Published in final edited form as:

Biometrics. 2010 March ; 66(1): 1–10. doi:10.1111/j.1541-0420.2009.01242.x.

Estimating HIV Incidence based on Combined Prevalence Testing

Raji Balasubramanian¹ and Stephen W. Lagakos²

Raji Balasubramanian: rbalasub@schoolph.umass.edu

¹Division of Biostatistics and Epidemiology, School of Public Health and Health Sciences
University of Massachusetts, 715 North Pleasant Street, Amherst, MA 01003

²Department of Biostatistics, Harvard University School of Public Health, 655 Huntington Avenue,
Boston, Massachusetts 02115, U.S.A

Summary

Knowledge of incidence rates of HIV and other infectious diseases is important in evaluating the state of an epidemic as well as for designing interventional studies. Estimation of disease incidence from longitudinal studies can be expensive and time-consuming. Alternatively, Janssen et al., 1998 proposed the estimation of HIV incidence at a single point in time based on the combined use of a standard and “detuned” antibody assay. This paper frames the problem from a longitudinal perspective, from which the maximum likelihood estimator of incidence is determined and compared with the Janssen estimator. The formulation also allows estimation for general situations, including different batteries of tests among subjects, inclusion of covariates, and a comparative evaluation of different test batteries to help guide study design. The methods are illustrated with data from an HIV interventional trial and a seroprevalence survey recently conducted in Botswana.

Keywords

clinical trials; cross sectional studies; incidence rate

1. Introduction

Estimates of HIV prevalence and incidence are widely used to assess the state of the epidemic as well as to evaluate the impact of medical and behavioral interventions. HIV prevalence can be estimated relatively easily from cross-sectional samples of the population, typically based on diagnostic tests for detecting the presence of HIV-specific antibodies in blood. In contrast, because HIV infections are silent events, estimation of HIV incidence is much more difficult. The most direct approach is through observational studies in which subjects are periodically monitored for HIV infection. However, such studies are time-consuming and expensive, and still may not provide precise enough estimates of incidence for certain applications, such as determining the required sample size for an HIV prevention study. Other approaches for estimating incidence can be less costly, but in general require additional assumptions. These include studies that estimate HIV incidence based on changes over time in prevalence (Podgor and Leske, 1986; Cleghorn et al., 1998; Williams et al.,

8. Supplementary Materials

Web Appendices referenced in Sections 2,3,4 and 7 are available under the Paper Information link at the Biometrics website
<http://www.biometrics.tibs.org>.

2001; Hallet et al., 2008), back calculation of HIV incidence based on AIDS incidence (Brookmeyer and Gail, 1994), and other approaches (Saidel et al., 1996).

An important advance in this area was given by Janssen et al. (1998), who made use of the fact that HIV infection is followed within a few weeks with the development of HIV-specific antibodies. By developing a “detuned” antibody assay designed to be less sensitive than the standard ELISA antibody assay, Janssen et al. (1998) identify individuals with a “recent” HIV infection, defined by a positive standard ELISA and a negative detuned ELISA test result, and then adapt methods proposed by Brookmeyer and Quinn (1995) to estimate the HIV incidence rate. In a subsequent paper, Kaplan and Brookmeyer (1999) developed a “snapshot” estimator of HIV incidence based on a sample obtained at a single time point, for which the method of Janssen et al. (1998) is a special case.

Two studies recently conducted in Botswana motivated our interest in this problem. One was a randomized clinical trial (Thior et al., 2006) aimed at preventing mother to child transmission of HIV; the other was a HIV serosurvey conducted in 2003 by the Debswana mining company (Executive Report, Debswana Diamond Company). Goals in both studies, which used standard and detuned ELISA testing, included the estimation of HIV incidence rate over time and in persons of different ages and job types.

The purpose of this paper is to build on the initial results of Janssen et al. (1998) by framing the problem using a maximum likelihood based approach, and from this deriving likelihood-based estimators of HIV incidence for more general settings, estimators of other quantities of interest in applying this approach, and methods for assessing the impact on incidence of multiple covariates. For typical HIV applications, the methods of Janssen et al. (1998) are shown to closely approximate the maximum-likelihood estimators. A particular advantage of this framework is that it allows for a natural extension of the methods to estimate incidence in settings where subjects are tested at different ages and time points, where testing may be based on different test batteries and, importantly, which can incorporate the effects of covariates. The methodology also provides a theoretical basis for comparing testing strategies with respect to the precision of incidence estimators and expected time between infection and detection.

In Section 2 we give the model for the underlying disease process, the observables and the resulting likelihood contributions. In Section 3 we develop likelihood equations under several sampling scenarios, based on the types of diagnostic tests that are administered and the time period during which they are administered. In Section 4 we undertake efficiency comparisons of various batteries of tests and discuss the implications for the design of studies of incidence based on combined prevalence testing. In Section 5 we extend the methods to assess the association between measured covariates and the incidence rate. We illustrate the methods in Section 6 with the Botswana studies that motivated our interest in this problem, and discuss some related issues in Section 7.

2. Model, Probability Elements, Observations and Likelihood Contributions

We conceptualize a person’s HIV history by a progression through 4 states, denoted S_0 , S_1 , S_2 , and S_3 . State S_0 represents the uninfected state, and extends from birth until the individual first develops detectable HIV antigen. State S_1 represents the “acute infection” state, in which the subject has detectable HIV antigen, but has not yet developed detectable HIV antibodies. S_2 denotes the “recent infection” state, where the subject has detectable (by standard ELISA) HIV antibodies, but does not yet have antibody levels detectable by the detuned ELISA. Finally, S_3 represents the “nonrecent infection” state in which the subject’s antibody levels are detectable by the detuned ELISA. Busch et al. (1995) provide a technical discussion and estimates for the various sojourn times (also referred to as “window period”)

associated with the states S_1 , S_2 and S_3 . Denoting the results of the nucleic acid antigen (A), standard ELISA (E), and detuned ELISA (D) tests as “+” and “-” for positive and negative, respectively, S_0 corresponds to (A-), S_1 to (A+, E-), S_2 to (E+, D-), and S_3 to (D+). Note that D+ implies that both A and E are positive, and E+ implies that A is also positive.

