BIOMETRICS 50, 675–688
September 1994

Estimation of Incidence of HIV Infection Using Cross-Sectional Marker Surveys

Glen A. Satten

Division of HIV/AIDS, National Center for Infectious Diseases,
Centers for Disease Control and Prevention, Atlanta, Georgia 30333, U.S.A.

and

Ira M. Longini, Jr.

Division of Biostatistics, School of Public Health,
Emory University, Atlanta, Georgia 30322, U.S.A.

SUMMARY

Methods of estimating the probability density function of infection times for a population, using serial cross-sectional measurements of a marker of disease progression, are presented. The infection time distribution may be calculated back to the beginning of the epidemic, if it is possible to sample individuals who were infected at the beginning of the epidemic; otherwise, under a Markov assumption, the infection time distribution may be calculated conditional on infection after sampling has begun. In either case, the proportion of prevalent cases infected in an arbitrary time interval between the onset and termination of sampling may be measured. Data from the San Francisco Men's Health Study are analyzed; the infection time distribution compares well with that estimated by Bacchetti (1990, *Journal of the American Statistical Association* **85,** 1002–1008) using stored sera from several San Francisco cohort studies.

1. Introduction

The estimation of HIV incidence is important in evaluating the current status of and predicting the future course of the HIV epidemic, as well as in assessing current and future health care needs of HIV-infected persons. Although AIDS cases are reported to the Centers for Disease Control and Prevention (CDC), there is no national reporting system for reports of HIV infection. As a result, estimates of HIV incidence are made using AIDS cases, through the method of back-calculation (Brookmeyer and Gail, 1988; Brookmeyer and Liao, 1990; Rosenberg and Gail, 1991). Since it is currently believed that the median time between HIV infection and onset of clinical AIDS is about 10 years, with a very small probability of progression from initial infection to AIDS in a short time (e.g., 2 years), estimates of recent HIV incidence obtained from back-calculation are considered unreliable.

The CDC, in collaboration with state and local health departments, as well as various other entities, conducts sentinel surveillance of HIV prevalence through a series of surveys conducted in selected populations. Many of these surveys are unlinked, anonymous surveys using blood drawn for other purposes, from which all personal identifiers have been removed before testing for HIV. These surveys provide a setting for collection of data of the type considered in this paper. Direct estimation of HIV incidence using only the observed patterns of prevalence from these surveys is often problematic. As patterns of incidence of HIV infection have clearly varied over time within the last decade (Brookmeyer, 1991), a steady-state assumption is not reasonable. However, many surveys find that prevalence is changing slowly on the yearly time-scale (Centers for Disease Control, 1991). Hence, the estimation of incidence from changes in prevalence alone is confounded

Key words: CD4 cell count; Cross-sectional surveys; Cubic splines; HIV; Incidence; Markers of disease progression; Markov model; Penalized maximum likelihood; San Francisco Men's Health Study.

by unmeasured factors, such as migration into or out of the target population for reasons related to having HIV infection.

In this paper, we consider the problem of estimating the incidence of HIV infection in a cross-sectional setting, in the absence of steady-state assumptions, and using data that may be subject to various sampling biases. To accomplish this, we will assume that we have collected data on a marker of disease progression, for which we have an understanding of the probabilistic relationship between marker values at different times. The marker of disease progression we will use is the CD4 cell count, although other markers may also be used.

The methods we present use only samples of infected individuals; hence, we cannot estimate incidence in the traditional sense of the number (or proportion) of susceptible individuals infected in a given time period. Instead, we will estimate the proportion of prevalent infection that has occurred in a given time interval. We can also estimate the probability density function (pdf) of infection times for the individuals infected up to the time of data collection.

In Section 2, we discuss the general properties we require of a marker and outline a Markov model for progression of CD4 counts (Longini et al., 1989, 1991). In Section 3, we develop the distributions required to analyze progression data. In Section 4, we present a method of analyzing cross-sectional marker surveys which requires that we estimate the pdf of infection times over the distant past. In Section 5, we present a model that allows estimation of current incidence only; unlike the method of Section 4, this model requires that the marker progression model be a Markov model. In Section 6, we present results of an analysis of data from the San Francisco Men's Health Study (SFMHS), and we discuss our results in Section 7. Technical details of implementation of the methods developed in Sections 4 and 5 using cubic splines are given in an Appendix.

2. Models of Marker Progression

In our usage, a marker of disease progression is any quantity that has a monotone trend throughout the course of disease, and for which we may obtain the probability distribution of marker values at time t_r, given that infection has occurred at time $t \leq t_r$. We will call this probability the "marker progression function," and will assume that it is known from cohort studies and is (relatively) invariant from population to population. This assumption is similar to the situation in back-calculation, where we assume that an incubation distribution for the time between initial infection and AIDS diagnosis is known from a cohort study, and this incubation distribution is then applied to other populations (Brookmeyer and Gail, 1988).

The marker we will consider in this paper is CD4 cell count. Various models for CD4 progression have been considered; we will consider only the Markov model of Longini et al. (1991), which is shown in schematic form in Figure 1. Seropositive individuals begin in stage 1, and progress through six stages of declining CD4 count, ultimately reaching clinical AIDS (stage 7) and death (stage 8). The time spent in each stage is assumed to be exponentially distributed, and no "backward progression" from a stage characterized by a lower CD4 count to a stage characterized by a higher CD4 count is allowed. This model has been fit to a number of cohort studies (Longini et al., 1989, 1991, 1992), from which estimates of the marker progression may be obtained. The stage occupied by an individual at time t, as determined by a CD4 measurement at time t, is denoted by CD4(t). We will consider exclusively marker information in this discrete (staged) form; however, our results are also applicable for a continuous marker model, with sums over stages replaced by integrals over CD4 values. The information obtained by fitting the Markov model to cohort data is a transition probability

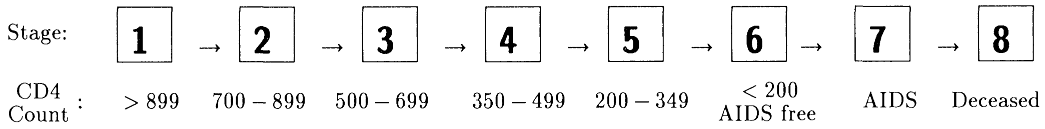

Figure 1. Schematic diagram of progression of seropositives through the stages of the Markov model for CD4 progression.

$$T_{ii'}(t_2, t_1) = \Pr[\text{CD4}(t_2) = i | \text{CD4}(t_1) = i'], \quad 1 \leq i' \leq i \leq 8, \quad t_1 \leq t_2.$$

The explicit form of the transition probability can be found in Longini et al. (1989). Since the waiting time in stage 1 is exponential, the "marker progression function" obtained for the Markov model is

$$\Pr[\text{CD4}(t_r) = i | \text{infection time} = t] = T_{i1}(t_r, t).$$

3. Probability Distributions of Interest in Analyzing Marker Data

A quantity of central interest in estimating infection incidence is the pdf of infection times, denoted by $I(t)$. In general, we may know $I(t)$ only up to a constant, since if the latest time we collect data is t_r, we may know only

$$\frac{I(t)}{\int_{-\infty}^{t_r} I(t')\,dt'} = \beta(t_r)I(t) \tag{1}$$

for $t \leqslant t_r$. The constant $\beta(t_r)$ will be inestimable in our work; hence, only quantities involving ratios of I will be estimable.

