

## SNAPSHOT ESTIMATORS OF RECENT HIV INCIDENCE RATES

EDWARD H. KAPLAN

*Yale University School of Management, New Haven, Connecticut*

RON BROOKMEYER

*Johns Hopkins University, Baltimore, Maryland*

(Received March 1996; revisions received December 1996 and July 1997; accepted August 1997)

The recent HIV incidence rate (or hazard rate for infection) is an important quantity for use in monitoring the HIV/AIDS epidemic, evaluating HIV prevention programs, and allocating HIV prevention resources. Direct measurement of HIV incidence is difficult and time consuming, while estimating HIV incidence via backcalculation (deconvolution) using AIDS incidence data and the (presumed known) HIV incubation time distribution yields little information about recent infection. We propose a method for estimating recent HIV incidence in a population via a single sample at a single point in time. The method relies on understanding the progression of certain markers such as CD4 counts following infection. The actual formulas derived for the point and interval estimates of HIV incidence are simple and easy to use while the sample sizes needed to implement our approach are reasonable. We present two applications of our approach, comparing our results to those obtained from more conventional methods where possible.

Monitoring the epidemic of human immunodeficiency virus (HIV) infection, the virus responsible for the acquired immune deficiency syndrome (AIDS), is an important public health activity (Brookmeyer and Gail 1994, Kaplan and Brandeau 1994a). However, such monitoring is difficult because in the United States and most other countries HIV infection is not a reportable condition. Knowledge of the incidence rate (or hazard) for HIV is also important for evaluating the efficacy of HIV prevention programs such as counseling and testing, safer-sex education, aggressive treatment of sexually transmitted diseases, drug treatment, and needle exchange (Auerbach et al. 1995, Kaplan 1995a, Normand et al. 1995). The efficient allocation of HIV prevention resources also requires estimates of the rate of new HIV infections (Holtgrave and Qualls 1995, Kaplan 1995b, Kaplan and Brandeau 1994b, Valdiserri et al. 1995).

Various methods have been proposed for estimating the rate of new HIV infections. One approach has been to rely on cohort studies that repeatedly test uninfected individuals over time, enabling the direct computation of HIV incidence rates as the ratio of the number of new infections detected to the number of person years of exposure (Breslow and Day 1987, Brookmeyer and Gail 1994, Keet 1995, Nelson et al. 1995, Winkelstein et al. 1987). Such cohort studies are expensive and time consuming, and they are also subject to significant follow-up biases depending upon subject retention rates. While such studies have provided invaluable information pertaining to new HIV

infection rates in certain populations, the practical drawbacks of cohort studies have prevented widescale implementation, particularly for use in the evaluation of HIV prevention programs.

Another proposal has been to examine changes in HIV prevalence (the fraction of the population currently infected) over time via cross-sectional samples (Brookmeyer and Gail 1994, Centers for Disease Control 1990). In areas where HIV incidence is rapidly increasing from negligible levels, such comparisons do provide good insight into new infection rates because early in an epidemic, population losses as a result of disease are minimal. However, in areas where the epidemic is more mature, changes in HIV prevalence provide little information regarding the rate of new infections because population losses as a result of AIDS are substantial. In the extreme case where HIV prevalence reaches a steady state (and new infections are cancelled by deaths in the population), differences in cross-sectional prevalence surveys approach zero irrespective of the underlying rate of new HIV infections.

Taking advantage of readily available data reporting AIDS (as opposed to HIV) incidence, along with natural history studies that estimate the probability distribution of the *incubation time* from HIV infection to the development of AIDS, the method of *backcalculation* has been employed to reconstruct historical HIV infection rates from reported AIDS incidence over time (Brookmeyer 1991; Brookmeyer and Gail 1988, 1994). Denoting  $a(t)$  as the expected number of AIDS cases that occur at time  $t$ ,

$h(t)$  as the expected number of new HIV infections that occur at time  $t$ , and  $f_I(t)$  as the probability density for the incubation time from infection until AIDS, the simplest backcalculation model dictates that

$$a(t) = \int_{-\infty}^t h(u) f_I(t-u) du. \quad (1)$$

More sophisticated backcalculation models allow the incubation time density to depend upon calendar time to allow for the treatment effects of AIDS therapy that have served to lengthen the time between infection and AIDS.

To estimate the historical incidence rates, one assumes some flexible form such as a step or spline function for  $h(t)$ , treats observed AIDS incidence data over time as realizations from some random process (e.g., Poisson) with mean value function given by Equation (1), and then estimates the parameters of  $h(t)$  using maximum likelihood or some other estimation technique. Unfortunately, because the probability of progressing to AIDS within the first few years following infection is very low, current AIDS data contain very little information regarding recent (as opposed to past) infection rates. The statistical uncertainty associated with recent HIV infection rates estimated from backcalculation is sufficiently large to render them unreliable for monitoring and evaluation purposes.

A different (and newer) set of approaches for estimating HIV incidence involves the study of immunological markers for HIV infection. Brookmeyer and Quinn (1995) presented a simple model based on the fact that following infection with the HIV virus, there is a “window period” during which HIV antibodies cannot be detected, but other markers of infection such as p24 antigen or viral culture would yield positive results. Focusing on p24 antigen, Brookmeyer and Quinn estimated that there was approximately a 22.5-day window (with a 95% confidence interval from 13–42 days) during which p24 antigen could be detected but HIV antibody could not be detected in HIV infected persons. Given the short duration of this period relative to changes in the rate of new HIV infections, the fraction of HIV antibody negative persons found to be p24 positive in a sample can be expressed as 22.5 days times the HIV incidence rate at the time of sampling, enabling estimation of the current incidence rate. Unfortunately, unless the HIV incidence rate is large, the sample sizes required for reasonably precise estimates are very large owing to the brevity of the window. An extension embedding this method in a cohort study that does not require an a priori estimate of the window duration and automatically adjusts for follow-up bias is reported in Brookmeyer et al. (1995).