For a subject born at calendar time t_0 , let T denote the calendar time of HIV infection, and let $f(t|t_0)$, $\lambda(t|t_0)$ and $F(t|t_0)$ denote the density, hazard and cumulative distribution functions of T at calendar time $t|t_0$, when the subject is age $t-t_0$. The sojourn times (i.e. window periods) in the states S_1 and S_2 are denoted by the random variables L_1 and L_2 , which we assume are independent of T but not of one another. We describe the joint distribution of (L_1, L_2) by the cumulative distribution functions $G_1(u) = P[L_1 \leq u]$ and $G_2(v|u) = P[L_2 \leq v|L_1 = u]$. We assume that L_j has support in the interval $[0, L_j^*]$, for $j = 1, 2$, and that $L_{12}^* = L_1^* + L_2^* < t$. For HIV infection, L_1^* and L_2^* are on the order of 2–6 weeks, and 5–7 months, respectively.

It is not difficult to show that the prevalence probabilities of the 4 states are given by:

$$\begin{aligned}\pi_0(t|t_0) &\stackrel{\text{def}}{=} P[\text{in } S_0 \text{ at time } t | \text{born at time } t_0] = 1 - F(t|t_0), \\ \pi_1(t|t_0) &\stackrel{\text{def}}{=} P[\text{in } S_1 \text{ at time } t | \text{born at time } t_0] = \int_{\max(0, t-L_1^*)}^t f(u|t_0) [1 - G_1(t-u)] du, \\ \pi_2(t|t_0) &\stackrel{\text{def}}{=} P[\text{in } S_2 \text{ at time } t | \text{born at time } t_0] = \int_{\max(0, t-L_{12}^*)}^t \int_{\max(0, t-u-L_2^*)}^{\min(t-u, L_1^*)} f(u|t_0) g_1(v) [1 - G_2(t-u-v|v)] dv du, \text{ and} \\ \pi_3(t|t_0) &= 1 - \pi_0(t|t_0) - \pi_1(t|t_0) - \pi_2(t|t_0)\end{aligned}$$

We focus on estimating $F(t|t_0)$ and $\lambda(t_0) = f(t|t_0)[1 - F(t|t_0)]$, the HIV prevalence and incidence rate at calendar time t , for someone born at calendar time t_0 . We later consider extensions of the methods when testing is done on subjects of different ages, where the proposed methods can be adapted without assumptions regarding the sojourn times in states S_1 , S_2 , S_3 . When $f(u|t_0)$ is constant in u for $u \in [t-L_{12}^*, t]$, say $f(u|t_0) = f(t_0)$, the incidence rate at time t is given by $\lambda(t|t_0) = f(t_0)[1 - F(t|t_0)]$ and the expressions for $\pi_j(t|t_0)$ simplify to

$$\begin{aligned}\pi_0(t|t_0) &= [1 - F(t|t_0)] \\ \pi_1(t|t_0) &= f(t_0)k_1(t) \\ \pi_2(t|t_0) &= f(t_0)k_2(t), \quad \text{where}\end{aligned}$$

$$k_1(t) = \int_{\max(0, t-L_1^*)}^t [1 - G_1(t-u)] du = E[L_1], \quad \text{and} \quad (1)$$

$$k_2(t) = \int_{\max(0, t-L_1^*-L_2^*)}^t \int_{\max(0, t-u-L_2^*)}^{\min(t-u, L_1^*)} g_1(v) [1 - G_2(t-u-v|v)] dv du. \quad (2)$$

When $L_1^* < t$, as will be the case in any practical setting, $k_1(t) = E[L_1]$. It is easily shown that $k_2(t) = E[L_2]$ when L_1 and L_2 are independent. $k_2(t)$ might be closely approximated by $E[L_2]$ when $\text{Var}(L_1) \ll \text{Var}(T)$ or when $\text{Var}(L_1) \approx 0$ because in both cases $T + L_1$ is approximately independent of L_2 .

For someone found to be in State j at time t , we are also interested in the elapsed time since they entered this state, say $Y_1 = t - T$ when $j = 1$ and $Y_2 = t - T - L_1$ when $j = 2$. The conditional density functions of Y_1 and Y_2 , derived in Web Appendix A, are

$$f_1(y|t_0, t) = \frac{[1-G_1(y)]f(t-y|t_0)}{\pi_1(t)} \quad \text{for } 0 \leq y \leq L_1^*; \text{ and}$$

$$f_2(y|t_0, t) = \frac{[1-G_2(y|t-y)] \int_0^{t-y} f(u|t_0)g_1(t-y-u)du}{\pi_2(t)} \quad \text{for } 0 \leq y \leq L_2^*.$$

When $f(u|t_0) = f(t_0)$ for $u \in [t-L_{12}^*, t]$, these simplify to

$$f_1(y|t_0, t) = \frac{[1-G_1(y)]}{E[L_1]} \quad \text{for } 0 \leq y \leq L_1^*,$$

which doesn't depend on t_0 or t , and

$$f_2(y|t_0, t) = \frac{[1-G_2(y|t-y)]}{k_2(t)} \quad \text{for } 0 \leq y \leq L_2^*,$$

which depends on t but not t_0 . The first two moments of Y_1 are

$$E[Y_1|t_0, \text{ in } S_1 \text{ at } t] = \frac{E[L_1^2]}{2E[L_1]}, \quad E[Y_1^2|t_0, \text{ in } S_1 \text{ at } t] = \frac{E[L_1^3]}{3E[L_1]}.$$

If L_1 and L_2 are independent, the first two moments of Y_2 can be shown to equal

$$E[Y_2|t_0, \text{ in } S_2 \text{ at } t] = \frac{E[L_2^2]}{2E[L_2]}, \quad E[Y_2^2|t_0, \text{ in } S_2 \text{ at } t] = \frac{E[L_2^3]}{3E[L_2]},$$

and thus independent of both t_0 and t . Similar expressions are derived by Kaplan and Brookmeyer (1999). Note that $E[Y_j|t_0, \text{ in } S_j \text{ at } t] = E[L_j]/2$ for $j = 1, 2$, reflecting a type of length-biased sampling.