In a survey taken at time t_r, even if sampling is independent of time of infection, we will not sample times of infection according to the pdf $I(t)$. We assume that we sample only those individuals who meet some set of eligibility requirements. We will denote the set of individuals at time t_r who meet the eligibility criteria for survey entrance as the set of "qualified" individuals. To see the importance of qualification, note that in a population survey (as opposed to an analysis of a database), a minimum requirement for qualification is being alive at the time the survey is conducted. To use existing marker model estimates, we will require our surveys to be such that being qualified corresponds to being in a subset of the stages of the Markov model. (If we were using a continuous model for CD4 decline, we would require that being qualified correspond to being in a certain set of ranges of CD4 values.) We will denote the set of qualified stages at time t_r by $Q(t_r)$. For example, the following are all valid choices of qualified individuals: AIDS-free $[Q(t_r) = \{\text{CD4}(t_r) \leqslant 6\}]$; alive $[Q(t_r) = \{\text{CD4}(t_r) \leqslant 7\}]$; or individuals having a CD4 count greater than 500 $[Q(t_r) = \{\text{CD4}(t_r) \leqslant 3\}]$.

Denote the probability of being qualified at time t_r given that infection occurred at time $t \leqslant t_r$ by

$$\begin{aligned} S(t_r|t) &= \Pr[\text{individual is qualified at time } t_r|\text{infection time} = t] \\ &= \sum_{i \in Q(t_r)} \Pr[\text{CD4}(t_r) = i|\text{CD4}(t) = 1]. \end{aligned} \tag{2}$$

The notation $S(t_r|t)$ was chosen because, if being qualified corresponds to being AIDS-free, then $S(t_r|t)$ is the survival function for the AIDS incubation time distribution.

We can now see that the pdf for infection times obtained in an unbiased sample at time t_r is

$$\Pr[t \leqslant \text{infection time} \leqslant t + dt|\text{qualified at } t_r] = \frac{S(t_r|t)I(t)dt}{\int_{-\infty}^{t_r} S(t_r|t')I(t')\,dt'}, \quad t \leqslant t_r. \tag{3}$$

Using our assumed condition on qualified individuals, the marker progression function for those individuals qualified at time t_r is given by

$$\begin{aligned} &\Pr[\text{CD4}(t_r) = i|\text{qualified at } t_r, \text{ infection time} = t] \\ &\quad = \frac{\Pr[\text{CD4}(t_r) = i|\text{infection time} = t]}{S(t_r|t)}, \quad i \in Q(t_r). \end{aligned} \tag{4}$$

The joint distribution of infection times and marker values, obtained from (3) and (4), is given by

$$\begin{aligned} &\Pr[\text{CD4}(t_r) = i, t \leqslant \text{infection time} \leqslant t + dt|\text{qualified at } t_r] \\ &\quad = \frac{\Pr[\text{CD4}(t_r) = i|\text{infection time} = t]I(t)dt}{\int_{-\infty}^{t_r} S(t_r|t)I(t)\,dt}, \quad i \in Q(t_r). \end{aligned}$$

The marginal distribution of qualified seropositives among the stages in Q is given by integrating the joint distribution of marker values and infection times over infection times. Hence, we find

$$p_i(t_r) \equiv \Pr[\mathrm{CD4}(t_r) = i|\text{qualified at } t_r], \quad i \in Q(t_r)$$

$$= \frac{\int_{-\infty}^{t_r} \Pr[\mathrm{CD4}(t_r) = i|\text{infection time} = t]I(t)\,dt}{\int_{-\infty}^{t_r} S(t_r|t)I(t)\,dt}. \tag{5}$$

Note that by equation (2), $p_i(t_r)$ is properly normalized for any t_r and any set of stages for the "qualifieds."

By working only with samples of seropositives, we cannot measure incidence in the usual sense of a change in numbers (or proportion) of uninfected individuals. Instead, we can measure incidence as a proportion of current prevalence using (1). Define the proportion of prevalent infection at time t_r that has occurred between times $t_{r'}$ and $t_{r''}$, where $t_{r'} \leq t_{r''} \leq t_r$, by

$$\pi(t_{r'}, t_{r''}, t_r) = \int_{t_{r'}}^{t_{r''}} \frac{S(t_r|t)I(t)}{\int_{-\infty}^{t_r} S(t_r|t')I(t')\,dt'}\,dt. \tag{6}$$

If an independent estimate of the prevalence at time t_r, denoted $\mathcal{P}(t_r)$, is available, then the incidence as a proportion of the uninfected population may be estimated by $\pi(t_{r'}, t_{r''}, t_r)\mathcal{P}(t_r)$.

4. Reconstructing *I*(*t*) Using Marker Surveys: The Backward Method

In many cases, it is possible to reconstruct the function $I(t)$ from marker survey data, using only data from the qualified seropositives sampled. Two assumptions are necessary: (i) that sampling of seropositives is independent of time since seroconversion, and (ii) that the beginning of the epidemic in the population is sufficiently recent that some seropositives infected at the onset of the epidemic are still qualified. The second condition ensures that we are not trying to estimate $I(t)$ for infection times we cannot have information about (e.g., for which there are no qualified seropositives). Since this method reconstructs $I(t)$ for times before the onset of sampling, we will refer to it as the backward method. If this condition does not hold, the methods described in Section 5 may still be used.

Suppose that we have conducted surveys at times $t_{r1} < t_{r2} < \cdots < t_{rK}$, and are sampling with replacement. In the kth survey we sample $N(t_{rk})$ qualified seropositives. We will denote the set of stages corresponding to being qualified at time t_{rk} by $Q(t_{rk})$. At each survey, each qualified seropositive can be assigned to one of the stages in $Q(t_{rk})$, so that the data may be summarized in the $K \times I$ table shown in Table 1. Thus, there are $n_i(t_{rk})$ individuals observed in each qualified stage $i \in Q(t_{rk})$, with

$$\sum_{i=1}^{I} n_i(t_{rk}) = N(t_{rk}), \quad k = 1, \ldots, K.$$

Note that the definition of "qualified" may change over time, leading to a different set of qualified stages $Q(t_{rk})$ at each time. If the definition of "qualified" changes over time, then some of the entries in Table 1 may be structural zeros.