Another approach combining ideas from backcalculation with markers for infection is reported by Satten and Longini (1994). They model immunological markers such as CD4 counts as functions of time since infection among HIV infected persons. Letting  $e(m|t)$  denote the expected

number of persons with a marker value of  $m$  in a random sample of antibody positive persons from the population at time  $t$ ,  $f_{M(t)}(m)$  denote the probability density for the marker measured  $t$  time units after infection, and  $h(t)$  represent the new HIV infection rate in the *sample* at time  $t$ , one obtains

$$e(m|t) = \int_{-\infty}^t h(u) f_{M(t-u)}(m) du, \quad (2)$$

which allows for statistical estimation of  $h(t)$  in a manner analogous to backcalculation. Satten and Longini consider a discrete state Markov model for the progression of CD4 counts, and they apply variations of Equation (2) to data collected among gay and bisexual men in San Francisco. An advantage of this approach is that one can obtain (scaled) estimates of HIV infection rates into the past. A disadvantage is that the entire marker progression process  $M(t)$  must be modeled, a difficult task given the enormous variability exhibited by immunological markers of HIV infection.

This paper presents a new approach to estimating recent HIV incidence. We retain the advantages associated with estimating HIV incidence from a single sample at a single point (snapshot) in time via very simple formulas. The idea is to construct an estimate based on easily observed markers of infection among already HIV antibody positive persons as suggested by Satten and Longini while avoiding the need to model the entire marker progression process, though our method can take advantage of such models when they exist. To do this, we construct an analogy to Brookmeyer and Quinn’s method by designing target regions for the markers of interest that are reflective of recent infection (testing positive for p24 antigen but negative for HIV antibody defines such a target region in Brookmeyer and Quinn’s approach). The duration of time spent by markers in our target regions will be on the order of 1 to 2 years, which greatly reduces the sample size needed to generate accurate estimates (note that the time from infection until the development of clinical AIDS is on the order of 10 years; Brookmeyer and Gail 1994). However, rather than obtaining estimates of current HIV incidence, our snapshot samples will produce estimates of incidence in the recent past. Our method does require obtaining a representative sample of the population in question, a challenging task in itself given the difficulty of establishing proper sampling frames for populations such as injecting drug users or sexually active gay men. However, this same sampling challenge confronts cohort models for incidence estimation. Given the ease of implementation of our new method, we believe this approach offers a pragmatic alternative to existing methods for use in monitoring, program evaluation and resource allocation studies.

Our paper proceeds as follows. In the next section, we outline our general approach to constructing snapshot incidence estimators, discuss their salient properties, and

present the point and interval estimates associated with this method. We illustrate our methods in §3 with two applications in different populations at risk for HIV infection. A summary and suggestions for future research appear in §4.

## 1. CONSTRUCTING SNAPSHOT ESTIMATORS: GENERAL APPROACH

### 1.1. The Marker Process and the Target Region

We assume the existence of a (possibly vector valued) marker  $M(t)$  where  $t$  denotes time from HIV infection. Possible markers could include CD4 count, percent CD4, platelet count, or combinations of these (Brookmeyer and Gail 1994). We now define  $\mathfrak{R}$  as the *target region* for the marker, and let

$$\phi(t) = \Pr\{M(t) \in \mathfrak{R}\} \quad (3)$$

denote the probability that the marker is in the target region at time  $t$  following HIV infection. The target region  $\mathfrak{R}$  is designed to correspond with marker levels indicative of recent infection. Possible target regions could include CD4 count of at least 900 cells/ $\mu\text{l}$ , percent CD4 greater than 35%, or combinations such as percent CD4 > 35% and platelet count at least 1200, in addition to HIV-positive infection status. We do *not* allow target regions to include an AIDS diagnosis, but this is not a limitation because the target region is intended to reflect recent infection. In the ensuing discussion, we assume that  $\phi(t)$  is known.

### 1.2. Snapshot Samples under Constant Incidence

Now imagine observing the HIV epidemic at an arbitrary moment in time. Of all those who have become infected in the past, what is the expected number with markers that are in the target region at the time our *snapshot* is taken? Let  $\lambda(t)$  denote the expected number of new infections in the population per unit time measured  $t$  time units *prior* to our snapshot—we refer to  $\lambda(t)$  as the infection rate—and let  $L$  denote the number of infected persons with markers in the target region at the time of the snapshot. Clearly,

$$E(L) = \int_0^{\infty} \lambda(t) \phi(t) dt, \quad (4)$$

for persons infected  $t$  time units ago will appear in the target region with probability  $\phi(t)$  (note that Equation (4) requires  $\phi(t)$  to be invariant with respect to the calendar date of infection). If the historical infection rate  $\lambda(t)$  is roughly constant and equal to  $\lambda$ , Equation (4) simplifies to

$$E(L) = \lambda \tau, \quad (5)$$

where

$$\tau = \int_0^{\infty} \phi(t) dt \quad (6)$$

is the expected *total* time spent by the marker in the target region prior to an AIDS diagnosis. Note that it is possible

for markers to exit and reenter the target region; all we require is that at any time following infection, the probability of being in the target region is known. However, in the special case where the marker spends  $T$  time units in the target region and never returns, then  $\phi(t) = \Pr\{T > t\}$ ,  $\tau = E(T)$ , and Equation (5) simplifies to Little's Theorem (Little 1961).