Suppose that an individual born at calendar time t_0 is tested at calendar time t using A, E, and D, and let the results of these tests be denoted r_1 , r_2 and r_3 , respectively, where $r_j = 1$ denotes a positive test and $r_j = 0$ denotes a negative test. To allow for settings where only a subset of the 3 tests are given, we use $r_j = -1$ to denote that test j is not given, for $j = 1, 2, 3$. Thus, for example, the observation $(r_1, r_2, r_3) = (-1, 1, 0)$ means that the subject was not given the antigen test, and tested positive for the standard ELISA and negative for the detuned ELISA. Table 1 gives the probabilities of the 19 possible outcomes based on all combinations of the 3 assays, administered at t . The horizontal lines between rows delineate different sampling strategies. For example, the first 2 rows represent use of only the antigen assay, whereas the last 4 rows represent the setting where all three assay results are determined.

3. Likelihood Functions and Inference

In this section, we consider two specific testing cases that arise in practice and develop the simplified form of the likelihood in each of these settings. We initially assume that all subjects are the same age and tested at the same time, and thus suppress t and t_0 in the notation. We later discuss and illustrate extensions of the methods when testing is done at different time points, or at one time point but for subjects of different ages. We also consider how specific parametric assumptions regarding the form of $f(\cdot)$ in the interval $(t-L_1^*-L_2^*, t)$ can further simplify the likelihood.

3.1 Subjects tested with different batteries of tests

Suppose each of n subjects may be tested with any of the seven testing strategies enumerated in Table 1. For example, all subjects might be given the standard and detuned ELISA assays, and a random sample are also given the antigen assay. Let n_{jkl} denote the number of subjects with test result jkl , where $j = -, 0$ or 1 to denote that A was not done, negative, or positive, respectively, and where k and l are defined similarly. For example, n_{-11} denotes the number of subjects that were tested using (E,D) and who were positive on both tests. Then if p_{jkl} denotes the probability associated with test outcome (jkl) , the likelihood function is

$$L = \prod_{\{j,k,l\}} p_{jkl}^{n_{jkl}} = (\pi_0 + \pi_1)^{n_{-0-} + n_{-00} + n_{000} + n_{00-}} \times (\pi_1 + \pi_2)^{n_{1-0}} (\pi_2 + \pi_3)^{n_{11-} + n_{-1-}} (\pi_0 + \pi_1 + \pi_2)^{n_{--0}} \times \pi_2^{n_{-10} + n_{110}} \pi_1^{n_{10-} + n_{100}} (\pi_1 + \pi_2 + \pi_3)^{n_{1--}} \pi_3^{n_{-1-1} + n_{-11} + n_{1-1} + n_{111}}, \quad (3)$$

subject to $\pi_0 + \pi_1 + \pi_2 + \pi_3 = 1$. Without any assumptions about the form of $f(\cdot)$, it cannot, in general, be disentangled from the functions G_j that contribute to the form of the directly observable prevalence π_{jkl} . We therefore consider different parametric forms for $f(\cdot)$ in order to simplify the likelihood in (3).

First assume that $f(u) = f$ for $u \in [t-L_1^*-L_2^*, t]$. The likelihood (3) can then be written as products of linear combinations of terms involving f and $\theta = F(t | t_0)$ as follows:

$$L = [1 - \theta + fk_1]^{n_{-0-} + n_{-00}} [1 - \theta]^{n_{00-} + n_{000} + n_{00-}} [fk_1 + fk_2]^{n_{1-0}} \times [\theta - fk_1]^{n_{11-} + n_{-1-}} [1 - \theta + fk_1 + fk_2]^{n_{--0}} [fk_2]^{n_{-10} + n_{110}} \times [fk_1]^{n_{10-} + n_{100}} [\theta]^{n_{1--}} [\theta - fk_1 - fk_2]^{n_{-1-1} + n_{-11} + n_{1-1} + n_{111}}, \quad (4)$$

where k_1 and k_2 are given in (1) and (2). For given k_1 and k_2 , the maximum likelihood estimator of $\theta = F(t | t_0)$ and f can be obtained from (4) by joint maximization of the likelihood using numerical techniques. Approximate confidence intervals and standard errors for MLEs of θ , f , or functions of these components, can be obtained from the Hessian of the log-likelihood function (Cox and Hinkley (1974)). Analytical expressions for the Hessian matrix are provided in equations (1)–(15) in Web Appendix B. Alternatively, the density function for someone born at t_0 could be assumed to be piece-wise constant, with $f(u | t_0) = f_1$ for $u \in (t-L_1^*, t]$ and $f(u | t_0) = f_2$ for $u \in (t-L_1^*-L_2^*, t-L_1^*)$. Expressions for π and simplified forms of the likelihood are presented in Web Appendix C.

3.2 All subjects tested with the same battery of tests

When all subjects are evaluated using the same battery of tests and $f(u) = f$ for $u \in [t-L_1^*-L_2^*, t]$, the resulting expressions for the likelihood when the battery consists of AED, AE, AD, and ED are, respectively:

$$L_{AED} = [1 - \theta]^{n_{000}} [fk_1]^{n_{100}} [fk_2]^{n_{110}} [\theta - fk_1 - fk_2]^{n_{111}} \quad (5)$$

$$L_{AE} = [1 - \theta]^{n_{00-}} [fk_1]^{n_{10-}} [\theta - fk_1]^{n_{11-}} \quad (6)$$

$$L_{AD} = [1 - \theta]^{n_{0-0}} [fk_1 + fk_2]^{n_{1-0}} [\theta - fk_1 - fk_2]^{n_{1-1}} \quad (7)$$

$$L_{ED} = [1 - \theta + fk_1]^{n_{-00}} [fk_2]^{n_{-10}} [\theta - fk_1 - fk_2]^{n_{-11}}. \quad (8)$$

The maximum likelihood estimator of θ , f can be obtained by joint maximization of the log-likelihood. In each of the above four cases, closed form solutions for the MLEs of f , θ and $\lambda = \frac{f}{1-\theta}$, exist and are presented in Table 2. For example, when using the (E,D) battery, the estimated incidence rate is given by the ratio of the number of recent infections to the quantity $n_{-00}k_2 - n_{-10}k_1$. Approximate confidence intervals and standard errors for MLEs of θ , f , or functions of these components, can be obtained from the Hessian of the log-likelihood function. Analytical expressions for the elements of the Hessian matrix are presented in equations (1)–(15) in Web Appendix B.