Table 1

Table of stage by survey time for qualified seropositives. At time t_{rk}, $n_i(t_{rk})$ of the qualified seropositives were observed to be in stage i.

Stage of seropositives	Time of survey			
	t_{r1}	t_{r2}	$\cdots$	t_{rK}
1	$n_1(t_{r1})$	$n_1(t_{r2})$	$\cdots$	$n_1(t_{rK})$
2	$n_2(t_{r1})$	$n_2(t_{r2})$	$\cdots$	$n_2(t_{rK})$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
I	$n_I(t_{r1})$	$n_I(t_{r2})$	$\cdots$	$n_I(t_{rK})$
	$N(t_{r1})$	$N(t_{r2})$	$\cdots$	$N(t_{rK})$

Conditional on $N(t_{rk})$, the number of seropositives sampled at time t_{rk}, the contribution to the likelihood from the data gathered at each survey (each column in Table 1), denoted $L(t_{rk})$, is multinomial, e.g.,

$$L(t_{rk}) = \frac{N(t_{rk})!}{\prod_{i \in Q(t_{rk})} n_i(t_{rk})!} \prod_{i \in Q(t_{rk})} p_i(t_{rk})^{n_i(t_{rk})}, \tag{7a}$$

where $p_i(t_{rk})$, $i \in Q(t_{rk})$ is given in (5). Recall that the $p_i(t_{rk})$'s are properly normalized for any time t_{rk} and any set $Q(t_{rk})$. The overall likelihood L is the product of the columnwise multinomials, i.e.,

$$L = \prod_{k=1}^{K} L(t_{rk}). \tag{7b}$$

As the marker progression function is known, the only unknown in L is $I(t)$; however, with only a finite number of stages in $Q(t_{rk})$, we must either choose $I(t)$ from a parametric family or regularize the problem in some way and make a semiparametric estimate of $I(t)$. We have adopted the second course, and have replaced the log-likelihood log[L] by a penalized log-likelihood L_p, given by

$$\log[L_p] = \log[L] - \alpha \int_{-\infty}^{t_{rK}} \left(\frac{d^2 I(t)}{dt^2}\right)^2 dt, \tag{8}$$

where α is a smoothing parameter chosen to give reasonable estimates of $I(t)$. Unless a functional form for $I(t)$ for times in the distant past is chosen, it is necessary to assume that there is some time t_0 before which no infection occurred in the population.

The penalized likelihood is maximized subject to constraints

$$\int_{-\infty}^{t_{rK}} I(t)\, dt = 1$$

and $I(t) \geqslant 0$. In the Appendix, we describe the mathematical details necessary to carry out penalized maximum likelihood estimation when $I(t)$ is chosen to be a cubic spline function.

Although the number of surveys in the method described above is arbitrary, it should be noted that the information on $I(t)$ recovered from a single survey may be very limited, particularly if $I(t)$ has a complicated shape. However, in some cases, a single survey may provide valuable information. For example, a survey among 21-year-olds, in which age a replaces time t, and for whom $I(a)$ could be assumed monotonically increasing, would provide information on the age of infection among teenagers.

5. Estimation of $I(t)$ Without Reconstruction of $I(t)$ in the Distant Past: The Forward Method

In this section, we present a method that allows estimation of $I(t)$ for recent times only. Unlike the methods developed in Section 4, the methods described here make heavy use of the Markov property. However, they allow relaxation of some of the requirements of Section 4, specifically that there be a time t_0 before which there were no infections in the population, and that some individuals infected early in the epidemic be qualified for sampling at the onset of data collection. On the other hand, they require at least two serial surveys. Since reconstruction of $I(t)$ is limited to times after the onset of sampling, we will refer to it as the forward method.

Suppose that we have collected data in the form described in Table 1. By using the Markov property, it is possible to estimate $I(t)$ only for $t_{r1} \leqslant t \leqslant t_{rK}$. To do this, we must estimate the initial condition of the Markov process; this is accomplished by considering the values of $p_i(t_{r1}) \equiv p_i$ as parameters in the model to be estimated. Using the Chapman–Kolmogorov equation, we can write $p_i(t_{rk})$ in terms of the p_i's at the earlier time t_{r1} as

$$p_i(t_{rk}) = \gamma(t_{rk}) \left[\int_{t_{r1}}^{t_{rk}} T_{i1}(t_{rk}, t) \beta I(t)\, dt + \sum_{i' \leqslant i} T_{ii'}(t_{rk}, t_{r1}) p_{i'} \right]. \tag{9}$$

The constant of proportionality $\gamma(t_{rk})$ is chosen at each time t_{rk} by normalizing the $p_i(t_{rk})$ over $i \in Q(t_{rk})$. Note that $p_i(t_{rk})$ for $k > 1$ depends on $\{p_i\}$ and $I(t)$ only in the range $t_{r1} \leqslant t \leqslant t_{rk}$. As before, $I(t)$ is known only up to a normalization constant. If we choose this value by imposing the condition that

$$\int_{t_{r1}}^{t_{rK}} I(t)\, dt = 1, \tag{10}$$

so that $I(t)$ is the infection time pdf *conditional on infection between times* t_{r1} *and* t_{rK}, then β is identifiable, and has the interpretation that

$$\beta = \frac{\text{Number of infections occurring in time interval } (t_{r1}, t_{rK})}{\text{Number of qualified seropositives at time } t_{r1}}.$$

The likelihood is the same product-multinomial as equation (7) of Section 4, except that the $p_i(t_{rk})$ are calculated using (9), with the result that the likelihood is a function of the parameter vector $\boldsymbol{\Theta}$ given by

$$\boldsymbol{\Theta} = (p_1, \dots, p_I, \theta_1, \dots, \theta_J, \beta),$$

where $\theta_1, \dots, \theta_J$ are parameters that determine $I(t)$. If regularization via penalized maximum likelihood is used, then the penalty is applied only over the range $t_{r1} \leq t \leq t_{rK}$. The likelihood is maximized subject to

$$\sum_{i \in Q(t_{r1})} p_i = 1$$

as well as (10). In the Appendix, we give details on how the resulting likelihood can be maximized using basis spline functions.