From Equation (5), the aggregate infection rate could thus be found from the ratio of  $E(L)/\tau$ , and an estimate of  $\lambda$  would follow from inserting the *observed* number of infected persons with markers in the target region in place of  $E(L)$ . Of course, one cannot observe all infected persons, and in addition, the quantity of greatest interest to epidemiologists and HIV prevention planners is the *incidence* (or hazard) of infection, defined as the expected number of new infections per uninfected person per unit time.

Addressing this latter point first, let  $g$  denote the per capita infection rate in the population (that is,  $g$  equals  $\lambda$  divided by the size of the population), and let  $p$  denote the probability that a randomly selected member of the population is infected with HIV (so  $p$  is the *prevalence* of HIV). The incidence of infection  $\iota$  is then defined by

$$\iota = \frac{g}{1-p}. \quad (7)$$

To estimate  $\iota$ , let  $\psi$  denote the probability that a randomly chosen person has a marker in the target region (and thus  $\psi$  equals  $E(L)$  divided by the population size). We can decompose  $\psi$  as

$$\psi = p \pi, \quad (8)$$

where  $\pi$  is the conditional probability that a person has a marker value in the target region, given that (s)he is HIV infected. With these definitions, we may divide both sides of Equation (5) by the size of the population to obtain

$$\psi = p \pi = g \tau. \quad (9)$$

The incidence of infection  $\iota$  is then given by

$$\iota = \frac{g}{1-p} = \frac{p}{1-p} \frac{\pi}{\tau}. \quad (10)$$

To use this equation, one would take a *snapshot sample* at an arbitrary moment in time, and estimate both  $p$  and  $\pi$  by their sample analogs (assuming that  $\tau$  is known). Even if the infection rate is not constant over time, Equation (10) will be approximately correct providing the historical infection rate  $\lambda(t)$  (and hence  $g(t)$ ) is roughly constant over those time periods where  $\phi(t)$  is appreciably greater than zero. This condition is plausible if  $\tau$  is sufficiently small.

### 1.3. Snapshot Samples under Changing Incidence: General Properties

The formula we have just derived estimates HIV incidence when such incidence is unchanging. While this is not an unreasonable assumption in some circumstances, in many at-risk groups HIV incidence is changing over time and the

constant incidence assumption is inappropriate (Brookmeyer and Gail 1994, Mann et al. 1992). However, the simplicity of Equation (10) invites a closer look. Again letting  $t$  be the time prior to the snapshot sample, suppose that the historical infection rate  $\lambda(t)$  (and hence the *per capita* infection rate  $g(t)$  found from normalizing  $\lambda(t)$  by the population size *at the time of sampling*) is arbitrary. By analogy with Equation (10), we now *define* the snapshot incidence rate to be  $\iota = p\pi/(1-p)\tau$  where  $p$  and  $\pi$  are the HIV prevalence and conditional probability of being in the target region given a positive HIV antibody test at the time of sampling. However, when HIV incidence is changing over time, the snapshot incidence rate  $\iota$  is not equal to the instantaneous incidence rate at the time of the snapshot sample; this latter rate equals  $g(0)/(1-p)$ . We now show that the snapshot incidence rate approximates the true incidence rate in the recent past.

As in Equation (9) the probability  $\psi$  that a person in the population at the time of a snapshot sample is both HIV infected and has a marker in the target region is given by

$$\psi = p\pi = \int_0^{\infty} g(t)\phi(t) dt. \quad (11)$$

Define

$$f_S(t) = \frac{\phi(t)}{\tau} = \frac{\phi(t)}{\int_0^{\infty} \phi(u) du} \quad \text{for } t > 0 \quad (12)$$

as the probability density of an important random variable we denote by  $S$  and discuss below. The snapshot incidence rate  $\iota$  then becomes

$$\iota = \frac{\int_0^{\infty} g(t)\phi(t) dt}{(1-p)\tau} = \frac{1}{1-p} \int_0^{\infty} g(t)f_S(t) dt = \frac{E[g(S)]}{1-p}, \quad (13)$$

which is the expected value of the per capita infection rate at  $S$  time units *before* sampling, scaled by the fraction of the population that is uninfected *at the time* of sampling.

To interpret  $\iota$ , consider the second order Taylor series expansion of the per capita incidence rate function  $g(t)$  about  $\mu_S = E(S)$ . Equation (13) then yields

$$\begin{aligned} \iota &\approx \frac{1}{1-p} \int_0^{\infty} \left( g(\mu_S) + g'(\mu_S)(t - \mu_S) \right. \\ &\quad \left. + \frac{g''(\mu_S)}{2} (t - \mu_S)^2 \right) f_S(t) dt \quad (14) \\ &= \frac{1}{1-p} \left( g(\mu_S) + \frac{g''(\mu_S)}{2} \sigma_S^2 \right), \end{aligned}$$

where  $\sigma_S^2 = \text{Var}(S)$ . Now, the incidence of infection  $\mu_S$  time units prior to the date of the snapshot sample is given by

$$\iota(\mu_S) = \frac{g(\mu_S)}{1-p(\mu_S)}, \quad (15)$$

where  $p(\mu_S)$  denotes the prevalence of HIV infection in the population  $\mu_S$  time units into the past. Comparing Equations (14) and (15), we see that if  $p(\mu_S) \approx p$  (a reasonable assumption if  $\mu_S$  is suitably small), the difference between the snapshot sample incidence rate  $\iota$  and the actual incidence  $\mu_S$  time units ago approximately equals

$$\Delta \iota = \frac{1}{1-p} \frac{g''(\mu_S)}{2} \sigma_S^2. \quad (16)$$

Note that if the historical per capita infection rate  $g(t)$  is linear in time, or if  $\sigma_S^2$  is small, then  $E[g(S)] \approx g[E(S)] = g(\mu_S)$ , and Equation (16) approaches zero.