As described in Section 3.1, the density function for someone born at t_0 could be assumed to be piece-wise constant. See Web Appendix C for details.

3.3 Large-Sample Properties and Connection to the Janssen et al. (1998) Estimator

For a given battery of tests, the consistency of the proposed estimators of $\lambda(t | t_0)$ follows from the fact that the proportions of subjects with different test results converge to the expressions for $\pi_j(t | t_0)$ (Table 1). Under mild regularity conditions, the standardized estimators of $\lambda(t | t_0)$ and $F(t | t_0)$ will converge to the Gaussian distribution as $n \rightarrow \infty$.

As above, suppose all subjects are born at time t_0 and tested at time t . When the (E,D) battery of tests is used, the MLE of incidence rate is (see Table 2):

$$\hat{\lambda} = \frac{n_{-10}}{n_{-00}k_2 - n_{-10}k_1}.$$

The estimated incidence rate proposed by Janssen et al. (1998) is given by

$$\hat{\lambda}_j = \frac{n_{-10}}{n_{-00}E[L_2]}.$$

Under the assumption that $L_1 = 0$, $\hat{\lambda}$ reduces to $\hat{\lambda}_j$. The Janssen et al. (1998) estimator can also be derived as a special case of the snapshot estimator (Kaplan and Brookmeyer, 1999), when the target region is assumed to be S_2 (i.e. the subject has detectable antibodies by E but is undetectable by the D assay). When L_1 and L_2 are independent, $E[L_2] = k_2(t)$. Also, for infections like HIV, $\lambda(t_0)k_1$ is much smaller than 1, in which case the Janssen et al. (1998) estimator should give similar results as the MLE $\hat{\lambda}$. The probability limit of the Janssen et al. (1998) estimator is

$$\hat{\lambda}_j = \frac{n-10}{n-00E[L_2]} = \frac{\frac{n-10}{n}}{\frac{n-00}{n}E[L_2]} \xrightarrow{P} \frac{\pi_2(t|t_0)}{[\pi_0(t|t_0)+\pi_1(t|t_0)]E[L_2]} = \frac{f(t_0)k_2(t)}{[1-F(t|t_0)+f(t_0)k_1]E[L_2]}.$$

Thus, when L_1 and L_2 are independent and $\lambda(t_0)k_1$ is much smaller than 1, the probability limit of the Janssen et al. (1998) estimator is very close to $\lambda(t_0)$. We performed a variety of simulations using the (E,D) battery, including some with L_1 and L_2 being dependent, and in almost all cases $\hat{\lambda}_J$ and $\hat{\lambda}$ gave very similar results.

4. Design Considerations

When designing a cross-sectional study to estimate incidence rates, one issue is the choice of batteries of tests (see Table 1) involving at least 2 of (A,E,D). If the window period in state S_2 were modifiable, say by changing the optical density threshold used to define a positive test, another issue is the optimal window period for the detuned ELISA (D). We assess these issues by using the analytical expressions for the variance of the MLEs of (f, θ) ((1)–(15) in Web Appendix B) to compare the relative efficiency of the estimate of disease incidence, λ , at calendar time t for the different test batteries. We also examine the expected values of the time spent in state j for a subject found to be in this state at time t . That is, for the AED and ED batteries, we consider the expected time in S_2 for someone found to be in state S_2 , and for the AD battery, we consider the expected times since HIV infection, given that the subject was found to be in S_1 or S_2 , respectively.

Suppose that $n = 100$ and that L_1 and L_2 have independent normal distributions with $E(L_1) = 35$ days, $\text{Var}(L_1) = 3.5$ and $\text{Var}(L_2) = 10$ days. Figure 1(a) presents variance of the MLE of λ for values of $E(L_2)$ ranging from 100 to 300 days, for the AE, AD, ED and AED batteries, where prevalence is assumed to be 0.10 and incidence rate (per year) to be 0.01. Figure 1(b) presents the expected sojourn time for AED, AD and ED batteries.

Fig. 1(a) shows that the AED battery results in the lowest variance. The variance of the estimate of λ for the AD battery coincides with that for AED, since the analytical expressions for the MLEs are identical (See Table 2). The variance of the MLE of λ obtained with the ED battery is higher than that for AED and AD - however, for larger values of $E(L_2)$, the variance approaches that of the AED and AD batteries. Thus, for larger values of $E(L_2)$, the asymptotic relative efficiency (ARE) comparing ED to the AED battery ($\frac{\text{Var}_{ED}}{\text{Var}_{AED}}$) approaches 1. The largest variance for the MLE of λ is obtained from the AE battery. Fig 1b shows the expected sojourn time in state j for a subject found to be in state j at time t for each of the 3 batteries. We see that as $E(L_2)$ increases, expected sojourn time in state j increases. Moreover, for detuned assays with larger values of $E(L_2)$ (i.e $E(L_2) > 150$ days), the expected sojourn time in state j increases more rapidly than the rate of decrease in the variance of the MLE of incidence.

We also examined settings where the true incidence rate (per year) was increased to 0.1 and the true prevalence rate was 0.30, which result in a decrease in variance of the estimate of incidence for all four test batteries. However, the efficiencies of the three two-test batteries relative to the AED battery are unaffected (results available upon request).

