BIOMETRICS 55, 507–513
June 1999

Fitting Semi-Markov Models to Interval-Censored Data with Unknown Initiation Times

Glen A. Satten

Division of HIV/AIDS Prevention—Surveillance and Epidemiology,
National Center for HIV, STD, and TB Prevention,
Centers for Disease Control and Prevention, Atlanta, Georgia, U.S.A.
email: gas0@cdc.gov

and

Maya R. Sternberg

Department of Biostatistics, Rollins School of Public Health, Emory University,
Atlanta, Georgia, U.S.A.

SUMMARY. In a semi-Markov model, the hazard of making a transition between stages depends on the time spent in the current stage but is independent of time spent in other stages. If the initiation time (time of entry into the network) is not known for some persons and if transition time data are interval censored (i.e., if transition times are not known exactly but are known only to have occurred in some interval), then the length of time these persons spent in any stage is not known. We show how a semi-Markov model can still be fit to interval-censored data with missing initiation times. For the special case of models in which all persons enter the network at the same initial stage and proceed through the same succession of stages to a unique absorbing stage, we present discrete-time nonparametric maximum likelihood estimators of the waiting-time distributions for this type of data.

KEY WORDS: Chain-of-events data; Doubly censored data; EM algorithm; HIV/AIDS; Initiation time; Interval censoring; Left censoring; Multistage model; Semi-Markov model; Truncation.

1. Introduction

When a semi-Markov model is fit to staged longitudinal data, two features may make an analysis difficult. First, observations on the exact times at which transitions between stages were made may not be available; often we have knowledge only of the stage each person is in at certain (possibly random) observation times. In this case, we say that the transition times are interval censored. Second, the initiation times (i.e., the times of first entry into the network) may also be unknown. In a semi-Markov model, the waiting times in each stage can follow an arbitrary distribution, the only assumption being that waiting times in different stages are independent. Hence, the hazard of moving to a subsequent stage depends on the length of time spent in the current stage. When transition times are known exactly, nonparametric maximum likelihood estimators of the waiting-time distributions are available (Lagakos, Sommer, and Zelen, 1978). If the initiation time is unknown but transition times are observed exactly, then the first observed transition can be used as an initiation time and the methods of Lagakos et al. also apply directly. If the initiation time is known but the transition times are interval censored, De Gruttola and Lagakos (1989) and Sternberg and Satten (1999) have given approaches in discrete time for specific types of networks. If the initiation time is unknown and subsequent transition times are interval censored, an analysis is sometimes still possible if a fixed time scale is available, i.e., if a preinitiation stage can be defined, if there is some time scale (typically calendar time) governing transitions from the preinitiation stage into stage 1, and if it is possible to specify a time 0 at which all persons can be assumed to be in the preinitiation stage. In this case, the methods of Gruttola and Lakakos (1989) or Sternberg and Satten (1999) can be used by considering the preinitiation stage to be the first stage in the network since the initiation time into this stage is known for all persons. However, if such a fixed-time-scale analysis is not available or is impractical, the only currently available methodology is a Markov model that assumes that transition hazards are independent of time-in-stage (Kalbfleisch and Lawless, 1985).

Because a Markov model makes strong parametric assumptions about the shape of the waiting-time distributions, it is important to develop alternative models. In this paper, we show how semi-Markov models can be fit to interval-censored data with unknown initiation times. For simplicity of presentation, we focus primarily on the case where all persons may only move through the same succession of stages to a unique absorbing state; we call such a model a unidirectional model. However, the ideas presented here can be applied to more complicated systems and can be used in parametric or nonparametric analyses.

An example of a unidirectional model occurs in analyzing data on times between successive events that occur in a distinct order. In persons recently infected with human immunodeficiency virus (HIV), infection is first detectable by measuring HIV-RNA by polymerase chain reaction (PCR); then p24 antigen (a core protein of HIV) becomes detectable; then antibodies to HIV are detectable by using an enzyme-linked immunoassay (EIA); then an indeterminate Western blot (an assay used as a confirmatory assay in conjunction with is an EIA) result develops; then a positive Western blot result develops in which the p31 band is missing; subsequently, a Western blot includes the p31 band. The goal of the analysis is to determine the distribution of times between each of these events using data obtained from repeat blood plasma donors.

Data such as these can be analyzed by using a multistage modeling approach, in which a transition between stages occurs at each time an additional quantity becomes detectable (see Table 1 for the definition of the stages). The natural initiation time for this application is the time a person is first infected with HIV; however, this time is rarely known in general and is not known for any person whose data is included in the analysis described in this paper. For persons for whom at least one specimen that is negative on all tests is available, the initiation time could be taken as the time of the earliest specimen (corresponding to study enrollment, although the data we use in this paper were retrospectively ascertained). For these data, the methods of Sternberg and Satten (1999) can be used. However, for persons whose earliest specimen is positive for one or more tests, this is not possible; the initiation time for these persons is not known. Because subsequent transition times are interval censored, it is not possible to use time of first entry into a later stage as an initiation time.