These features provide an interpretation for our snapshot incidence estimator:  $\iota$  is approximately the HIV incidence rate  $\mu_S$  time units prior to sampling. Since the random variable  $S$  dictates the extent to which  $\iota$  reaches into the past, we refer to  $S$  as the *shadow* cast by the snapshot sample. Note that in the special case where the marker spends  $T$  time units in the target region and never returns,  $S$  corresponds to the backward recurrence time in an equilibrium renewal process with interarrival times distributed as  $T$ , and thus  $\mu_S = E(T^2)/2E(T)$  and  $\sigma_S^2 = E(T^3)/3E(T) - \mu_S^2$  as is well known (Cox 1962).

#### 1.4. Bias in Snapshot Incidence Rates

The validity of interpreting  $\iota$  as the incidence of infection  $\mu_S$  time units before sampling depends upon the size of the estimation bias. The absolute relative error (ARE) from estimating  $\iota(\mu_S)$  by  $\iota$  is given by

$$\text{ARE} = \left| \frac{\iota - \iota(\mu_S)}{\iota(\mu_S)} \right| \approx \left| \frac{E(g(S)) - g(\mu_S)}{g(\mu_S)} \right|, \quad (17)$$

with the approximation following from  $p \approx p(\mu_S)$ . First, suppose one is willing to assume that the per capita infection rate  $g(t)$  follows the quadratic

$$g(t) = g(\mu_S) \left( 1 \pm r(t - \mu_S) \pm \frac{1}{2} r^2 (t - \mu_S)^2 \right) \quad (18)$$

for some  $r > 0$ . Equation (17) implies that the ARE in our estimate equals  $r^2 \sigma_S^2 / 2$ . Given an acceptable relative error tolerance (e.g., 20%) along with the value of  $\sigma_S^2$ , one can determine the smallest value of  $r$  that would lead to an unacceptably large error and judge whether such a value of  $r$  is plausible given current knowledge. If the implied value of  $r$  is believed to be improbably large, then one can feel reasonably comfortable using the snapshot incidence rate to estimate the incidence of infection  $\mu_S$  time units earlier.

Of course,  $g(t)$  might not follow a quadratic curve. Consider an exponential model

$$g(t) = g(\mu_S) e^{r(t - \mu_S)} \quad \text{for } r, t > 0. \quad (19)$$

Then the ARE from Equation (17) is given by

$$\text{ARE} = E(e^{r(S - \mu_S)}) - 1. \quad (20)$$

As an example, if the shadow  $S$  itself is exponentially distributed with mean  $\mu_S$  (as in Section 3.1 below), then

$$\text{ARE} = \frac{e^{-r\mu_S}}{1 - r\mu_S} - 1. \quad (21)$$

For values of  $r\mu_S \leq 0.49$ , the ARE is less than 20%.

One should therefore exercise caution when employing snapshot incidence estimates. Sensitivity analyses of the form illustrated above provide one approach to judging the reasonability of the results.

### 1.5. Point and Interval Snapshot Incidence Estimates

We now turn briefly to statistical estimation issues. The snapshot incidence estimate is given by

$$\hat{i} = \frac{\hat{p}}{1 - \hat{p}} \frac{\hat{\pi}}{\hat{\tau}}, \quad (22)$$

where  $\hat{p}$  is the observed proportion in the sample who are HIV infected ( $\hat{p} = n^+/n$  where  $n^+$  is the observed number of individuals seropositive for HIV antibodies out of the  $n$  who are tested),  $\hat{\pi}$  is the observed proportion of HIV infected individuals with markers in the target region  $\mathfrak{R}$  ( $\hat{\pi} = n_{\mathfrak{R}}/n_s$  where  $n_s$  is the number of HIV infected individuals who are screened for the marker, and  $n_{\mathfrak{R}}$  is the number of those screened who are found with markers in the target region), and  $\hat{\tau}$  is the estimated mean time spent by the marker in the target region. Taking logarithms we have

$$\log(\hat{i}) = \log\left(\frac{\hat{p}}{1 - \hat{p}}\right) + \log(\hat{\pi}) - \log(\hat{\tau}). \quad (23)$$

We compute  $\nu$ , the estimated sampling variance of  $\log(\hat{i})$ , conditional on two quantities: (i) the snapshot sample size  $n$ ; and (ii) the observed fraction of HIV infecteds who are screened for the marker,  $n_s/n^+$ , which is an ancillary statistic. Application of the delta method (Bishop et al. 1975) to Equation (23) yields

$$\begin{aligned} \nu &= \frac{1}{(1 - \hat{p})\hat{p}n} + \frac{1 - \hat{\pi}}{\hat{\pi}n_s} + \frac{\widehat{\text{var}}(\hat{\tau})}{\hat{\tau}^2} \\ &= \frac{n}{n^+(n - n^+)} + \frac{n_s - n_{\mathfrak{R}}}{n_s n_{\mathfrak{R}}} + \frac{\widehat{\text{var}}(\hat{\tau})}{\hat{\tau}^2}, \end{aligned} \quad (24)$$

where  $\widehat{\text{var}}(\hat{\tau})$  is the estimated variance of  $\hat{\tau}$  as derived from marker studies. An asymptotic  $100 \times (1 - \alpha)\%$  confidence interval for  $\hat{i}$  (with  $n$  and  $n_s$  large) is then given by the interval

$$[\hat{i}e^{-z_{\alpha/2}\sqrt{\nu}}, \hat{i}e^{z_{\alpha/2}\sqrt{\nu}}], \quad (25)$$

where  $z_{\alpha/2}$  is the  $\alpha/2$  critical value of the standard normal distribution. This confidence interval procedure does not assume that all HIV infected individuals are tested for the marker, but rather that some predetermined arbitrary fraction of infected persons are evaluated. This may be important if, for example, the assay for the marker is expensive. The methods described here do assume that those tested for the marker are a random sample of all HIV infected individuals in the snapshot sample.