5. Covariates

Suppose $\lambda(u | t_0) = (t_0)$ for $u \in (t-L_{12}^*, t]$, $\mathbf{Z}$ is a vector of covariates measured on each subject, and we want to assess the association between $\mathbf{Z}$ and the HIV incidence rate. Consider the proportional hazards model

$$\lambda(u|\mathbf{Z}) = \lambda(t_0)e^{\beta\mathbf{Z}}. \quad (9)$$

For a set of observations using any combination of diagnostic tests, the parameters of this model can be estimated by standard likelihood methods, using the likelihood equations (4)–(8). In general, the MLE of β cannot be expressed in closed form, but can be easily found using standard numerical methods.

Alternatively, consider the random variable $Y(t)$ defined to be j if the subject is in state j at time t , for $j = 0, 1, 2, 3$, and define the odds function

$$\mathcal{O}(j|j, k) = \frac{P[Y(t)=j|Y(t)=j \text{ or } k]}{1 - P[Y(t)=j|Y(t)=j \text{ or } k]} = \frac{\pi_j(t)}{\pi_k(t)}. \quad (10)$$

It follows from the expressions for the $\pi_j(t)$ that

$$\mathcal{O}(1|0, 1) = \lambda(t_0)k_1 \quad (11)$$

$$\mathcal{O}(2|0, 2) = \lambda(t_0)k_2 \quad (12)$$

$$\mathcal{O}(1, 2|0, 1, 2) = \lambda(t_0)[k_1 + k_2] \quad (13)$$

$$\mathcal{O}(2|0, 1, 2) = \frac{\lambda(t_0)k_2}{1 + \lambda(t_0)k_1}, \quad (14)$$

Thus, if $\mathcal{O}(j|j, k; \mathbf{Z})$ denotes the odds for a subject with covariate vector $\mathbf{Z}$, and if k_1 and k_2 do not depend on $\mathbf{Z}$, it follows from (9) that

$$\begin{aligned} \log\{\mathcal{O}(1|0, 1; \mathbf{Z})\} &= \log\{\lambda(t|\mathbf{Z})k_1\} = \log\{k_1\lambda(t_0)\} + \beta\mathbf{Z} = \alpha_1 + \beta\mathbf{Z}, \\ \log\{\mathcal{O}(2|0, 2; \mathbf{Z})\} &= \log\{\lambda(t|\mathbf{Z})k_2\} = \log\{k_2\lambda(t_0)\} + \beta\mathbf{Z} = \alpha_2 + \beta\mathbf{Z}, \text{ and} \\ \log\{\mathcal{O}(1, 2|0, 1, 2; \mathbf{Z})\} &= \log\{\lambda(t|\mathbf{Z})[k_1 + k_2]\} = \log\{[k_1 + k_2]\lambda(t_0)\} + \beta\mathbf{Z} = \alpha_{12} + \beta\mathbf{Z}, \end{aligned}$$

$\log\{\mathcal{O}(2|0, 2; \mathbf{Z})\} = \log\{\lambda(t|\mathbf{Z})k_2\} = \log\{k_2\lambda(t_0)\} + \beta\mathbf{Z} = \alpha_2 + \beta\mathbf{Z}$, and $\log\{\mathcal{O}(1, 2|0, 1, 2; \mathbf{Z})\} = \log\{\lambda(t|\mathbf{Z})[k_1 + k_2]\} = \log\{[k_1 + k_2]\lambda(t_0)\} + \beta\mathbf{Z} = \alpha_{12} + \beta\mathbf{Z}$, where $\alpha_1 = \log\{k_1\lambda(t_0)\}$, $\alpha_2 = \log\{k_2\lambda(t_0)\}$, and $\alpha_{12} = \log\{[k_1 + k_2]\lambda(t_0)\}$. The regression coefficient β in (9) can be estimated by fitting a logistic regression model to results of certain batteries of tests. For example, when (A,E) is used, one can discard observations for which $(A, E) = (+, +)$, regard subjects with $(A, E) = (+, -)$ as “successes” and subjects with $(A, E) = (-, -)$ as “failures”, and analyze the results using logistic regression applied to equation (11).

Similarly, if (A,E,D) is used, one can apply a logistic regression model to equation (12), by discarding observations where $(A, E, D) = (+, -, -)$ or $(+, +, +)$, regarding recent infections $[(A, E, D) = (+, +, -)]$ as successes and regarding uninfected subjects $[(A, E, D) = (-, -, -)]$ as failures. Alternatively, one could use equation (13), discarding observations where $(A, E, D) = (+, +, +)$, regarding acute or recent infections $[(A, E, D) = (+, -, -)$ or $(+, +, -)]$ as successes and regarding uninfected subjects $[(A, E, D) = (-, -, -)]$ as failures. For each logistic model, the estimated intercept can be combined with an estimate of k_1 and/or k_2 to obtain an estimate of the “baseline” hazard $\lambda(t_0)$.

In settings such as HIV, when $\lambda(t_0)k_1 \ll 1$, logistic models can also be applied based on (14) when using the (E,D) battery, by approximating

$$\frac{\lambda(t_0)k_2}{1+\lambda(t_0)k_1} \approx \lambda(t_0)k_2, \text{ so that}$$

$$\log\{\mathcal{O}(2|0, 1, 2; \mathbf{Z})\} \approx \log\{\lambda(t|\mathbf{Z})k_2\} = \log\{k_2\lambda(t_0)\} + \beta\mathbf{Z} = \alpha_2 + \beta\mathbf{Z}.$$

Thus, by discarding observations where $(E, D) = (+, +)$, regarding recent infections $[(E, D) = (+, -)]$ as successes and antibody negative $[(E, D) = (-, -)]$ subjects as failures, one can estimate β using logistic regression. Alternatively, one can avoid this approximation by fitting the logistic model, and then solving the equations

$$\log\left(\frac{\lambda_Z k_2}{1 + \lambda_Z k_1}\right) = \alpha + \beta\mathbf{Z} \quad (15)$$

for λ_Z for each value of $\mathbf{Z}$.