Although we have not estimated $I(t)$ before sampling begins, equation (5) still holds at t_{r1}. Hence, we may express the proportion of infection in stages Q^* at time $t_{rk''}$ that occurred between times t_{rk} and $t_{rk'}$ (where $t_{r1} \leq t_{rk} \leq t_{rk'} \leq t_{rk''} \leq t_{rK}$), given in (6), by

$$\pi(t_{rk}, t_{rk'}, t_{rk''}) = \frac{\beta \displaystyle\int_{t_{rk}}^{t_{rk'}} S(t_{rk''}|t)I(t)\, dt}{\beta \displaystyle\int_{t_{r1}}^{t_{rk'}} S(t_{rk''}|t)I(t)\, dt + \sum_{i \in Q^*} \sum_{i' \leq i} T_{i,i'}(t_{rk''}, t_{r1})p_{i'}}. \tag{11}$$

It is also possible to change the criteria for qualification during the course of the study as in Section 4. In this case, it is important to note that the set of stages in Q^* in equation (11) is not limited to the set of qualifieds at time $t_{rk''}$, but may include any stages for which an estimate of p_i, the stage occupation probability at time t_{r1}, is available. For example, suppose that the set of qualifieds includes all stages from 1984 through 1987, and only stages 1–3 after 1987 (so as to remove any effect of treatment). It is still possible to calculate the proportion of people in all stages in 1991 who were infected from 1989 through 1990.

6. Example: The San Francisco Men's Health Study

As an example of our methodology, we present an analysis of the CD4 cell count data collected in the SFMHS, which is a population-based cohort study comprised of a random sample of 1,045 homosexual and bisexual men aged 25–54 years at enrollment (Winkelstein et al., 1987). Follow-up has been conducted approximately every 6 months since mid-1984; we have discretized these follow-up times to 14 surveys conducted up to January 1991. These "survey times" are given in Table 2, as well as the numbers of men in each stage at each survey time. A total of 451 HIV-infected men were observed; of these, 43 seroconverted during the course of the study. The average number of exams per HIV-positive person was 4.4.

We have assumed that the stage-by-sampling time table of Table 1, given in Table 2 for the SFMHS data, has remained representative of the CD4 cell counts of the original population of men sampled at the time the cohort was assembled. Hence, the primary effects of the difference between the SFMHS design and the serial cross-sectional survey design will be reflected in the variance estimate of $\hat{I}(t)$, as well as in the aging of the target population.

Table 3 is adapted from Longini, Clark, and Karon (1993). They fitted the Markov model, described in Figure 1, to the HIV progression data from the SFMHS. The estimated mean occupancy times in Table 3 are the reciprocals of the estimated transition rates. To assess the effect of treatment on our incidence estimates, we considered two analyses. Analysis A uses the full data of Table 2; analysis B uses all stages in Table 2 up to 1987 and only stages 1–3 after 1987 (corresponding

Table 2
Number of seropositives by stage and sampling time for the SFMHS

Survey number k	Time t_{rk}	Stage[a] 1	2	3	4	5	6	Total
1	1/91	12	22	36	24	21	22	137
2	7/90	25	26	41	38	38	29	197
3	1/90	18	28	39	44	40	28	197
4	7/89	18	20	48	51	37	34	208
5	1/89	26	24	46	53	53	22	224
6	7/88	19	23	57	50	52	21	222
7	1/88	10	28	61	53	67	35	254
8	7/87	17	33	62	71	69	25	277
9	1/87	25	23	54	49	40	22	213
10	7/86	41	49	74	78	41	17	300
11	1/86	37	52	84	62	27	19	281
12	7/85	36	48	91	76	50	21	322
13	1/85	40	72	89	46	27	11	285
14	7/84	45	47	81	38	18	4	233

[a] Stages are as given in Figure 1.

Table 3
Stages and occupancy times (fitted in the SFMHS)[a]

CD4 stages	Mean occupancy time[b] (SFMHS) (months)	Cumulative (months)
(1) >899	20.79	20.79
(2) 700–899	17.33	38.12
(3) 500–699	29.15	67.28
(4) 350–499	23.92	91.20
(5) 200–349	20.12	111.31
(6) <200	12.42	123.74

[a] Adapted from Table 3 in Longini et al. (1992).

[b] Estimated from all the observed transitions in the data, regardless of treatment status.

to CD4 $\geq$ 500). Treatment here refers to any combination of zidovudine or pentamidine. In 1987, only 6.4% of men were on treatment (Longini, Clark, and Karon, 1993).

In Figure 2, we show our estimate of the infection time pdf for the population represented by the SFMHS, calculated using the methods of Section 4, for analyses A and B. We have assumed a t_0 of January 1978. If we choose an earlier t_0 (January 1970), approximately 7% of the distribution would fall before January 1978 (depending on the amount of smoothing). The value of the penalty parameter was 3.0×10^8 for each case; this value was chosen to give results of reasonable but not excessive smoothness. On the same figure, we also show the estimate of $I(t)$ given by Bacchetti (1990) for the homosexual/bisexual population of the city of San Francisco, obtained by truncated survival analysis techniques applied to time-of-seroconversion data from three cohorts. We have scaled Bacchetti's estimate, $I_b(t)$, for ease of comparison, so that the total probability represented by $I_b(t)$ is .90; this value is the average of the probability of seroconversion before January 1989 computed using analyses A and B. Finally, Figure 2 also shows an estimate of $I(t)$ using all CD4 stages, but using only survey data from the final six surveys (July 1988 through January 1991).

In Figure 3, we show two estimates of the infection time pdf obtained by the methods of Section 5, for the time period July 1989 to January 1991. The first analysis uses all stages in Table 2 up to 1987 and only stages 1–3 after 1987 (corresponding to CD4 $\geq$ 500). The second analysis uses stages 1–6 for the July 1984 survey, and only stages 1–3 thereafter. To facilitate comparison with the results in Figure 2, we have scaled $I(t)$ so that the area under $I(t)$ is .233, which is the area under the $I(t)$ estimate produced by analysis B, when integrated from July 1984 to January 1991. In the same figure, we have also shown the 95% "confidence intervals" obtained by the percentile method from 100 bootstrap iterates (Efron, 1982) of analysis B, drawn from Table 2. A penalty value of 3×10^8 was used for all analyses. These intervals should be viewed with some caution, since they represent the variability that would be inherent in the data if they had been collected as a series of cross-