In Section 2, we outline our new approach to multistage data in which transition times are interval censored and initiation times are missing. In Section 3, we modify the results of Sternberg and Satten (1999) to obtain nonparametric estimates of the waiting-time distributions in discrete time for unidirectional models with data such as those described above. The results of applying our methodology to the HIV data discussed above are presented in Section 4. Section 5 contains concluding remarks.

2. Semi-Markov Models with Unknown Initiation Time

Consider data that arise from a J-stage unidirectional semi-Markov model such as the seven-stage model in Figure 1. We consider first the case where the initiation time is known and corresponds to time zero for each individual i. For the ith individual, let t_{ij} denote the time of transition from stage j to $j+1$, so that $k_{ij} = t_{ij} - t_{ij-1}$ is the waiting time in stage j (where t_{i0}, the initiation time, is zero). To establish connection with the methods of Sternberg and Satten (1999), we assume that $f_j(k) = \Pr[k_{ij} = k]$ is a discrete distribution with mass at the integers $0, 1, \ldots, K_j$ independent of i. To allow for loss to follow-up, let $j(i)$ denote the last stage that the ith individual was observed to have reached.

A person who starts in stage 1 and reaches stage J provides interval-censored data on all $J-1$ stage-to-stage transition times. A person who starts in stage 1 but reaches only stage $j(i) < J$ provides interval-censored data on $j(i)-1$ transitions and a right-censored observation on the transition from stage $j(i)$ to stage $j(i)+1$. Hence, each person who starts in stage 1 contributes information to $m_i \equiv \min[j(i), J-1]$ transitions. For the ith person, t_{ij} is known only to be in the interval $A'_{ij} = [l'_{ij}, u'_{ij}]$ if $j < j(i)$ or $A'_{ij} = (l'_{ij}, \infty)$ if $j = j(i) < J$.

If the initiation time is known, then the likelihood of the $f_j(k)$'s given the interval-censored data is

$$\mathcal{L}'_i = \sum_{t_1 \epsilon A'_{i1}} \cdots \sum_{t_{m_i} \epsilon A'_{im_i}} \prod_{j=1}^{m_i} f_j(t_j - t_{j-1}), \qquad (1)$$

Table 1

Discrete-time nonparametric and Markov analyses of plasmapheresis donor data

					Mean waiting time estimates (days)		
					DT-NPMLE[a]		
Stage	PCR	P24 (ever +)	EIA	Western blot	Estimate b_2, b_3, and b_4	$b_2, b_3, b_4 =$ uniform	Markov (95% CI)
8	−	−	−	−	57.0[b]	56.7[b]	55.8 (37.5, 83.5)
2	+	−	−	−	3.4	4.0	5.0 (3.1, 8.1)
3	+	+	−	−	9.2	8.2	5.3 (3.7, 7.7)
4	+	+	+	−	3.8	3.0	3.2 (2.1, 4.8)
5	+	+	+	Indeterminate	4.1	4.3	5.6 (3.8, 8.1)
6	+	+	+	+ (no p31 band)	43.2	43.9	69.5 (39.7, 121.7)
7	+	+	+	+	∞	∞	∞ —

[a] Discrete time nonparametric maximum likelihood estimate.

[b] Waiting times are for stage 8 in the model of Figure 2.

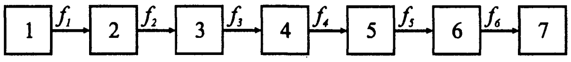

Figure 1. Unidirectional model with seven stages and six waiting-time distributions $f_1, \ldots, f_6$.

where $t_0 \equiv 0$. When $A'_{im_i} = (l_{im_i}, \infty)$, the last sum in (1) must be interpreted as $\Sigma_{t_{m_i} \epsilon A'_{im_i}} f_{m_i}(t_{m_i} - t_{m_i-1}) = S_{m_i}(l_{im_i} - t_{m_i-1})$ and the survival function must be calculated by using $S_j(k) = 1 - \Sigma_{k' \leq k} f_j(k')$ to properly account for the case where the longest waiting time in stage j is from a right-censored observation. As some transition times may be censored into the same interval (i.e., $l'_{j-1} = l'_j$ and $u'_{j-1} = u'_j$ for at least one j), note when writing (1) that $f_j(k) = 0$ whenever $k < 0$. If time is measured continuously, then the sums in equation (1) are replaced by integrals. Sternberg and Satten (1999) considered maximization of (1) subject to the constraints $\Sigma_{k=0}^{K_j} f_j(k) + S_j(K_j) = 1$ by using an EM algorithm or self-consistency approach.

In the case where the initiation time is unknown, let γ_d denote the (unknown) amount of time spent in stage d before study enrollment among people first seen in stage d and assume γ_d has distribution $\rho_d(\gamma_d)$. Let d_i denote the first stage in which the ith observation was seen. Measure t_{id_i} and all subsequent times relative to the time of study enrollment (i.e., the time that the person was first seen) rather than relative to the unknown initiation time. A person first seen in stage d_i contributes information on $q_i = m_i - d_i + 1$ transitions (counting the first transition from stage d_i to $d_i + 1$). Let τ_{ij} denote the jth of these transitions. For $1 \leq j \leq q_i$, let $\delta_{ij} = 1$ if the transition from stage $d_i + j - 1$ to stage $d_i + j$ is observed. Hence, $\delta_{ij} = 0$ implies first that the ith person was lost to follow-up, having last been seen in stage $d_i + j - 1 < J$; thus, if $\delta_{ij} = 0$, it must also be the case that $j = q_i$. Let A_{ij} be the interval that contains τ_{ij}, $1 \leq j \leq q_i$, with $A_{ij} = [l_{ij}, u_{ij}]$ if $\delta_{ij} = 1$ and $A_{ij} = (l_{ij}, \infty)$ if $\delta_{ij} = 0$.