Equation (24) highlights several practical points. First, the variance  $\nu$  declines as the fraction of infected persons with markers in the target region,  $\pi$ , increases. Other things being equal,  $\pi$  is increasing in the rate of new infections. In situations of low HIV incidence, then, the precision of a snapshot estimate will suffer, and very large samples are required to obtain narrow confidence bands. Of course, this is also true of direct cohort incidence estimates: The total person-years of exposure must be large to reliably estimate low incidence levels. Second,  $\pi$  is increasing in  $\tau$  (for Equation (11) implies that  $p\pi = E(g(S))\tau$ ). Larger values of  $\tau$  thus provide greater precision in the snapshot incidence estimate. Third,  $\nu$  becomes very large as the HIV prevalence  $p$  approaches 0 or 1. In both of these situations, it is difficult to estimate the log odds of infection ( $\log(p/(1 - p))$ ). Finally,  $\nu$  grows as uncertainty about  $\tau$  increases, reflecting the importance of good marker studies in producing snapshot incidence estimates.

### 1.6. Considerations for the Design of Target Regions

The discussion to this point has taken the target region  $\mathfrak{R}$  and its associated  $\phi(t)$  function as given. If raw data describing the progression of one or more markers over time from infection are available for study, however, then how should one design a target region? We propose three criteria for selecting a target region from a list of alternatives based on the preceding analysis. First, the mean shadow  $\mu_S$  must be small enough to ensure the relevance of whatever estimate is obtained. If  $\mu_S$  is too large (say, greater than 2 years), then the incidence estimate is too dated to be useful. Second, plausible models for the size of the estimation bias (as discussed in §2.4 above) should lead to acceptable relative errors. Subject to satisfying these two conditions, the third criterion is to make  $\tau$  large, because doing so increases the value of  $\pi$ , which reduces the variance of the estimated incidence rate as discussed above. One way to operationalize this procedure is to consider target regions of the form  $\mathfrak{R}_m \equiv \{M : M > m\}$ , and then to consider our criteria for different values of  $m$ . This approach is illustrated in §3.2.

## 2. APPLICATIONS

### 2.1. Gay/Bisexual Men in San Francisco

The San Francisco Men's Health Study (Winkelstein et al. 1987) followed a probability sample of men aged 25 through 55 from those parts of San Francisco most affected by the AIDS epidemic. Recruitment began in June 1984, leading to the enrollment of 1,045 men. Data from these subjects were collected in waves every six months over an eight-year period. Of interest here are data reporting HIV infection and CD4 counts among gay and bisexual men in the study.

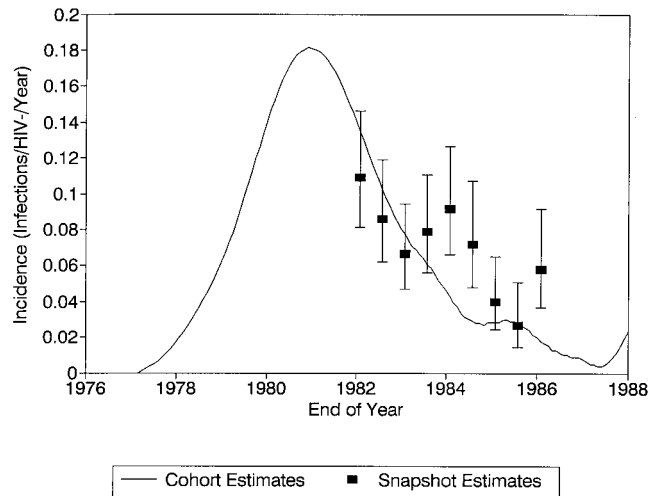
To implement our snapshot incidence estimator, we consider CD4 count as the marker of interest, and define the target region  $\mathfrak{R}$  to be CD4 count  $\geq 900$  cells/ $\mu$ l. A

discrete state Markov model for CD4 counts has been constructed by Longini et al. (1991) by partitioning continuous CD4 measurements into six CD4 regions, with a seventh state reserved for an AIDS diagnosis and an eighth state for death. The model assumes that upon infection, one enters the initial state corresponding to CD4 count  $\geq 900$ , and then progresses sequentially through the other five CD4 states to AIDS and death. Thus, in this model, once the marker leaves the target region, it never returns.

Given the Markov assumption, the total time spent in the target region is exponentially distributed. Using data from the San Francisco Men's Health Study (SFMHS), Longini et al. estimated that the mean time until CD4 drops below 900 equals 1.73 years (with a standard error of 0.2 years), which implies that the expected total time spent in the target region ( $\tau$ ), the expected value of the shadow ( $\mu_S$ ), and the standard deviation of the shadow ( $\sigma_S$ ) equal 1.73 years due to the Markov assumption. Assuming exponential decline in the rate of new infections over time coupled with a willingness to tolerate relative errors of 20%, Equation (21) suggests that this model should function well providing the exponential rate  $r$  does not exceed  $0.49/1.73 = 0.28 \text{ yr}^{-1}$ . If the rate of new infections followed the quadratic model of Equation (18), then the ARE would fall within 20% providing  $r \leq 0.37 \text{ yr}^{-1}$ . A snapshot incidence estimate under these circumstances should thus provide good information regarding the HIV incidence rate about 21 months prior to sampling.