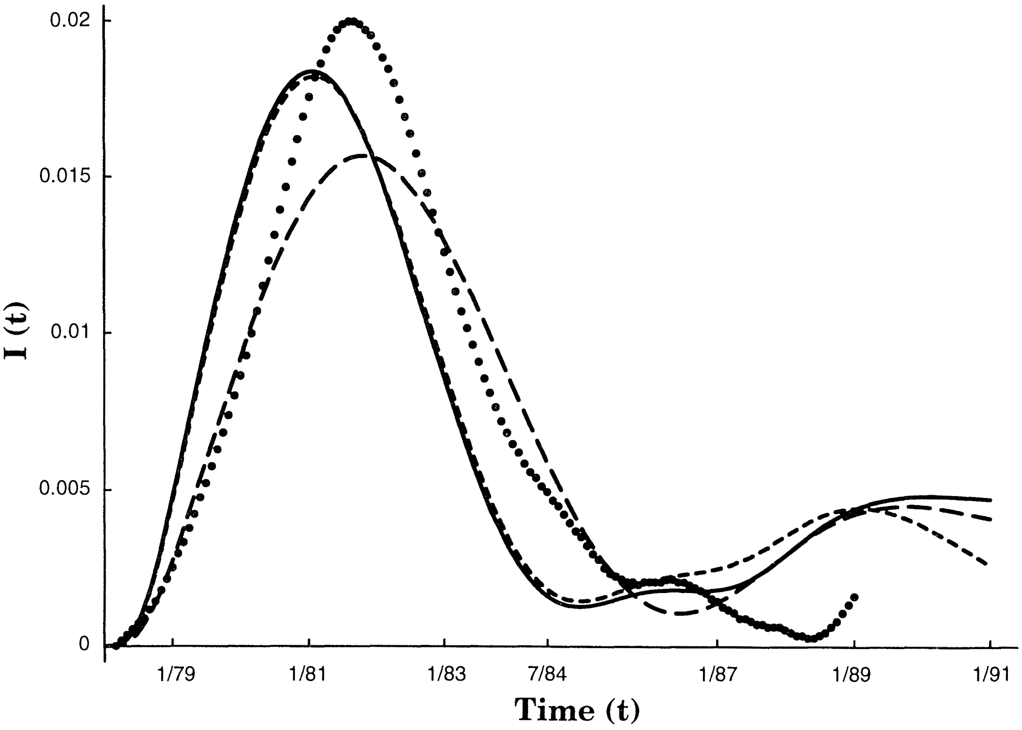

Figure 2. Estimates of infection time pdf $I(t)$ from backward method, and Bacchetti. *Solid line:* All stages (analysis A). *Short dashes:* All stages until 1987, stages 1–3 after 1987 (analysis B). *Long dashes:* Latest 3 years of data only. *Dots:* (Scaled) infection time pdf from Bacchetti (1990).

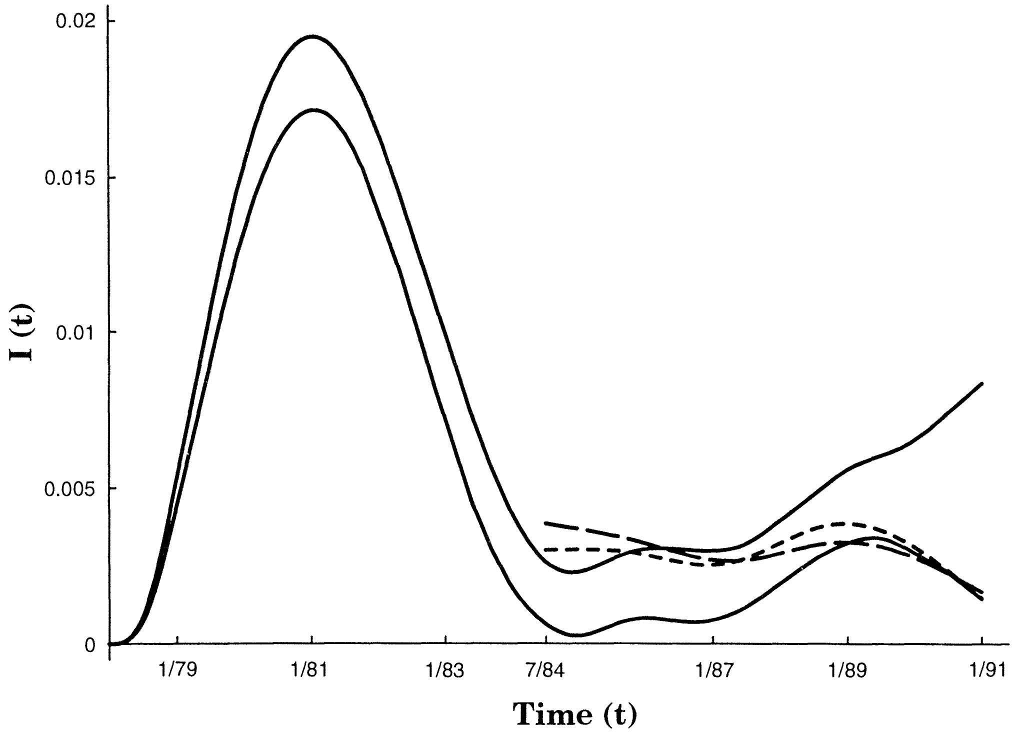

Figure 3. *Solid lines:* Pointwise bootstrap confidence bands for infection time pdf $I(t)$ from the backward method. *Short dashes:* Scaled infection time pdf for forward method, using all stages until 1987, stages 1–3 after 1987. *Long dashes:* Scaled infection time pdf for forward method, using all stages for July 1984 survey, and stages 1–3 thereafter.

sectional surveys. In addition, the penalty part of the likelihood may be considered as a prior distribution on $I(t)$; this strong prior favoring smoothness also decreases the variability of $\hat{I}(t)$. We can, however, conclude that estimates using the methodology of Section 5 appear to be consistent with those obtained using the methodology of Section 4.

In Table 4, we present estimates of the proportions of men seroprevalent in January 1991 who were infected during three periods: 1990; July 1984 through December 1989; and before July 1984. These proportions are given for five analyses, using various rates and stages. There is fairly good agreement among the analyses; if we estimate the proportion infected during the course of the study (July 1984 to January 1991) to be 40%, and note that 137 seropositive men were tested in January 1991, we estimate about 55 seroconversions, as compared to 43 observed. Unfortunately, this comparison must be regarded with caution, as both estimates of the number seroconverted are subject to error: Our estimate of 55 seroconversions assumes that all seropositives were tested in January 1991, whereas the observed 43 seroconversions counts only those actually tested by January 1991. We can conclude, however, that there is an overall agreement. Given these small rates of seroconversion (on the order of two every 3 months), no conclusion on differences in the shape of the estimates of $I(t)$ we have presented in Figures 2 and 3 can be made.