If it is reasonable to assume that the γ_d's are independent of the subsequent transition times (i.e., that the time at which someone was first observed is not predictive of their subsequent progression), then conditional on being seen in stage d_i at first observation, the contribution to the likelihood from the ith person's subsequent data is

$$\mathcal{L}_i = \int_0^\infty d\gamma_{d_i} \sum_{\tau_1 \in A_{i1}} \cdots \sum_{\tau_{q_i} \in A_{iq_i}} \rho_{d_i}(\gamma_{d_i}) f_{d_i}(\tau_1 + \gamma_{d_i}) \times \prod_{j=2}^{q_i} f_{j+d_i-1}(\tau_j - \tau_{j-1}).$$

This assumption holds, e.g., if the unknown initiation times follow a time-dependent Poisson process. If we define

$$b_d(\tau) = \int_0^\infty d\gamma\, \rho_d(\gamma) f_d(\tau + \gamma),$$

then we can write $\mathcal{L}_i$ as

$$\mathcal{L}_i = \sum_{\tau_1 \in A_{i1}} \cdots \sum_{\tau_{q_i} \in A_{iq_i}} b_{d_i}(\tau_1) \times \prod_{j=2}^{q_i} f_{j+d_i-1}(\tau_j - \tau_{j-1}). \tag{2}$$

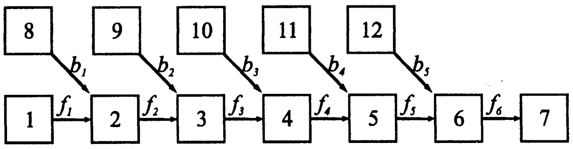

Figure 2. Expanded network for the seven-stage unidirectional model of Figure 1. Five additional stages (stages 8–12) have been added for persons first seen in stages 1–5 at study enrollment.

Note that (2) has the same form as (1) except that all times are measured relative to the time of first observation, f_{d_i} has been replaced by b_{d_i}, q_i replaces m_i, and the numbering of stages has been altered.

Although there is some relationship between the b_j's and f_j's, we suggest the following approach: Estimate each distribution b_j as a separate nuisance function. To estimate the b_j's as separate nuisance functions, we replace the model of Figure 1 by that in Figure 2. Only persons whose initiation times are known are started in stage 1. All other persons first seen in stage d with unknown initiation times are taken to enter the network in stage $d + J$ with known initiation time given by their time of study enrollment. The key observation is that the likelihood obtained using this model is given by (2). Hence, replacing the model of Figure 1 with that of Figure 2 provides a framework for estimating the b_j's and f_j's, which is still a multistage model, although a somewhat more complicated one; the advantage is that all persons may be considered to have known initiation time. The network for any multistage model could be expanded by adding additional stages for observations first seen an unknown time after their entry into the network. We will refer to the network of stages after addition of these extra stages as the expanded version of the original network. Because the mechanism generating observation times is independent of the initiation and transition times, the exclusion of individuals first seen in stage j from estimation of f_j does not result in biased estimation of f_j, although some efficiency may be lost. However, these individuals can only contribute to estimation of f_j if the distribution of their initiation times is known, while our proposal leads to valid inference when such information is absent or unreliable.

An important feature of the expanded version of a unidirectional model is that, once the initial stage for an observation is specified, the data for that observation follow a single predetermined sequence of stages. This feature allows the methods of Sternberg and Satten (1999) to be easily generalized to give nonparametric estimates of the f_j's (as well as the b_j's) in discrete time. This is considered in Section 3. Note also that there is no point in adding an additional stage for persons first seen in stage $J - 1$ at study enrollment, as these individuals do not contribute information to the estimation of any of $f_1, \ldots, f_{J-1}$.

An alternate, but closely related, approach can be used if additional assumptions hold. If the unknown initiating times among persons first seen in stage d can be taken to follow a Poisson process, then the time of their first observed transition (from stage d into stage $d + 1$) also follows a Poisson process. If the censoring intervals A_{i1} for these individuals are all small enough that the intensity of this process can be taken to be constant, then the marginal distribution of t_{i1} may be taken to be uniform in A_{i1}. If we take b_d to be the uniform distribution, then the contribution to the likelihood from these observations is proportional to (2).