More generally, suppose that subjects are tested at J different times, say $t_1, t_2, \dots, t_J$ and (9) applies to each t_j , but with a possibly different baseline rate, then the same arguments as above can be used, leading to a logistic regression model with covariates $(\mathbf{Z}, W_2, W_3, \dots, W_J)$, where $W_j = 1$ if a subject is tested at time t_j and $W_j = 0$ otherwise, for $j = 2, \dots, J$. Homogeneity of the HIV incidence rate over the different testing times can be assessed by testing the null hypothesis that the coefficients of the $J - 1$ indicator variables are all equal to 0. Alternatively, a time trend in HIV incidence can be fit and tested by defining a scalar covariate, say W , to denote time. For example, taking $W = d(t_j - t_1)$ for $j = 1, 2, \dots, J$ assumes a linear trend in HIV incidence. Note that these inferences about covariate associations are invariant to the choice of L_2 , and thus do not depend on accurately knowing L_2 . The proposed methods for covariate adjustments exhibit similarities to the incidence estimator proposed in an earlier paper by Brookmeyer et al. (1995) that does not also require knowledge of L_2 . However, the methods described in Brookmeyer et al. (1995) require data collected from both a single time point as well as from a longitudinal follow up study.

6. Example

We illustrate methods proposed in this paper using two studies recently conducted in Botswana - (1) the 'MASHI' clinical trial (Thior et al., 2006) for preventing mother to child transmission of HIV, and (2) the 2003 HIV serosurvey conducted by the Debswana Mining Company (Executive Report, Debswana Diamond Company). All subjects were given the (E,D) battery. Stratum specific incidence rates were calculated based on the analytical expressions in Table 2 and predicted incidence rates were calculated from logistic regression analysis following the methodology outlined in Section 5. In the analyses presented below, k_1 and k_2 were assumed to be 35 and 200 days, respectively. We present results from each study and for the combined data.

The Mashi trial consisted of data on 2479 women. Incidence rates were calculated within each stratum defined by age group and calendar time period (details available upon request). Predicted cumulative incidence rates (based on equation (15)) were obtained from the analysis of the subset of data from subjects who were (E-, D-) or (E+, D-) (See Table 3). A quadratic effect of time was found to be statistically significant and is thus included in the model. The results revealed that age groups 20-29 and 40+ have higher odds of recent HIV infection with odds ratios of 2.53 and 2.41 respectively, as compared to subjects under 20 years of age. Figure 2 shows stratum specific incidence rates and the predicted cumulative

incidence rates (per year) from the logistic regression model as a function of age group and calendar time. The predicted incidence rate shows a non-linear relationship with calendar time, and fits the stratum specific estimates well, except for the 20–29 age group for the September 2001–December 2001 time period, where the number of subjects ($n=30$) and events ($n=7$) were small. An interaction between age and calendar time (linear term) was marginally significant ($p = 0.07$, likelihood ratio test) - however, this did not have a substantial impact on the predicted incidence rates (details available upon request).

The Debswana Mining Company dataset included data on 3339 men and women screened in June 2003. Stratum specific incidence rates were calculated within each stratum defined by age group, job type and gender (details available upon request). Predicted cumulative incidence rates were obtained in a similar fashion as in the MASHI dataset (See Table 3). There were no significant interactions between age, job type, and gender. Estimated incidence declined with age group, males had a lower estimated infection risk than females, and non-contractors had a lower estimated risk than contractors. Figure 3 shows stratum specific incidence rates and the predicted cumulative incidence rates (per year) from the logistic regression model as a function of age group, job type and gender. The predicted estimates appear to fit the stratum specific estimates well.

A combined analysis by integrating data from both cohorts was carried out by considering the covariates age, calendar time, gender and job type in a logistic regression analysis, with pregnant women participating in MASHI considered as a separate job category (See Table 3). The association of calendar time with infection risk is similar to that seen in the analysis of MASHI data, and the associations of gender and job type (Contractor Versus Non-contractor) are similar to those for the Debswana cohort. MASHI participants (job type = other) have a much lower incidence than participants in the Debswana study. The integrated analysis leads to a somewhat different association between age group and risk than the individual analyses. With only main effects terms, the integrated analysis suggests that the youngest age group (10–19 years) has the lowest odds of recent HIV infection. However, as expected from the results of the cohort specific analyses, an interaction between age and job type was statistically significant, reflecting the differences between cohorts with respect to the trends in incidence across age groups (results available upon request).

7. Discussion

The methods developed in this paper are not restricted to the detuned ELISA assay, but can also be applied to the recently-developed BED capture enzyme immunoassay, which potentially may enjoy some advantages over the detuned assay originally considered by Janssen et al. (1998) (Parekh et al., 2002; Hu et al., 2003; McDougal et al., 2006). The methods can be applied to estimation of the incidence of any disease or infection for which there are diagnostic tests with varying sensitivity, such as assessing the safety of donated tissue or blood (Zou et al., 2004; Fang et al., 2003).

In some applications, the diagnostic test being used to identify a subject's state may be imperfect. If the diagnostic test has either imperfect specificity or imperfect sensitivity, but not both, the methods introduced in this paper can be easily extended, and closed-form estimators of incidence are still available for the settings reflected in Table 3. Details can be found in Web Appendix D. For more general settings, likelihood equations are easily developed but do not have closed-form solutions.

In practice, perhaps the most important limitation of cross-sectional methods to estimate HIV incidence is the fact that L_2 may not be known reliably. This clearly can impact the accuracy of the estimated incidence, and thus more longitudinal studies to better characterize L_2 are needed. Fortunately, the methods presented in Section 5 for assessing the importance

of a covariate do not depend on knowing L_2 , as it only arises in intercept terms of the logistic regression analyses.

Finally, we note that the 4-state model we assume does not account for the fact that, in practice, subjects can die at any point, and that the risk of death is increased among persons with later-stage, symptomatic HIV infection. For example, suppose that following entrance into State 3 (detuned assay positive), subjects are asymptomatic for a period of time, during which their risk of death is not modified, but that upon the development of symptoms, the risk of death is increased. Then, as shown in the Web Supplement E, the prevalence functions $\pi_j(t|t_0)$ developed in this paper are distorted. However, using this same model, it is shown that the ML and Janssen estimates of incidence developed in Section 3 remain valid, provided that HIV testing is based on sampling asymptomatic subjects.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are grateful to the referees for their helpful comments, the Debswana Mining Company who funded the 2003 Seroprevalence Survey and Max Essex and the investigators of the MASHI study for their permission to use the data. This research was supported in part by grant AI24643 from the National Institutes of Health.