Table 4
Proportion of 1991 *seroprevalent men by time of infection for various analyses*

	% of 1991 seroprevalent men infected in time interval:		
Analysis	1/90–12/90	7/84–12/89	Before 7/84
CD4 ⩾ 500 after 1987 forward method	4.4	31.1	64.5
All stages forward method	7.5	30.1	62.4
CD4 ⩾ 500 after 1984 forward method	4.8	34.7	60.6
CD4 ⩾ 500 after 1987 backward method	6.4	30.5	63.1
All stages backward method	9.4	27.8	62.8

Bacchetti (1990) has estimated that there were 20,060 seroconversions in the city of San Francisco before July 1984. Using analysis B, we estimate the proportion of seroconversions from July 1989 to January 1991 to be .233, corresponding to an estimate of 20,060/(1 − .233) = 26,154 seroconversions by January 1991. The cumulative number of AIDS cases diagnosed before January 1991 in homosexual/bisexual men in the city of San Francisco, adjusted for reporting delay (Karon, Devine, and Morgan, 1989), is 7,219 (Centers for Disease Control, unpublished data); hence, we estimate about 18,940 AIDS-free seropositives in San Francisco. Using the range of values in column 1 of Table 4, we may estimate approximately .044 × 18,940 ≈ 830 to .094 × 18,940 ≈ 1,780 new infections occurred in the population represented by the SFMHS during 1990.

7. Discussion

In this paper, we have shown that it is possible to reconstruct the HIV incidence history of a population from a series of cross-sectional surveys of a marker of disease progression. We have constructed several such estimates for the population of men for which the SFMHS is representative. We obtain overall agreement with Bacchetti's estimate, obtained from stored sera data from several studies, for the incidence history of the HIV epidemic in the San Francisco homosexual and bisexual population. It should be noted that any disagreement with Bacchetti's estimate may reflect a difference in populations sampled; since only the hepatitis B vaccine study has stored serum for the period 1978 to 1984, Bacchetti's estimate in this time period is based primarily on that study. Our estimates differ slightly from Bacchetti's, as we predict a peak incidence in January 1981, compared to his predicted peak during July and August 1981.

Gail, Rosenberg, and Goedert (1990) have shown that treatment may have an important effect on the rate at which seropositive individuals progress to AIDS. We have also considered the possible effect of treatment on our estimates of incidence history. This can be accomplished in a number of ways within our methodology: by excluding those individuals in CD4 stages where treatment is possible; by using a time-dependent marker progression function (which accounts for the phase-in of treatment over time) (Brookmeyer and Liao, 1990; Brookmeyer, 1991; Rosenberg, Gail, and

Carroll, 1992; and Longini et al., 1992); or by allowing separate states for those on and off treatment, with transitions to the treatment stages allowed only after 1987. However, our results in Figure 2 show that our analysis B, in which treatment is accounted for by excluding those individuals with CD4 counts over 500 after 1987, is in substantial agreement with analysis A, in which treatment is ignored. In addition, we have reproduced analysis B, using rates determined from only those individuals known to not be on treatment (Longini, Clark, and Karon, 1993). The results of this analysis were in very close agreement with those of analysis B.

The methods we have presented use data gathered on infected individuals. By using only infected individuals, we avoid bias in our results that could result from over- or undersampling seropositives, so long as this bias is unrelated to stage of infection. In some cases, this assumption can also be relaxed further so that sampling is required to be unbiased only with respect to the distribution of the marker over a restricted range of marker values. For example, we may be willing to believe that we have an unbiased sample with respect to CD4 counts only for those individuals with CD4 $\geq$ 500 at the time of sampling. In general, the smaller the range of marker values we are willing to trust, the more the Markov assumption is required.

We have presented two methods of estimating the incidence history in a population; each method has advantages and disadvantages. The backward method, presented in Section 4, estimates the pdf of infection times over all time in the epidemic, and is useful if a non-Markov model of marker progression is used. However, we must assume that there is some time t_0 before which no infection could have occurred, and we must be able to observe some individuals infected around time t_0. Because the late stages of disease (corresponding to low CD4 counts) are the oldest infections, the backward method works best if as many stages are included in the definition of ''qualified'' as possible (at least in some of the surveys). The forward method, presented in Section 5, does not require that the late stages be included among the qualifieds, since the pdf of infection times is not reconstructed back to t_0. In some cases, however, it may be of intrinsic interest to know this pdf over the longest possible range of times.

Both methods assume that there is no differential migration of seropositives, i.e., that the populations sampled in any pair of surveys are governed by the same pdf of infection times. If this is not the case, then differential migration of seropositives has occurred. In general, we expect differential migration to occur predominantly in individuals with late-stage infection, since recently infected people are less likely to know their serostatus. This is particularly likely if sampling is carried out in a health care setting (e.g., at a clinic or a hospital). Hence, the forward method may be preferred if differential migration is expected to be important, since it is easily implemented using only the first three stages. For the backward method, the analysis that uses all stages at the first survey time to estimate the initial stage occupation probabilities, then uses only stages 1–3 thereafter, may be the most insensitive to differential migration. An initial strategy in assessing whether differential migration is a factor in an analysis is to compare results obtained using different numbers of stages.

For the forward method, in estimating the initial stage occupation probabilities as parameters in the model, we have ignored the fact that there is a ''true'' distribution of infection times governing the epidemic for times before the onset of sampling. This true infection time distribution may not be a good continuation of the infection time distribution estimated after sampling has begun, in the sense that we may unknowingly have allowed a jump or other unnatural behavior in the infection time pdf at the time of the first survey. This danger, however, is minimized as a fairly wide variety of infection time pdfs can probably give rise to a given set of initial stage occupation probabilities.

The primary method of reconstruction of the infection time pdf for the HIV epidemic is back-calculation. Our methodology differs from back-calculation in several important ways. Back-calculation requires an exhaustive sequence of AIDS cases dating back to a time t_0 before which it may be assumed that there were no AIDS cases. In a population in which such records are not available—for example, in many third-world countries—back-calculation cannot be used, although our methodology could be implemented. An important feature of back-calculation is the estimation of the total number of infected individuals. We are unable to estimate this quantity, since we have assumed that we have only a random sample of seropositive individuals. However, if we also had available a complete list of AIDS cases arising during the course of sampling, our methodology could be easily modified to yield the total number infected, by using a product likelihood in which the first term in the product is our likelihood and the second term the likelihood for back-calculation (Brookmeyer and Gail, 1988).

Our methodology requires a knowledge of the probabilistic law determining the change in marker level over time. This information is summarized in what we have called the marker progression function. Although CD4 counts are highly variable at the individual level, our method requires only understanding CD4 change at the population level. By focusing on the proportion of the sample falling in discrete CD4 stages (e.g., tabular data of the form shown in Table 1), much of this individual variation will average out in the observed data. Similarly, the model of CD4 progression we have used has been fit to smoothed data by imposing "persistence criteria" on the observed transitions in the cohort (Longini et al., 1991).