3. Nonparametric Estimation for Unidirectional Models

In this section, we generalize the methods of Sternberg and Satten (1999) to allow nonparametric estimation of the waiting-time distributions in discrete time for unidirectional models when data transition times are interval censored and initiation times are missing. If there are J stages in the original model, the extended model described in Section 2 has $2J-3$ waiting-time distributions. Let

$$\phi_j(k) = \begin{cases} f_j(k) & 1 \le j \le J-1 \\ b_{j-J+1}(k) & J \le j \le 2J-3, \end{cases}$$

and let $\Psi_j(k) = 1 - \Sigma_{k' \le k}\, \phi_j(k')$ be the survival function for the distribution ϕ_j. For the model in Figure 2, ϕ_j is the waiting-time distribution in stage $j + I[j \ge J]$, where $I[C] = 1$ if C is true and 0 otherwise. Define a vector $p_i = (p_{i1}, p_{i2}, \dots, p_{iq_i})$, which has kth component given by the subscript of ϕ corresponding to the distribution that governs the kth transition for the ith person; two examples follow. Let $\mathcal{R}_j$ denote the set of all indices i for which $p_{ij'} = j$ for some j'. Then $i \in \mathcal{R}_j$ implies that the ith observation contributes information on ϕ_j. Recall that d_i and $j(i)$ are the first and last stages in which the ith person was seen, $m_i \equiv \min[j(i), J-1]$, and $q_i \equiv m_i - d_i + 1$.

To clarify our notation, consider two hypothetical persons who follow the model shown in Figures 1 and 2. Person 1 was in stage 4 at study enrollment and was followed to stage 7. For this person, $d_1 = 4$, $j(1) = 7$, $m_1 = 6$, so there is information on $q_1 = 3$ transitions. Because this person had an unknown time in stage 4 prior to study enrollment, he enters the network in stage 11. For the model in Figure 2, the three waiting-time distributions this observation contributes information to are $\phi_{10} = b_4$, ϕ_5, and ϕ_6, so that $p_1 = (10, 5, 6)$. Finally, no transition times are right censored, so $\delta_{11} = \delta_{12} = \delta_{13} = 1$. The time of the transition from stage 11 to stage 5 is in the interval A_{11}, the time of the transition from stage 5 to stage 6 is in the interval A_{12}, and the time of the transition from stage 6 to stage 7 is in A_{13}. The values of A_{12}, A_{13}, A_{13}, δ_{11}, δ_{12}, and δ_{13} make up the data D_1. Person 2 had a known initiation time and hence enters stage 1 at study enrollment but had reached only stage 4 at the last time she was seen. In this case, $d_2 = 1$ and $m_2 = 4$, so there is information on $q_2 = 4$ transitions. For the model of Figure 2, the four waiting-time distributions this observation contributes information to are ϕ_1, ϕ_2, ϕ_3, and ϕ_4, so $p_2 = (1, 2, 3, 4)$. The first three transition times are interval censored, so $\delta_{21} = \delta_{22} = \delta_{23} = 1$, but the transition time from stage 4 to stage 5 is right censored, so $\delta_{24} = 0$. The intervals A_{2j} and right-censoring indicators δ_{2j} make up the data D_2. Note that we do not consider the transitions from stage 5 to stage 6 or stage 6 to stage 7 for this person, as these transition times are censored into the interval $[0, \infty)$ and hence contribute no information.

With this notation, the ith person's contribution to the likelihood can be written as

$$\mathcal{L}_i = \sum_{\tau_1 \in A_{i1}} \cdots \sum_{\tau_{q_i} \in A_{iq_i}} \phi_{p_{i1}}(\tau_1) \prod_{j=2}^{q_i} \phi_{p_{ij}}(\tau_j - \tau_{j-1}); \tag{3}$$

the total likelihood is given by $\mathcal{L} = \Pi_{i=1}^{n} \mathcal{L}_i$. Following the results in Sternberg and Satten (1999), the likelihood $\mathcal{L}$ can be maximized using the EM algorithm. For each observation i and transition j for which $\delta_{ij} = 1$, define

$$\tilde{f}_{p_{ij}}(k \mid D_i) = \begin{cases} \dfrac{\sum\limits_{\tau_1 \epsilon A_{i1}} \cdots \sum\limits_{\tau_{q_i} \epsilon A_{iq_i}} I(\tau_j - \tau_{j-1} = k) \prod\limits_{j'=1}^{q_i} \phi_{p_{ij'}}(\tau_{j'} - \tau_{j'-1})}{\sum\limits_{\tau_1 \epsilon A_{i1}} \cdots \sum\limits_{\tau_{q_i} \epsilon A_{iq_i}} \prod\limits_{j'=1}^{q_i} \phi_{p_{ij'}}(\tau_{j'} - \tau_{j'-1})} \\ \quad \text{if } \max(0, l_{ij} - u_{ij-1}) \le k \le u_{ij} - l_{ij-1}, \\ 0 \quad \text{otherwise,} \end{cases} \tag{4}$$

where we adopt the convention that $k_0 = l_{i0} = u_{i0} := 0$. If $i \in \mathcal{R}_j$ and the ith observation is not right censored in stage j, then the value of $\tilde{f}_j(k \mid D_i)$ is the conditional probability that the ith observation has waiting time k in stage j, given the data D_i.