From codebooks of the SFMHS (Survey Research Center 1988–1990) we have obtained data describing the prevalence of HIV among gay and bisexual men involved in the study for the first nine waves of data collection. Also, Satten and Longini (1994) report the number of gay and bisexual men in the SFMHS observed with CD4 counts corresponding to the states of the Markov model over time; we focus our attention on those with CD4  $\geq 900$ .

Given these data and Equations (22) through (25), we have computed the point and 95% confidence interval estimates for HIV incidence among gay and bisexual men in those San Francisco neighborhoods represented by the SFMHS. These estimates are plotted against the date of sample construction less 21 months to account for the expected value of the shadow. Also shown in Figure 1 are cohort-based estimates of HIV infection over time among gay and bisexual men in San Francisco as modeled by Bacchetti (1990). Figure 1 shows that our snapshot estimates broadly conform to the incidence pattern estimated directly from seroconversion data over time; six of our nine confidence intervals include the cohort-based incidence figures, while in the remaining three instances our estimates are on the high side. Possible reasons for such disagreement include nonrandom sampling of CD4 counts among men in the SFMHS (e.g., undersampling of men who developed AIDS in the study), underestimation of  $\tau$  due to chance, bias in the snapshot estimates, or statistical error in the cohort incidence estimates themselves. Further, the model of Longini et al. that produced  $\tau = 1.73$



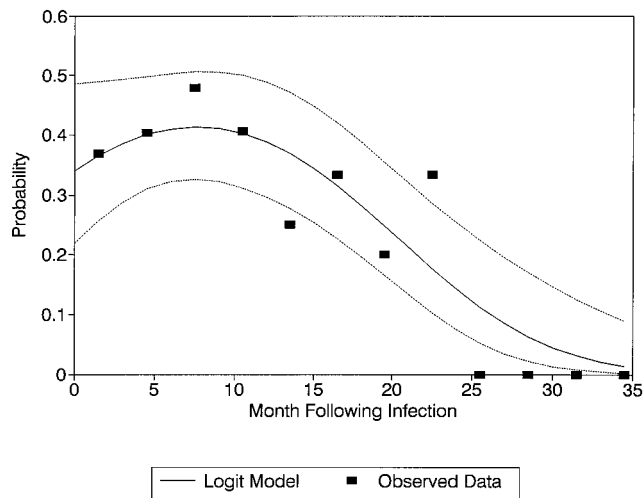
**Figure 1.** Annual incidence of infection among gay/bisexual men in San Francisco. (The snapshot estimates are shown together with the 95% confidence intervals (vertical bars), along with the cohort-based estimates reported by Bacchetti (1990).)

years assumes that all newly infected individuals begin with CD4 counts in the target region (CD4  $\geq 900$ ). However, it is possible that a proportion of newly infected individuals may have CD4 counts  $< 900$ . Further research on the natural history of immunological markers will refine such models and the resulting estimates of  $\tau$ . Nonetheless, the conformance we have achieved appears satisfactory when the ease with which our estimates were constructed is contrasted against the time and difficulty of mounting a large cohort study.

## 2.2. Drug Injectors in Baltimore

As part of a study of the maturity of HIV infection among drug injectors in Baltimore and the Bronx, Alcabes et al. (1994) collected data reporting HIV antibody status and two markers among drug injectors participating in ongoing cohort studies. For that subset of subjects who became infected during the course of these studies, data are available at various points following the estimated time of infection (which is midway between the last negative and first positive HIV antibody test). The marker we consider here is percent-CD4 (%CD4), defined as the fraction of lymphocytes with CD4+ T cell receptors (platelet counts, the other marker measured, reflect weak movement with time since infection relative to %CD4). We defined our target region  $\mathfrak{R}$  as %CD4  $> 35\%$ , and examined the available data provided by Dr. Alcabes (which covered 253 visits by 109 different drug injectors).

A logistic model for  $\phi(t) = \Pr\{\%CD4 > 35\% \text{ at time } t \text{ following infection}\}$  was fit to these data using ordinary maximum likelihood estimation, though generalized estimating equations to account for multiple visits by the same



**Figure 2.** The estimated logistic  $\phi(t)$  function corresponding to the target region  $\%CD4 > 35\%$  for drug injectors in Baltimore and the Bronx studied by Alcabes et al. (1994), along with 95% confidence intervals. (Also shown is the observed fraction of subjects with  $\%CD4 > 35\%$  grouped by three-month intervals.)

individuals (Zeger and Liang 1986) could also be employed. As shown in Figure 2, this model fits the data rather well. The estimated equation is given by

$$\phi(t) = 1/(1 + \exp(0.6649 - 0.0946t + 0.0061t^2)), \quad (26)$$

with  $t$  measured in months. From this model, we deduce that  $\tau = 8.72$  months (with a standard error of 0.99 months),  $\mu_S = 11.65$  months, and  $\sigma_S = 7.46$  months. Equation (17) suggests that the relative error corresponding to this model should be within 20% providing  $r \leq \sqrt{0.2 \times 2/(7.46/12)^2} = 1.017 \text{ yr}^{-1}$  in the quadratic epidemic model of Equation (18), a rather large rate. Alternatively, Equation (20) suggests that for the exponential epidemic model of Equation (19), the ARE is within 20% providing  $r \leq 0.926 \text{ yr}^{-1}$ , which is also large. Snapshot incidence estimates using  $\%CD4 > 35\%$  as a target region should thus produce estimates that correspond to HIV incidence roughly one year in the past.