The example shown in this paper was the SFMHS, a population-based cohort study. We plan to implement these methods in the setting of serial cross-sectional studies with data collected from the CDC Family of Surveys (Centers for Disease Control, 1990). Unfortunately, CD4 measurements were not made in most of these surveys, and CD4 cells cannot be measured in stored serum. We are investigating the usefulness of other markers, particularly β_2-microglobulin (Fahey et al., 1990), which can be measured in stored serum, and will be reporting our progress in that area elsewhere. In addition, a stage model using total lymphocyte count has been reported (Le et al., 1991); although total lymphocyte count is not available from stored serum, it is routinely collected in many medical settings. Thus, we believe that cross-sectional marker surveys will be an important source of HIV incidence information in analyses conducted at CDC.

Acknowledgements

The authors thank John Karon for helpful discussions, and Scott Clark for computer support. The data analyzed in this paper were provided by the SFMHS, which is supported by Contract N01-AI-82515 from the National Institute of Allergy and Infectious Diseases. We also thank Kimberly Page-Bodkin for extracting the data used in our analyses.

Résumé

Les méthodes permettant d'estimer la fonction de densité de probabilité des temps d'infection à partir de mesures transversales répétées de marqueurs de progression d'une maladie sont présentées. S'il est possible de disposer de prélèvements de sujets infectés dès le début de l'épidémie, la distribution des temps d'infection peut être calculée en remontant au début de celle-ci. Sinon, sous l'hypothèse d'un modèle Markovien, la distribution conditionnelle des temps d'infection peut être calculée à partir du début des prélèvements. Dans les deux cas, la proportion de cas prévalents infectés dans un intervalle de temps arbitraire situé entre le début et la fin des prélèvements peut être mesurée. Les données de l'Etude sur la Santé Masculine à San Francisco ont été analysées: la distribution des temps d'infection est tout à fait comparable à l'estimation proposée par Bacchetti (1990, *Journal of the American Statistical Association* **85,** 1002–1008), basée sur des sérums conservés au cours de plusieurs études de cohortes réalisées à San Francisco.

References

Bacchetti, P. (1990). Estimating the incubation period of AIDS by comparing population infection and diagnosis patterns. *Journal of the American Statistical Association* **85,** 1002–1008.

Brookmeyer, R. (1991). Reconstruction and future trends of the AIDS epidemic in the United States. *Science* **253,** 37–42.

Brookmeyer, R. and Gail, M. H. (1988). A method for obtaining short-term projections and lower bounds on the size of the AIDS epidemic. *Journal of the American Statistical Association* **83,** 301–308.

Brookmeyer, R. and Liao, J. (1990). Statistical modeling of the AIDS epidemic for forecasting health care needs. *Biometrics* **46,** 1151–1163.

Centers for Disease Control (1990). Special Section—The Sentinel HIV Seroprevalence Surveys. *Public Health Reports* **105,** 113–171.

Centers for Disease Control (1991). *National HIV Serosurveillance Summary: Results Through 1990.* Atlanta, Georgia: U.S. Department of Health and Human Services, Public Health Service.

de Boor, C. (1978). *A Practical Guide to Splines.* New York: Springer-Verlag.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* Philadelphia, Pennsylvania: Society for Industrial and Applied Mathematics.

Fahey, J. L., Taylor, Jeremy M. G., Detels, R., Hofmann, B., Melmed, R., Nishanian, P., and Giorgi, J. V. (1990). The prognostic value of cellular and serologic markers in infection with human immunodeficiency virus Type 1. *New England Journal of Medicine* **322,** 166–172.

Gail, M. H., Rosenberg, P. S., and Goedert, J. J. (1990). Therapy may explain recent deficits in AIDS incidence. *Journal of Acquired Immune Deficiency Syndromes* **3,** 296–306.

Karon, J. M., Devine, O. J., and Morgan, W. M. (1989). Predicting AIDS incidence by extrapolating from recent trends. In *Mathematical and Statistical Approaches to AIDS Epidemiology*, C. Castillo-Chavez (ed.). Lecture Notes in Biomathematics, **83,** 58–88. Berlin: Springer-Verlag.

Le, N. D., Schechter, M. T., Le, T. N., Craib, K. J. P., and Montaner, J. S. G. (1991). Use of the Markov model to estimate the waiting times for the proposed WHO clinical staging of HIV infection in a cohort of homosexual men. Abstract presented in the 7th International Conference on AIDS, Florence, Italy, Vol. 2, p. 32.

Longini, I. M., Byers, R. H., Hessol, N. A., and Tan, W. Y. (1992). Estimating the stage-specific numbers of HIV infection using a Markov model and back-calculation. *Statistics in Medicine* **11,** 831–843.

Longini, I. M., Clark, W. S., Byers, R. H., Ward, J. W., Darrow, W. W., Lemp, G. F., and Hethcote, H. W. (1989). Statistical analysis of the stages of HIV infection using a Markov model. *Statistics in Medicine* **8,** 831–843.

Longini, I. M., Clark, W. S., Gardner, L. I., and Brundage, J. F. (1991). The dynamics of CD4+ T-lymphocyte decline in HIV-infected individuals: A Markov modeling approach. *Journal of Acquired Immune Deficiency Syndromes* **4,** 1141–1147.

Longini, I. M., Clark, W. S., and Karon, J. (1993). The effect of routine use of therapy in slowing the clinical course of human immunodeficiency virus (HIV) infection in a population-based cohort. *American Journal of Epidemiology* **137,** 1229–1240.

Rosenberg, P. S., Gail, M. H., and Carroll, R. J. (1992). Estimating HIV prevalence and projecting AIDS incidence in the United States: A model that accounts for therapy and changes in the surveillance definition of AIDS. *Statistics in Medicine* **11,** 1633–1655.

Rosenberg, P. S. and Gail, M. H. (1991). Back-calculation of flexible linear models of the human immunodeficiency virus infection curve. *Applied Statistics* **40,** 269–282.

Winkelstein, W., Jr., Lyman, D. M., Padian, N., Grant, R., Samuel, M., Wiley, J. A., Anderson, R. E., Lang, W., Riggs, J., and Levy, J. A. (1987). Sexual practices and risk of infection by the human immunodeficiency virus. *Journal of the American Medical Association* **257,** 321–325.

Received March 1992; *revised September* 1992; *accepted January* 1993.

Appendix

In this appendix, we give details for carrying out the maximization of the likelihoods developed in Sections 4 and 5. We consider first the backward method, described in Section 4, taking $I(t)$ in the class of cubic spline functions with knot sequence $\boldsymbol{\tau}$. $I(t)$ is further restricted such that $I(t) = 0$ for $t \leqslant t_0$ with $d^nI(t)/dt^n|t_0 = 0$, $n = 1, 2$, where t_0 is some time before which it is reasonable to assume that no infection is possible. If it is not possible to find such a time t_0, then the methods in Section 5 may be used; otherwise, a parametric form for $I(t)$, when $t \leqslant t_0$, could be assumed. No assumptions on $I(t)$ or its derivatives are made at t_r. Assumptions about the behavior of $I(t)$ and its derivatives are carried in the choice of knot sequence $\boldsymbol{\tau}$. For cubic splines, our assumptions on $I(t)$ correspond to a knot sequence of the form

$$\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_{J+4}),$$

where $\tau_1 = t_0$, $\tau_2 > t_0$, $\tau_J < t_r$, and $\tau_{J+1} = \tau_{J+2} = \tau_{J+3} = \tau_{J+4} = t_r$. For more details on splines and knot sequences, see de Boor (1978).