To account for right-censored observations, for each observation i and transition j for which $\delta_{ij} = 0$ and $q_i > 1$, define

$$\tilde{g}_{p_{iq_i}}(k \mid D_i) = \begin{cases} \dfrac{\sum\limits_{\tau_1 \epsilon A_{i1}} \cdots \sum\limits_{\tau_{q_i-1} \epsilon A_{iq_i-1}} I(\tau_{j-1} = l_{iq_i} - k) \Psi_{p_{iq_i}}(l_{iq_i} - \tau_{q_i-1})}{\sum\limits_{\tau_1 \epsilon A_{i1}} \cdots \sum\limits_{\tau_{q_i-1} \epsilon A_{iq_i-1}} \Psi_{p_{iq_i}}(l_{iq_i} - \tau_{q_i-1}) \prod\limits_{j'=1}^{q_i} f_{p_{ij'}}(\tau_{j'} - \tau_{j'-1})} \\ \quad \times \dfrac{\prod\limits_{j'=1}^{q_i} \phi_{p_{ij'}}(\tau_{j'} - \tau_{j'-1})}{\sum\limits_{\tau_1 \epsilon A_{i1}} \cdots \sum\limits_{\tau_{q_i-1} \epsilon A_{iq_i-1}} \Psi_{p_{iq_i}}(l_{iq_i} - \tau_{q_i-1}) \prod\limits_{j'=1}^{q_i} f_{p_{ij'}}(\tau_{j'} - \tau_{j'-1})} \\ \quad \text{if } l_{iq_i} - l_{iq_i-1} \le k \le l_{iq_i} - u_{iq_i-1}, \\ 0 \quad \text{otherwise.} \end{cases} \tag{5}$$

For the case $q_i = 1$ and $\delta_{i1} = 0$, let $\tilde{g}_{p_{i1}}(k \mid D_i) = I(k = l_{i1})$. If $i \in \mathcal{R}_j$ and the ith observation is right censored in stage j, then $\tilde{g}_j(k \mid D_i)$ is the distribution of right-censoring times for the ith observation in stage j. Calculation of the values $\tilde{f}_j(k \mid D_i)$ and $\tilde{g}_j(k \mid D_i)$ make up the E step of the EM algorithm. Note that a different set of $\tilde{f}_j(k \mid D_i)$'s and $\tilde{g}_j(k \mid D_i)$'s are calculated for observations that have different values of p_i. For example, for observation 1 described above, we would calculate $\tilde{f}_{10}(k \mid D_1)$, $\tilde{f}_5(k \mid D_1)$, and $\tilde{f}_6(k \mid D_1)$; for observation 2, we would calculate $\tilde{f}_1(k \mid D_2)$, $\tilde{f}_2(k \mid D_2)$, $\tilde{f}_3(k \mid D_2)$, and $\tilde{g}_4(k \mid D_2)$.

The M step of the EM algorithm consists of constructing the Kaplan–Meier estimator obtained using the conditional probabilities $\tilde{f}_j(k \mid D_i)$ and $\tilde{g}_j(k \mid D_i)$. For each j from 1 to $2J-3$, define $n_{jk} = \Sigma_{i \in \mathcal{R}_j} \tilde{f}_j(k \mid D_i)$, $N_{jk} = \Sigma_{k' \ge k}\, n_{jk}$, $c_{jk} = \Sigma_{i \in \mathcal{R}_j} \tilde{g}_j(k \mid D_i)$, and $C_{jk} = \Sigma_{k' \ge k}\, c_{jk}$. For each $1 \le j \le 2J-3$, the Kaplan–Meier estimator of Ψ_j and ϕ_j obtained by using the fractional masses $\tilde{f}_j(k \mid D_i)$ and $\tilde{g}_j(k \mid D_i)$ are given by

$$\Psi_j(k) = \prod_{k' \leq k} \left(1 - \frac{n_{jk}}{N_{jk} + C_{jk}}\right), \qquad 0 \leq k \leq K_j, \qquad (6)$$

and

$$\phi_j(k) = \Psi_j(k-1) - \Psi_j(k), \qquad 0 \leq k, \leq K_j, \qquad (7)$$

where $\Psi_j(-1) \equiv 1$. To implement the EM algorithm, an initial choice for the ϕ_j's must be made: denote these by ϕ_j^0. Then the required values of $\tilde{f}_j(k \mid D_i)$ and $\tilde{g}_j(k \mid D_i)$ are calculated by using (4) and (5). Then new values of the ϕ_j^1's are obtained using (6) and (7). The process is repeated until $\max_{1 \leq j \leq 2J-3} \Sigma_{k=0}^{K_j} |\phi_j^{r+1}(k) - \phi_j^r(k)|$ is less than some preset tolerance ϵ (we used $\epsilon = 10^{-8}$ for the analyses in Section 4).

If some or all of the b_j's are to be assumed uniform, as described at the end of Section 2, then the initial distributions chosen for the corresponding ϕ_j's must be the uniform distribution. In the subsequent EM steps, these distributions are then never updated, so the values of $\tilde{f}_j(k \mid D_i)$ and $\tilde{g}_j(k \mid D_i)$ for these values of j need not be calculated.