To estimate HIV incidence, we turn to the AIDS Link to the Intravenous Experience (ALIVE) Study, an ongoing cohort study of drug injectors in Baltimore (and the source of some of the  $\%CD4$  data to which Equation (26) was fit). This study recruited 2,960 drug injectors between February 1988 and March of 1989. Direct estimates of HIV incidence in this cohort based on the ratio of new infections to person-years of exposure yielded an annual incidence rate of 4.25% from mid-1988 through mid-1989 (Nelson et al. 1995).

Of the 2,960 subjects initially recruited into this study, 2,247 were HIV antibody negative at enrollment. However, 31.8% of these seronegative subjects dropped out of the study over time, rendering follow-up bias a serious

concern in considering the incidence estimate reported above. For example, if those at greatest risk of infection are also those least likely to remain in the study, then incidence rates based on the cohort study will be biased downwards. Also, those who continued to participate in the study received counseling in concert with HIV antibody test results which could also have deflated HIV incidence below pre-study levels.

Given that 713 of the 2,960 subjects tested positive upon enrollment in the study, we estimate HIV prevalence as  $713/2,960 = 0.24$  among Baltimore drug injectors in mid-1988. Of these 713, 634 were screened for  $\%CD4$ , and 140 of these 634 reported  $\%CD4$  values greater than 35%. Using our derived value of  $\tau = 8.72$  months, Equations (22) through (25) produce an estimated incidence rate of 9.6% with a 95% confidence interval ranging from 7.2% through 12.7%. We earlier estimated that the expected shadow equals 11.65 months, suggesting that the HIV incidence rate among Baltimore drug injectors was much higher in mid-1987 than as estimated from the ALIVE cohort in the following year.

As discussed in §2.6, it is possible to perform an additional sensitivity analysis in this example, as the raw data enable the estimation of  $\phi(t)$  functions for different target regions. Analogous to Equation (26), we fit quadratic logistic curves to target regions defined by  $\mathfrak{R}_m \equiv \%CD4 > m$  for  $m = 30, 31, \dots, 40$ . Table I reports the resulting values of  $\tau$ ,  $\mu_S$ ,  $\sigma_S$ ,  $n_{\mathfrak{R}}$ , and  $\hat{i}$  for each target region. First, note the almost four-fold variation in  $\tau$  from 15.5 months through 4.2 months as  $m$  increases from 30 to 40. Second, note the relative stability of both  $\mu_S$  ( $\approx 12$  months) and  $\sigma_S$  ( $\approx 8$  months). This suggests two things. First, the precision of the incidence estimate can be improved by choosing a  $\%CD4$  cutoff of 30 without compromising the relevance or the bias of the estimate. Second, if our theory is correct, the numerical values of the incidence estimates should be quite close given the common values of  $\mu_S$  and  $\sigma_S$ . Table I reveals that this is in fact the case: The incidence estimates range from 9.1% through 10.5%. Note that this result required a four-fold drop in  $n_{\mathfrak{R}}$ , the number of initial HIV positives with markers in the target region, to correspond to the four-fold drop in  $\tau$  as  $m$  increased from 30 to 40. That the ratios of  $n_{\mathfrak{R}}$  to  $\tau$  remain roughly constant is an important validation of our theory, given that these quantities were derived from two different data sets (initial HIV positives and initial HIV negatives, respectively).

That our snapshot incidence estimates are roughly twice as high as the cohort estimate is understandable given that our estimate is of incidence in 1987—before the ALIVE study began—and given the high dropout rate from the ALIVE study that could be a significant source of follow-up bias. However, it may also be that while the ALIVE cohort was formed for the purpose of studying the natural course of HIV infection in the population of drug injectors, this cohort study itself may have contributed to reducing the rate of new HIV infections via ongoing counseling of study participants.

**Table I**  
Target Region Characteristics and Snapshot Incidence Estimates for Drug Injectors in Baltimore

$m(\mathfrak{R} \equiv \%CD4 > m)$	$\tau$ (months)	$\mu_S$ (months)	$\sigma_S$ (months)	$n_{\mathfrak{M}}$	$\hat{i}$ (% yr <sup>-1</sup> )	95% Confidence Interval
30	15.5	13.9	9.0	244	9.4	7.6–11.7
31	14.5	14.0	9.0	229	9.4	7.5–11.9
32	13.6	13.8	9.1	208	9.1	7.2–11.5
33	11.6	12.3	7.8	189	9.8	7.7–12.5
34	9.7	12.2	7.7	167	10.3	7.9–13.4
35	8.7	11.7	7.5	140	9.6	7.2–12.7
36	7.0	11.6	7.1	120	10.2	7.5–13.9
37	5.8	11.0	6.8	103	10.5	7.5–14.7
38	5.2	11.5	7.4	87	10.1	7.0–14.6
39	4.6	11.6	7.8	78	10.2	6.8–15.3
40	4.2	12.3	7.8	68	9.8	6.4–15.1

### 3. SUMMARY AND FUTURE RESEARCH

We have presented a new method for estimating recent HIV incidence via a single sample. The virtue of our approach lies in its ease of implementation, enabling the rapid determination of recent HIV infection rates. The drawbacks are that we obtain an estimate of HIV incidence lagged into the past as opposed to the current rate of new infections, and the method is susceptible to bias depending upon the underlying shape of the true epidemic curve.

While we believe the method can already be used in concert with existing marker studies as illustrated by our two applications, we also believe that some fine tuning is possible. Although our examples have involved immunological markers of HIV, our methods are applicable to the newer virological markers measured by polymerase chain reaction. The employment of multiple markers of recent infection might serve to sharpen our  $\phi(t)$  functions and reduce both  $\mu_S$  and  $\sigma_S^2$  without reducing  $\tau$ . However, whatever gains in accuracy could be achieved in this manner would need to be balanced against the cost increases entailed by observing multiple markers. Finally, though motivated by the need to estimate HIV incidence, our method should be applicable to incidence estimation for any infectious disease where the time from infection until the appearance of symptoms is relatively long and markers of disease progression are available.