References

- Brookmeyer, R.; Gail, M. AIDS epidemiology: a quantitative approach. Oxford University Press; 1994.
- Brookmeyer R, Quinn T. Estimation of current human immunodeficiency virus incidence rates from a cross-sectional survey using early diagnostic tests. American Journal of Epidemiology. 1995; 141:166–172. [PubMed: 7817972]
- Brookmeyer R, Quinn T, Shepherd M, Mehendale S, Rodrigues J, Bollinger R. The AIDS epidemic in India: A new method for estimating current human immunodeficiency virus (HIV) incidence rates. American Journal of Epidemiology. 1995; 142:709–713. [PubMed: 7572940]
- Busch M, Lee L, Satten G, Henrard D, Farzadegan H, Nelson K, Read S, Dodd R, Petersen L. Time course of detection of viral and serologic markers preceding human immunodeficiency virus type 1 seroconversion: implications for screening of blood and tissue donors. Transfusion. 1995; 35:91–97. [PubMed: 7825218]
- Cleghorn F, Jack N, Murphy J, Edwards J, Mahabir B, Paul R, O'Brien T, Greenberg M, Weinhold K, Bartholomew C, Brookmeyer R, Blattner W. Direct and indirect estimates of HIV-1 incidence in a high prevalence population. American Journal of Epidemiology. 1998; 147:834–839. [PubMed: 9583713]
- Cox, DR.; Hinkley, DV. Theoretical Statistics. Chapman and Hall; London: 1974.
- Fang C, Field S, Busch M, Heyns A. Human immunodeficiency virus-1 and hepatitis C virus RNA among South African blood donors: estimation of residual transfusion risk and yield of nucleic acid testing. Vox Sanguinis. 2003; 85:9–19. [PubMed: 12823725]
- Hallet T, Zaba B, Todd J, Lopman B, Mwita W, Biraro S, Gregson S, Boerma J. Estimating incidence from prevalence in generalised hiv epidemics: Methods and validation. PLoS Medicine. 2008; 5:611–621.
- Hu D, Vanichseni S, Mock P, Young N, Dobbs T, Byers R, Choopanya K, Griensven F, Kitayaporn D, McDougal S, Tappero J, Mastro T, Parekh B. HIV type 1 incidence estimates by detection of recent infection from a cross-sectional sampling of injection drug users in Bangkok: use of the IgG capture BED enzyme immunoassay. AIDS Research and Human Retroviruses. 2003; 19:727–730. [PubMed: 14585202]
- Janssen R, Satten GA, Stramer S, Rawal B, O'Brien T, Weiblen B, Hecht F, Jack N, Cleghorn J, Kahn J, Chesney M, Busch M. New testing strategy to detect early HIV-1 infection for use in incidence

- estimates and for clinical and prevention purposes. *Journal of the American Medical Association*. 1998; 280:42–48. [PubMed: 9660362]
- Kaplan K, Brookmeyer R. Snapshot estimators of recent HIV incidence rates. *Operations Research*. 1999; 47:29–37.
- McDougal J, Parekh B, Peterson M, Branson B, Dobbs T, Ackers M, Gurwith M. Comparison of HIV type 1 incidence observed during longitudinal follow-up with incidence estimated by cross-sectional analysis using the BED capture enzyme immunoassay. *AIDS Research and Human Retroviruses*. 2006; 22:945–952. [PubMed: 17067263]
- Parekh B, Kennedy M, Dobbs T, Pau C, Byers R, Green T, Hu D, Vanichseni S, Young N, Choopanya K, Mastro T, McDougal S. Quantitative detection of increasing HIV type 1 antibodies after seroconversion: A simple assay for detecting recent HIV infection and estimating incidence. *AIDS Research and Human Retroviruses*. 2002; 18:295–307. [PubMed: 11860677]
- Podgor M, Leske M. Estimating incidence from age-specific prevalence for irreversible diseases with differential mortality. *Statistics in Medicine*. 1986; 5:573–578. [PubMed: 3823665]
- Saidel T, Sokal D, Rice J, Buzingo T, Hassig S. Validation of a method to estimate age-specific HIV incidence rates in developing countries using population-based seroprevalence data. *American Journal of Epidemiology*. 1996; 144:214–223. [PubMed: 8686690]
- Thior I, Lockman S, Smeaton L, Shapiro R, Wester C, Heymann S, Gilbert P, Stevens L, Peter T, Kim S, Widenfelt E, Moffat C, Ndase P, Arimi P, Kebaabetswe P, Mazonde P, Makhema J, McIntosh K, Novistky V, Lee T, Marlink R, Lagakos S, Essex M. Breastfeeding with infant zidovudine prophylaxis for 6 months versus formula feeding for reducing HIV transmission and infant mortality: A randomized trial in botswana. *Journal of the American Medical Association*. 2006; 296:794–805. [PubMed: 16905785]
- Williams B, Gouws E, Wilkinson D, Karim S. Estimating HIV incidence rates from age prevalence data in epidemic situations. *Statistics in Medicine*. 2001; 20:2003–2016. [PubMed: 11427956]
- Zou S, Dodd RY, Stramer SL, Strong M. Probability of viremia with HBV, HCV, HIV and HTLV among tissue donors in the united states. *NEJM*. 2004; 351:751–759. [PubMed: 15317888]

Figure 1.

(a) Variance of estimates of λ , when $\theta = 0.1$ and true incidence rate (per year) = 0.01. Solid line corresponds to the AED and AD batteries. The dotted line corresponds to the AE battery; Dashed line corresponds to the ED battery; (b) Expected sojourn times for the three test batteries that include the detuned ELISA test. Solid line depicts the expected sojourn time in State 2 ($E(Y_2 | t_0, \text{ in } S_2 \text{ at } t)$). The dashed line depicts the expected time since HIV infection conditional on the subject being in States 1 or 2 at time t ($E(t - T | t_0, \text{ in } S_1 \text{ or } S_2 \text{ at } t)$).