The computational burden of calculating the $\tilde{f}_j(\cdot \mid D_i)$'s and $\tilde{g}_j(\cdot \mid D_i)$'s by using (4) and (5) grows exponentially with the number of stages J. However, both $\tilde{f}_j(\cdot \mid D_i)$ and $\tilde{g}_j(\cdot \mid D_i)$ can be calculated recursively. The recursion relations for these calculations are a simple modification of those given in Sternberg and Satten (1999) and are given in Sternberg (1997). Using these recursion relations, the computational burden in calculating the $\tilde{f}_j(\cdot \mid D_i)$'s and $\tilde{g}_j(\cdot \mid D_i)$'s grows linearly with J.

As first noticed by Turnbull (1976), the nonparametric maximum likelihood estimator for interval-censored data may be undefined in some regions. For a two-stage unidirectional model, Turnbull gave an algorithm for identifying such undefined regions. For the discrete-time EM algorithm, undefined regions are manifested as intervals in which the form of the waiting-time distribution depends on the initial distributions ϕ_j^0 used to start the EM algorithm. These regions can be identified by starting the EM algorithm using different initial distributions ϕ_j^0 and then looking for intervals where the total mass is conserved but is distributed differently inside the interval (after first confirming that the likelihood is the same in all cases). Further discussion can be found in Sternberg and Satten (1999).

4. Example: Time-Course of Events in Early HIV Infection

An example of our proposal is provided by an analysis of the time-course of events in early HIV infection (Busch et al., 1996). We obtained data on 50 seroconverting plasmapheresis donors, each of whom made a series of plasma donations after becoming infected with HIV. Each stored sample from each donor's donations were tested for presence of HIV-RNA by PCR, the presence of p24 antigen (a core protein of HIV) by an EIA, presence of antibodies to HIV using an HIV-1/2 combination EIA (a highly sensitive version of the assay used for screening for HIV), and HIV positivity by using a Western blot assay (which detects antibodies to specific viral proteins and is typically used as a confirmatory assay when determining HIV status). For the Western blot results, we distinguished between negative samples, indeterminate samples, and positive samples that did and those that did not contain antibodies to the viral protein p31. (For details on the tests, see Busch et al. [1997].) Note that plasma products are heat treated and that these donations posed no danger of HIV infection to the recipients (although persons with known HIV infection are excluded as plasma donors).

As the sequence of events described above occurs in a distinct order, these data can be described using the seven-stage unidirectional model shown in Figure 1. The definitions of the stages are given in the first five columns of Table 1. The natural initiation time for this problem is the moment of infection, which would correspond to entry into stage 1. To apply the results of Section 2, we added extra stages to account for the people with missing initiation times; the extended network of stages that results is shown in Figure 2. Since the time of infection is unknown for all persons in our study, the distribution f_1 is not estimable, and all persons who were seen prior to development of a positive HIV-RNA begin in stage 8. For this reason, the waiting time from a negative result on all tests to HIV-RNA positivity that is reported in Table 1 corresponds to the mean time from study enrollment (i.e., the time of the earliest available specimen that is negative on all tests) to development of a positive HIV-RNA result rather than the mean time from infection to an HIV-RNA positivity. Note that some individuals categorized into stage 8 at study enrollment may not yet be HIV infected. This affects the interpretation of b_1 only, not the mechanics of our proposed method.

For most HIV-infected people, after an initial rise, p24 antigen levels again drop to levels that cannot be detected using the standard assay. In this analysis, a person is considered p24 antigen positive if they have ever had a previous positive p24 antigen test result with a positive HIV-RNA result. One person was seen initially PCR positive with no other tests positive and, subsequently, 26 days later with positive HIV-1/2 EIA and a positive Western blot (but with the p31 band missing) but a negative p24 antigen result. It was assumed that the p24 antigen developed and was lost between these two observations. A second person who had an observation that was PCR and p24 antigen positive, EIA negative but with a very weak but discernible p24 band in their Western blot (corresponding to an indeterminate Western blot result) was considered to be in stage 3 for this analysis.

We first fit a Markov model to these data. However, the results summarized in Table 1 indicate the length of waiting times varies by over an order of magnitude; in this situation, the Markov assumption can result in misleading conclusions. Hence, we sought a nonparametric estimate of the waiting-time distributions. While the methods of Sternberg and Satten (1999) can be used for the persons initially in stage 8, only 25 of the 50 people had an initial specimen that was negative on all tests.

The number of persons entering the extended network in stages 8, 9, 10, and 11 were 25, 6, 17, and 1, respectively. No persons were observed to be initially in stage 5, so stage 12 was removed from the model and b_5 was not estimated. Data from one person was omitted from the analysis because they were first seen in stage 6. The average censoring interval width was 24 days (range, 2–231 days). A total of 245 transitions were observed: 103 transition times were censored into intervals containing no other transition time for the same individual (including 36 instances of right censoring), 58 were censored into intervals containing two transition times for the same individual, 39 were censored into intervals containing three transition times for the same individual, 24 were censored into intervals containing four transition times for the same individual, 15 were censored into intervals containing five transition times for the same individual, and in one instance an individual had six transition times censored into the same interval. The number of observations contributing to estimation of $\phi_2, \ldots, \phi_{10}$ (i.e., the cardinality of $\mathcal{R}_2, \ldots, \mathcal{R}_{10}$) was 25, 31, 47, 48, 45, 25, 6, 17, and 1, respectively.