Empirical estimates of recent HIV incidence rates remain elusive in spite of their unquestionable importance and value for monitoring the AIDS epidemic, evaluating HIV prevention programs, and efficiently allocating resources. We believe that the methods offered herein provide a pragmatic approach toward closing this informational gap.

### ACKNOWLEDGMENT

The authors thank Dr. Peter Bacchetti for the HIV incidence curve employed in Section 2.1, and Dr. Philip Alcades for the %CD4 data employed in §2.2. This research was supported in part by the Societal Institute for the Mathematical Sciences via Grant DA09531 from the

National Institute on Drug Abuse, and the Lady Davis Fellowship Trust, Jerusalem, Israel.

### REFERENCES

- Alcades, P., A. Muñoz, D. Vlahov, G. Friedland. 1994. Maturity of human immunodeficiency virus infection and incubation period of acquired immunodeficiency syndrome in injecting drug users. *AEP* 4 17–26.
- Auerbach, J. D., C. Wypijewska, H. K. H. Brodie (Eds.). 1995. *AIDS and Behavior: An Integrated Approach*. National Academy Press, Washington, DC.
- Bacchetti, P. 1990. Estimating the incubation period of AIDS by comparing population infection and diagnosis patterns. *JASA* 85 1002–1008.
- Bishop, Y. M. M., S. E. Fienberg, P. W. Holland. 1975. *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA.
- Breslow, N. E., N. E. Day (Eds.). 1987. *Statistical Methods in Cancer Research, Volume 2: Design and Analysis of Cohort Studies*. International Agency for Research on Cancer, Lyon, France.
- Brookmeyer, R. 1991. Reconstruction and future trends of the AIDS epidemic in the United States. *Science* 253 37–42.
- , M. H. Gail. 1988. A method for obtaining short-term projections and lower bounds on the size of the AIDS epidemic. *JASA* 83 301–308.
- , —. 1994. *AIDS Epidemiology: A Quantitative Approach*. Oxford University Press, Oxford.
- , T. C. Quinn. 1995. Estimation of current human immunodeficiency virus incidence rates from a cross-sectional survey using early diagnostic tests. *Amer. J. Epidemiology* 141 166–172.
- , —, M. Shepherd, S. Mehendale, J. Rodrigues, R. Bollinger. 1995. The AIDS epidemic in India: a new method for estimating current human immunodeficiency virus (HIV) incidence rates. *Amer. J. Epidemiology* 142 709–713.
- Centers for Disease Control. 1990. Estimates of HIV prevalence and projected AIDS cases: summary of a workshop, October 31–November 1, 1989. *Morbidity and Mortality Weekly Report* 39 110–119.
- Cox, D. R. 1962. *Renewal Theory*. Methuen and Co. Ltd., London.

- Holtgrave, D. R., N. L. Qualls. 1995. Threshold analysis and programs for prevention of HIV infection. *Medical Decision Making* **15** 311–317.
- Kaplan, E. H. 1995a. Probability models of needle exchange. *Oper. Res.* **43** 558–569.
- . 1995b. Economic analysis of needle exchange. *AIDS* **9** 1113–1119.
- , M. L. Brandeau (Eds.). 1994a. *Modeling the AIDS Epidemic: Planning, Policy and Prediction*. Raven Press, New York.
- , ———. 1994b. AIDS policy modeling by example. *AIDS* **8** (suppl 1), S333–S340.
- Keet, R. 1995. HIV-1 seroconversion and its aftermath in homosexual men: studies on acquisition of HIV-1 and natural history of HIV-1 infection. Ph.D. Dissertation. University of Amsterdam, Amsterdam, The Netherlands.
- Little, J. D. C. 1961. A proof of the queueing formula  $L = \lambda W$ . *Oper. Res.* **9** 383–387.
- Longini, I. M., W. S. Clark, L. I. Gardner, J. F. Brundage. 1991. The dynamics of CD4+ T-lymphocyte decline in HIV infected individuals: a Markov modeling approach. *J. AIDS* **4** 1141–1147.
- Mann, J., D. J. M. Tarantola, T. W. Netter (Eds.). 1992. *AIDS in the World*. Harvard University Press, Cambridge, MA.
- Nelson, K. E., D. Vlahov, L. Solomon, S. Cohn, A. Muñoz. 1995. Temporal trends of incident human immunodeficiency virus infection in a cohort of injecting drug users in Baltimore, Md. *Arch. Intern. Med.* **155** 1305–1311.
- Normand, J., D. Vlahov, L. E. Moses (Eds.). 1995. *Preventing HIV Transmission: The Role of Sterile Needles and Bleach*. National Academy Press, Washington, DC.
- Satten, G. A., I. M. Longini, Jr. 1994. Estimation of incidence of HIV infection using cross-sectional marker surveys. *Biometrics* **50** 675–688.
- Survey Research Center. 1988–1990. *Codebook for the San Francisco Men's Health Study* (Waves 1–9). Survey Research Center, University of California, Berkeley.
- Valdiserri, R. O., T. V. Aultman, J. W. Curran. 1995. Community planning: a national strategy to improve HIV prevention programs. *J. Community Health* **20** 87–100.
- Winkelstein, W. Jr., M. Samuel, N. Padian, J. A. Wiley, W. Lang, R. E. Anderson, J. A. Levy. 1987. The San Francisco men's health study: III. Reduction in human immunodeficiency virus transmission among homosexual/bisexual men, 1982–86. *Amer. J. Pub. Health* **76** 685–689.
- Zeger, S. L., K. Y. Liang. 1986. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42** 121–130.