Figure 2.
MASHI Study - stratum specific and predicted incidence rates from logistic model by age and calendar time.

Figure 3. Debswana Study - stratum specific and predicted incidence rates from logistic model by age, gender and job type.

Table 1

Probabilities associated with 19 possible outcomes based on battery of tests used at calendar time t in a subject born at time t_0

Test Battery *	Test Result	Probability	State
A	(0, -1, -1)	$\pi_0(t t_0)$	uninfected
	(1, -1, -1)	$\pi_1(t t_0) + \pi_2(t t_0) + \pi_3(t t_0)$	acute, recent or nonrecent
E	(-1, 0, -1)	$\pi_0(t t_0) + \pi_1(t t_0)$	uninfected or acute
	(-1, 1, -1)	$\pi_2(t t_0) + \pi_3(t t_0)$	recent or nonrecent
D	(-1, -1, 0)	$\pi_0(t t_0) + \pi_1(t t_0) + \pi_2(t t_0)$	uninfected, acute or recent
	(-1, -1, 1)	$\pi_3(t t_0)$	nonrecent
A, E	(0, 0, -1)	$\pi_0(t t_0)$	uninfected
	(1, 0, -1)	$\pi_1(t t_0)$	acute
	(1, 1, -1)	$\pi_2(t t_0) + \pi_3(t t_0)$	recent or nonrecent
A, D	(0, -1, 0)	$\pi_0(t t_0)$	uninfected
	(1, -1, 0)	$\pi_1(t t_0) + \pi_2(t t_0)$	acute or recent
	(1, -1, 1)	$\pi_3(t t_0)$	nonrecent
E, D	(-1, 0, 0)	$\pi_0(t t_0) + \pi_1(t t_0)$	uninfected or acute
	(-1, 1, 0)	$\pi_2(t t_0)$	recent
	(-1, 1, 1)	$\pi_3(t t_0)$	nonrecent
A, E, D	(0, 0, 0)	$\pi_0(t t_0)$	uninfected
	(1, 0, 0)	$\pi_1(t t_0)$	acute
	(1, 1, 0)	$\pi_2(t t_0)$	recent
	(1, 1, 1)	$\pi_3(t t_0)$	nonrecent

* A=antigen test, E=ELISA, D=detuned ELISA

Table 2

Analytical expressions for the MLEs of f , θ and λ when all subjects are tested with the same battery of tests. n denotes the total number of subjects screened. Here $k_1 = E(L_1)$ and, if L_1 and L_2 are independent, $k_2 = E(L_2)$.

Battery of tests	θ	$\hat{f}$	$\hat{\lambda}$
A, E, D	$\frac{n_{100}+n_{110}+n_{111}}{n}$	$\frac{n_{100}+n_{110}}{(k_1+k_2)n}$	$\frac{n_{100}+n_{110}}{(k_1+k_2)n_{000}}$
A, E	$\frac{n_{10-}+n_{11-}}{n}$	$\frac{n_{10-}}{k_1 n}$	$\frac{n_{10-}}{k_1 n_{00-}}$
A, D	$\frac{n_{1-0}+n_{1-1}}{n}$	$\frac{n_{1-0}}{(k_1+k_2)n}$	$\frac{n_{1-0}}{(k_1+k_2)n_{0-0}}$
E, D	$\frac{k_2(n_{-10}+n_{-11})+k_1 n_{-10}}{k_2 n}$	$\frac{n_{-10}}{k_2 n}$	$\frac{n_{-10}}{k_2 n_{-00}-k_1 n_{-10}}$

Table 3

Results from logistic regression analysis

Covariate	(n ₀₀ , n ₁₁ , n ₁₀)	Odds Ratio	95% CI	p value Wald (LRT)
MASHI cohort				
Calendar time (days)				
Linear term		1.03	(1.009, 1.042)	(0.0007)
Quadratic term		0.99996	(0.99993, 0.99998)	0.003
Age (years)				
< 19	(391,49,9)	1	-	0.001
20–29	(997,378,56)	2.53	(1.24, 5.18)	(0.009)
30–39	(354,148,9)	1.11	(0.44, 2.84)	-
40+	(70,14,4)	2.41	(0.72, 8.13)	0.01
Debswana cohort				
Job type				
Contractor	(688,313,18)	1	-	0.13
Non contractor	(1849,445,26)	0.60	(0.31, 1.12)	(0.40)
Gender				
Female	(565,129,11)	1	-	-
Male	(1972,629,33)	0.73	(0.36, 1.49)	0.38
Age (years)				
< 19	(20,0,2)	1	-	(0.23)
20–29	(713,11,15)	0.25	(0.05, 1.17)	-
30–39	(717,352,14)	0.25	(0.05, 1.24)	0.08
40+	(1087,255,13)	0.16	(0.03, 0.82)	0.09
Integrated analysis				
Age (years)				
< 19	(411,49,11)	1	-	(0.06)
20–29	(1710,529,71)	1.99	(1.03, 3.82)	-
30–39	(1071,500,23)	1.28	(0.60, 2.73)	0.04
				0.52

Covariate	(n ₀₀ , n ₁₁ , n ₁₀)	Odds Ratio	95% CI	p value Wald (LRT)
40+	(1157, 269, 17)	1.23	(0.52, 2.91)	0.63 (0.002)
Calendar time				
Linear term		1.02	(1.01, 1.04)	0.006
Quadratic term		0.99997	(0.99994, 0.99999)	0.003 (0.37)
Gender				
Female	(2468, 718, 89)	1	-	-
Male	(1972, 629, 33)	0.71	(0.35, 1.45)	.34 (0.002)
Job type				
Contractor	(688, 313, 18)	1	-	-
Non contractor	(1849, 445, 26)	0.54	(0.28, 1.04)	0.06
Other (Mashii subjects)	(1812, 589, 78)	1.6×10^{-5}	$(1.2 \times 10^{-8}, 0.02)$	0.003