The EM algorithm was started using a uniform distribution with $K_j = 300$ days for each stage. To identify regions where the waiting-time distributions were not defined, we also performed analyses in which the initial distribution for all but one stage was the discrete uniform distribution and used the discrete triangular distribution with mode zero for the remaining stage. We found that the form of $\phi_{10} = b_4$ is not defined in the interval $[4, 6]$, although the total mass assigned to that interval is.

Because plasma donors make frequent donations, the intervals A_{i1} for donors first seen in stages 2, 3, and 4 were short relative to the time scale on which changes in the intensity of new HIV infections occurs (the maximum lengths were 14, 9, and 5 days, respectively). Hence, we also estimated the waiting-time distributions assuming that b_2, b_3, and b_4 were uniform. For this analysis, we find that all f_j's are defined at integer values of time.

The mean stage occupation times for the Markov and nonparametric analyses are compared in Table 1. Confidence intervals for the mean occupation times for the Markov model were calculated by inverting the observed information matrix evaluated at the maximum likelihood estimates. The estimated mean waiting times for the first five stages are fairly similar in the three models; however, the mean waiting time in the last stage estimated for the Markov model is more than half again the mean estimated for the two nonparametric methods. The waiting-time distributions $b_1, f_2, \ldots, f_6$, obtained when b_2, b_3, and b_4' are estimated as well, are shown in Figure 3, along with the corresponding distributions obtained by fitting the Markov model. The waiting-time distributions obtained assuming b_2, b_3, and b_4 are uniform were similar (results not shown).

Because the HIV epidemic in the U.S. had a fairly well-defined onset, it is possible to analyze these data using a fixed time scale, i.e., defining stage 1 to correspond to being uninfected and then assuming that all persons were in stage 1 at the beginning of the epidemic (e.g., January 1, 1980, could be used as such a date). Taking this time as the initiation time for all persons, we could use the methods of Sternberg and Satten (1999) to analyze these data. However, this in essence requires a reconstruction of the epidemic curve (i.e., distribution of times of infection) using only the 50 seroconverters in this study. Because the number of mass points required to describe the waiting time in stage 0 is so much larger than the number of study participants, we can expect little borrowing of strength (i.e., the times of first PCR positivity for the 25 persons with data available before PCR seroconversion will be of little help in determining the calendar dates of PCR

