. *PLoS ONE*. 2013; 8:e82772. [PubMed: 24386116]
- Laeyendecker O, Brookmeyer R, Cousins MM, Mullis CE, Konikoff J, Donnell D, Celum C, Buchbinder SP, Seage GR III, Kirk GD, Mehta SH, Astemborski J, Jacobson LP, Margolick JB,

- Brown J, Quinn TC, Eshleman SH. HIV incidence determination in the United States: A multiassay approach. *Journal of Infectious Diseases*. 2013; 207:232–239. [PubMed: 23129760]
- Mastro TD. Determining HIV incidence in populations: Moving in the right direction. *Journal of Infectious Diseases*. 2013; 207:204–206. [PubMed: 23129757]
- Suligoi B, Massi M, Galli C, Sciandra M, Di Sora F, Pezzotti P, Recchia O, Montella F, Sinicco A, Rezza G. Identifying recent HIV infections using the avidity index and an automated enzyme immunoassay. *Journal of Acquired Immune Deficiency Syndromes*. 2003; 32:424–428. [PubMed: 12640201]

Figure 1. Example testing algorithm to define the early disease stage

An infected sample is excluded from the early disease stage if it does not meet one or more of the four criteria. The following units are used for the component assays: CD4 cell count: cells/mm³; BioRad-Avidity assay: percentage (avidity index); LAg-Avidity: normalized optical density units; viral load: copies/mL. Abbreviations: LAg-Avidity: limiting antigen avidity enzyme immunoassay (Duong et al., 2012) and BioRad-Avidity: BioRad-Avidity assay (Suligoi et al., 2003).

Figure 2. Panel A: Distributions of μ used in sample size calculations and Panel B: W/I_c achieved per number of uninfected samples when $I_c = 1\%$ dependent on distributions of μ in Panel A

Figure 3. Underlying epidemic curves for four simulated scenarios

Figure 4. Daily epidemic changes

Births occur at a rate of 0.004 percent of the uninfected part of the population

Deaths from causes other than HIV occur at a rate of 0.002 percent of the population

Deaths specifically from HIV occur according to Weibull distribution with shape parameter 2.5 and scale parameter 14 so that $h(x)$ is $0.00000013x^{1.5}$ for x is in days

The number of new infection in the population is determined by multiplying the size of the population by $g(t)/365$

Table 1

Needed uninfected samples at the first time point multiplied by initial incidence, $u_1 I_1$, and, in parenthesis, total uninfected samples needed multiplied by the incidence at the first time point, $((u_1 + u_2) I_1)$.

	$r = 1.2$	$r = 1.5$	$r = 2$	$r = 3$	$r = 5$
$s = 2$					
i	898.7 (2696.0)	166.4 (499.0)	50.5 (151.5)	17.0 (50.8)	6.2 (18.6)
ii	907.7 (2723.1)	168.2 (504.5)	51.1 (153.2)	17.2 (51.4)	6.3 (18.8)
iii	1049.0 (3147.0)	195.2 (585.6)	59.6 (178.7)	20.2 (60.4)	7.5 (22.3)
iv	442.8 (1328.2)	82.0 (246.0)	24.9 (74.7)	8.4 (25.1)	3.1 (9.2)
$s = 1$					
i	1196.6 (2393.1)	221.2 (442.3)	67.4 (134.8)	22.2 (44.4)	8.2 (16.3)
ii	1208.8 (2417.5)	223.5 (447.0)	68.1 (136.2)	22.5 (44.9)	8.3 (16.5)
iii	1396.2 (2792.3)	259.2 (518.4)	79.0 (158.0)	26.3 (52.6)	9.8 (19.6)
iv	589.6 (1179.1)	109.0 (218.0)	33.2 (66.4)	11.0 (21.9)	4.1 (8.1)
$s = \frac{1}{2}$					
i	1793.3 (2689.9)	331.1 (496.6)	100.4 (150.6)	33.8 (50.6)	12.1 (18.1)
ii	1811.0 (2716.5)	334.3 (501.5)	101.5 (152.2)	34.1 (51.1)	12.2 (18.3)
iii	2090.3 (3135.4)	386.6 (579.8)	117.7 (176.5)	39.3 (59.0)	14.1 (21.1)
iv	883.4 (1325.1)	163.1 (244.7)	49.5 (74.2)	16.7 (25.0)	6.0 (8.9)

i $\mu = 0.40$

ii $\mu \sim \text{Gamma}(120.77, 0.003)$

iii $\mu \sim \text{Gamma}(8, 0.05)$

iv $\mu \sim \text{Gamma}(500, 0.002)$

r is ratio of the incidence at the second time point to the incidence at the first time point when the alternative hypothesis is true

s is the ratio of the needed uninfected samples at the second time point to the needed uninfected samples at the first time point

Table 2
Comparison of incidences at years 8- ψ and 10- ψ (1) and years 8- ψ and 10- ψ (2) for four epidemic scenarios

	True Incidence	ASI	CI coverage	Power	n
(1)					
Epidemic A	0.80% and 0.30%	0.80% and 0.29%	0.94 and 0.92	0.92	9,984
Epidemic B	0.41% and 1.08%	0.42% and 1.05%	0.95 and 0.95	0.87	7,595
Epidemic C	0.66% and 1.01%	0.65% and 1.01%	0.95 and 0.95	0.90	29,557
Epidemic D	0.21% and 0.94%	0.22% and 0.96%	0.91 and 0.96	0.90	4,848
(2)					
Epidemic A	0.65% and 0.30%	0.64% and 0.29%	0.95 and 0.92	0.92	17,920
Epidemic B	0.62% and 1.08%	0.65% and 1.05%	0.96 and 0.94	0.82	18,383
Epidemic C	0.83% and 1.01%	0.83% and 1.01%	0.95 and 0.95	0.90	123,545
Epidemic D	0.45% and 0.94%	0.45% and 0.96%	0.95 and 0.96	0.90	12,960

ψ is estimated to be 0.48 years

ASI: Average simulated incidence over all bootstraps and trials

Epidemics A, B, C, and D are the epidemic scenarios labeled in Figure 3

CI coverage is the percentage of the 999,000 incidences covered by the confidence interval created according to the formula in Section 4

Power is calculated according to the formula in Section 5 with equal sample sizes

n is the needed uninfected samples