With our assumptions, $I(t)$ can be written as

$$I(t) = \sum_{j=1}^{J} \theta_j M_j[t, \ell, \boldsymbol{\tau}], \tag{A.1}$$

where the θ_j's are unknown coefficients (to be estimated) and $M_j[t, \ell, \boldsymbol{\tau}]$ is the jth M-spline of order ℓ with knot sequence $\boldsymbol{\tau}$. For cubic splines, $\ell = 4$. If we define the quantities

$$\kappa_{ijk} = \int_{-\infty}^{t_{rk}} \Pr[\mathrm{CD4}(t_{rk}) = i|\text{Infection time} = t] M_j[t, \ell, \boldsymbol{\tau}]\, dt,$$

then the penalized log-likelihood (8) can be written as

$$\log[L_p] = \sum_{k=1}^{K} \sum_{i\in Q(t_{rk})} n_i(t_{rk}) \log\left[\frac{\sum_{j=1}^{J} \kappa_{ijk}\theta_j}{\sum_{i\in Q(t_{rk})} \sum_{j=1}^{J} \kappa_{ijk}\theta_j}\right] - \alpha \sum_{j=1}^{J} \sum_{j'=1}^{J} \theta_j A_{jj'}\theta_{j'} + C,$$

where

$$A_{jj'} = \int_{-\infty}^{t_{rK}} \frac{d^2 M_j[t, \ell, \boldsymbol{\tau}]}{dt^2} \cdot \frac{d^2 M_{j'}[t, \ell, \boldsymbol{\tau}]}{dt^2}\, dt$$

and C is a constant. Since $I(t)$ is known only to a constant, maximization of the log-likelihood must be carried out subject to a constraint. A convenient choice is

$$\sum_{j=1}^{J} \theta_j = 1, \tag{A.2}$$

which is equivalent to requiring that

$$\int_{-\infty}^{t_{rK}} I(t)\, dt = 1.$$

We may ensure that $I(t) \geqslant 0$ by maximizing the likelihood subject to

$$\theta_j \geqslant 0, \quad j = 1, \ldots, J. \tag{A.3}$$

In implementing the forward method described in Section 5, we model $I(t)$ between times t_{r1} and t_{rK} as a linear combination of M-splines of the form (A.1). To avoid making any assumptions on $I(t)$ or its derivatives at either t_{r1} or t_{rK}, for cubic splines we use a knot sequence of the form

$$\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_{J+4}),$$

where $\tau_1 = \tau_2 = \tau_3 = \tau_4 = t_{r1}$, $\tau_5 > t_0$, $\tau_J < t_{rK}$, and $\tau_{J+1} = \tau_{J+2} = \tau_{J+3} = \tau_{J+4} = t_{rK}$. Given our choice of $\boldsymbol{\tau}$, $M_j[t, \ell = 4, \boldsymbol{\tau}] = 0$ for $t < t_{r1}$ for all j. Hence, we may retain the above definitions of κ_{ijk} and $A_{jj'}$ with the modification that the lower limits in the integrals defining κ and A are changed from $-\infty$ to t_{r1}. With these results, we may rewrite (9) as

$$p_i(t_{rk}) = \frac{\displaystyle\sum_{j=1}^{J} \kappa_{ijk}\beta\theta_j + \sum_{i' \leqslant i} T_{i,i'}(t_{rk}, t_{r1})p_{i'}}{\displaystyle\sum_{i \in Q(t_{rk})} \left(\sum_{j=1}^{J} \kappa_{ijk}\beta\theta_j + \sum_{i' \leqslant i} T_{i,i'}(t_{rk}, t_{r1})p_{i'} \right)}. \tag{A.4}$$

As before, the likelihood is maximized subject to constraint (A.2), which here corresponds to

$$\int_{t_{r1}}^{t_{rK}} I(t)\, dt = 1,$$

and (A.3). In terms of the κ_{ijk}'s and θ_j's, the proportion of infection in stages Q^* at time $t_{rk''}$ that occurred between times t_{rk} and $t_{rk'}$ (where $t_{rk} \leqslant t_{rk'} \leqslant t_{rk''}$), given in equation (11) for the forward method, can be expressed as

$$\pi(t_{rk}, t_{rk'}, t_{rk''}) = \frac{\displaystyle\sum_{i \in Q^*} \sum_{j=1}^{J} \sum_{i' \leqslant i} (T_{i,i'}(t_{rk''}, t_{rk'})\kappa_{ijk'} - T_{i,i'}(t_{rk''}, t_{rk})\kappa_{i'jk})\beta\theta_j}{\displaystyle\sum_{i \in Q^*} \left(\sum_{j=1}^{J} \kappa_{ijk''}\beta\theta_j + \sum_{i' \leqslant i} T_{i,i'}(t_{rk''}, t_{r1})p_{i'} \right)}.$$

In the forward method, as a model for $I(t)$ is needed only for the time interval during which sampling is carried out, it may be sufficient to assume a polynomial form for $I(t)$ in the interval (t_{r1}, t_{rK}). Splines also provide a convenient basis set for this type of calculation, as the constraints (A.2) and (A.3) ensure the positivity and normalization of $I(t)$. To fit a Jth-order polynomial (e.g., a polynomial of degree $J - 1$), the expansion (A.1) may be used with order $\ell = J$. A knot sequence of length $2J$ is used, where $\tau_1 = \tau_2 = \cdots = \tau_J = t_{r1}$ and $\tau_{J+1} = \tau_{J+2} = \cdots = \tau_{2J} = t_{rK}$. With this choice for $\boldsymbol{\tau}$, the expansion (A.1) reduces to a polynomial of order J in the interval (t_{r1}, t_{rK}). The definition of κ_{ijk} and expression (A.4) remain unchanged; in this case, the unpenalized likelihood should be used.

Estimation of the parameters in both the forward and backward methods was performed using a FORTRAN program on an IBM compatible PC 486/33. The constrained maximization was performed using the IMSL subroutine dlcong. To ensure that a global maximum was attained, in a number of cases we used several randomly generated starting values; convergence to the same value within numerical accuracy held in every case. We found that the starting point $\theta_j = 1/J$ gave satisfactory results.