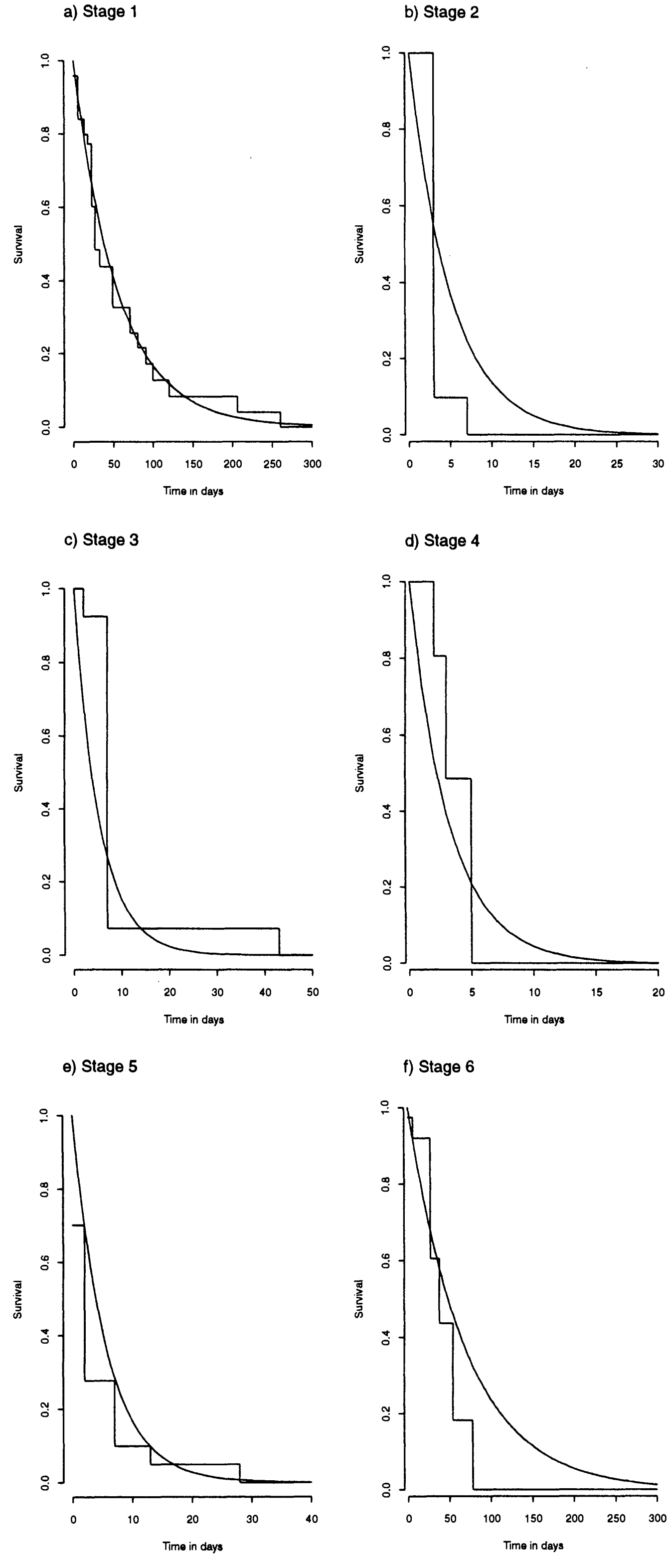

Figure 3. Waiting-time distributions for stages 8, 2, 3, 4, 5, and 6 estimated using the nonparametric method (step functions) and using a Markov model (smooth curves).

seroconversion among persons whose first sample was PCR positive).

5. Discussion

Longitudinal multistage data are inherently complex; when transition times are interval censored and initiation times unknown, some assumptions are required to extract any information from the data. When initiation times are unknown, and no fixed-time-scale analysis is available, the only analysis tool previously available was to fit a Markov chain. Because model fitting with data such as we have considered here is sometimes the only way to begin exploring and visualizing the data, it is important to have models that allow a flexible view of the data. In addition, alternate methodology must be available to properly handle instances in which the Markov assumption is not appropriate. The nonparametric estimates we have developed can help explore the data and can confirm the results of a Markov analysis or indicate important features that the Markov analysis missed.

Even when a fixed-time-scale analysis is available, the new approach presented here may be preferable. One such case was discussed in Section 4; if the fixed-time-scale analysis requires calendar time analysis of a small amount of data gathered over a long period of time, there may be no gain in statistical efficiency over our new approach (note that computational efficiency decreases with both the number of mass points estimated and the number of stages). Another case is when it is known or suspected that the distribution of initiation times varies by the stage first observed. As a hypothetical example, consider the data presented in Section 4. If persons first seen with PCR-negative results were from plasma centers in San Francisco while those first seen with PCR-positive results were from the Midwest, then the fixed-time-scale analysis would lead to misleading results, as the distribution of times of infection for San Francisco is known to differ from that in the Midwest. The methodology of Sections 2 and 3 would nonetheless remain valid in this case. However, when a fixed-time-scale analysis is available, appropriate, and does not require estimating an initiation-time distribution that is not well characterized by the amount of data available, it constitutes a more efficient use of the available data than the methodology proposed here.

In this paper, we have considered unidirectional semi-Markov models in detail. However, auxiliary stages of the type we considered for the unidirectional model can be added to any network, which allows construction of a likelihood when including data from persons whose initiation times are unknown, even when transition times are interval censored.

ACKNOWLEDGEMENT

MRS was supported by NIH grant T32-AI07442.

Résumé

Dans un modèle semi-markovien, la possibilité d'avoir une transition entre étapes dépend du temps écoulé dans l'étape courante, mais est indépendants du temps passé dans les autres étapes. Si le temps initial (temps d'entrée dans le processus) n'est pas connu pour certains individus, et si les temps de transition sont des données censurées par intervalle (c'est-à-dire si les temps de transition ne sont pas connu exactement, mais seulement pour appartenir à un intervalle), alors le temps que ces individus passent dans une étape quelconque n'est pas connu. Nous montrons comment ajuster un modèle semi-markovien à des données censurées par intervalle avec des temps initiaux inconnus. Pour le cas particulier des modèles dans lesquels tous les individus entrent dans le processus à la même étape initiale, et progressent par la même successions d'étapes vers une unique étape absorbante, nous présentons des estimateurs non paramétriques du maximum de vraisemblance à temps discret pour les distributions des temps d'attente.

References

Busch, M. P., Herman, S. A., Henrard, D. R., et al. (1996). Time course and kinetics of HIV viremia during primary infection (abstract #38). Third National Conference on Human Retroviruses and Related Opportunistic Infections, Washington, D.C.

De Gruttola, V. and Lagakos, S. W. (1989). Analysis of doubly-censored survival data, with application to AIDS. *Biometrics* **45,** 1–11.

Kalbfleisch, J. D. and Lawless, J. F. (1985). The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association* **80,** 863–871.

Lagakos, S. W., Sommer, C. J., and Zelen, M. (1978). Semi-Markov models for partially censored data. *Biometrika* **65,** 311–318.

Sternberg, M. (1997). Discrete time nonparametric estimation for chain of events data subject to interval censoring and truncation. Thesis, Emory University, Atlanta, Georgia.

Sternberg, M. and Satten, G. A. (1999). Discrete-time nonparametric estimation for chain-of-events data subject to interval censoring and truncation. *Biometrics* **55,** 179–187.

Turnbull, B. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B* **38,** 290–295.

Received September 1997. *Revised May* 1998.
Accepted July 1998.