Biostatistics (2007), **8**, 2, *pp.* 252–264
doi:10.1093/biostatistics/kxl006
Advance Access publication on June 1, 2006

Statistical methods for panel data from a semi-Markov process, with application to HPV

MINHEE KANG*, STEPHEN W. LAGAKOS

Department of Biostatistics, Harvard University School of Public Health, 655 Huntington Avenue, Boston, MA 02115, USA
mkang@hsph.harvard.edu

SUMMARY

Continuous-time, multistate processes can be used to represent a variety of biological processes in the public health sciences; yet the analysis of such processes is complex when they are observed only at a limited number of time points. Inference methods for such panel data have been developed for time homogeneous Markov models, but there has been little research done for other classes of processes. We develop likelihood-based methods for panel data from a semi-Markov process, where transition intensities depend on the duration of time in the current state. The proposed methods account for possible misclassification of states. To illustrate the methods, we investigate a three- and a four-state models in detail and apply the results to model the natural history of oncogenic genital human papillomavirus infections in women.

Keywords: Human papillomavirus; Misclassification; Multistate process; Natural history.

1. INTRODUCTION

Continuous-time, multistate stochastic processes provide a useful framework for many studies of event history data (Commenges, 1999, 2002; Hougaard, 1999; Andersen and Keiding, 2002). Most research in continuous-time, discrete-state processes has been probabilistic, and inference about processes using independent realizations from a group of individuals has been based almost entirely on settings where sample paths are continuously observed (cf. Andersen and Borgan, 1985).

However, in many instances, observations consist of the states of the individual processes at discrete time points, with no information about the types and times of events between observation times. For example, when state transitions of a process are silent events—such as the onset of an early stage of a disease before symptoms—the sample paths are observed infrequently, often resulting from diagnostic tests given during patient visits to their caregivers. Inference methods for such panel data from multistate processes have been limited. Kalbfleisch and Lawless (1985) proposed methods for the analysis of panel data for Markov models with time homogeneous transition intensities; Kay (1986), Andersen (1988), and Gentleman *and others* (1994) have applied these methods to cancer, diabetes, and HIV. Inference based

*To whom correspondence should be addressed.

© The Author 2006. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oxfordjournals.org.

Downloaded from https://academic.oup.com/biostatistics/article/8/2/252/230485 by guest on 21 August 2022

on Markov models in such settings is greatly simplified, because the discrete-time process observed at prespecified time points forms a Markov process.

In many applications, however, the Markov assumption is not appropriate because the transition intensities depend on the elapsed time in the current state. For instance, in modeling the natural history of human papillomavirus (HPV), which is known to cause almost all cervical cancers, the Markov assumption would not account for the strong association between infection duration and progression to cervical abnormality (Stoler, 2000). Sexually acquired genital HPV infections in women are often transient and recurring, and are usually resolved by the host prior to the onset of symptoms. However, epidemiology studies show that when an infection by certain HPV types persists for a long period of time, it can eventually lead to clinical conditions such as high-grade cervical intraepithelial neoplasia (CIN 2 and 3) and, ultimately, cervical cancer (Stoler, 2000). When the underlying process is not a time homogeneous Markov process, the observed discrete-time process will in general have a complex structure. A further complication that can arise is that the observations of the states of the process are subject to misclassification.

Motivated by the HPV studies, this paper considers inference for continuous-time semi-Markov processes in settings where sample paths of individuals are observed only at a finite number of prespecified times, possibly with misclassification errors in the observed states. We show that evaluation of likelihood functions can be greatly simplified when the transition intensity from at least one of the states of the underlying process is time homogeneous. Section 2.1 introduces the underlying and the observed processes and outlines the general approach to likelihood methods. Section 3 presents the likelihood contributions in detail for a nonprogressive three-state process. Section 4 applies the proposed methods to a recent HPV trial and a simulated study.

2. Inference for a K-state semi-Markov process

2.1 *Underlying process and observed data*

Suppose that $X(\cdot) = \{X(t), t \geqslant 0\}$ denotes a continuous-time process with K states, denoted $1, \ldots, K$. Let the random variables $\sigma_0, \sigma_1, \ldots$ denote the initial and subsequent consecutive states occupied by the process, and let τ_n represent the sojourn time between the $(n-1)$th and nth states, for $n = 1, \ldots$. Thus, $X(\cdot)$ is equivalent to

$$\sigma_0, \tau_1, \sigma_1, \ldots, \tau_n, \sigma_n, \ldots.$$

The process is semi-Markov if the sequence $\{\sigma_0, \sigma_1, \ldots\}$ of consecutively occupied states forms a simple Markov chain, and the sojourn times τ_n between consecutively occupied states are independent random variables with distributions that depend only on the adjoining states (cf. Cox and Miller, 1977). The probabilistic properties of a semi-Markov process can be characterized by the transition intensities, or cause-specific hazard functions, among states. Suppose that the process enters state i at its nth transition. Then the transition intensity functions out of state i, say $\lambda_{ij}(\cdot)$, are

$$\lambda_{ij}(t) = \lim_{h \downarrow 0} \frac{P[\tau_{n+1} < t + h, \sigma_{n+1} = j \mid \tau_{n+1} \geqslant t, \sigma_n = i]}{h},$$

where t denotes the elapsed time from entrance into state i, for $i, j = 1, \ldots, K$, $i \neq j$, and $n = 1, 2, \ldots$. Such dependence on the duration (t) in the current state (σ_n) makes semi-Markov processes distinct from Markov processes.

The probabilistic properties of semi-Markov processes have been studied extensively (Cox and Miller, 1977), and inferences for settings where sample paths are continuously observed or right censored can be made by extensions of methods for ordinary failure time data (cf. Lagakos *and others*, 1978). Inferences for unidirectional, or progressive, processes that lead to interval-censored data have also been developed

Downloaded from https://academic.oup.com/biostatistics/article/8/2/252/230485 by guest on 21 August 2022

(Sternberg and Satten, 1999). However, we consider settings where the independent realizations of the process $X(\cdot)$ corresponding to different subjects are observed only at a finite number of prespecified time points and the process may be bidirectional (nonprogressive). Consider the values of $X(\cdot)$ at the fixed times $0 = v_0 < v_1 < v_2 < \cdots < v_M$, and let $\mathbf{X} = (X_0, X_1, \ldots, X_M)$, where $X_m = X(v_m)$. Despite the simple probabilistic form of the underlying semi-Markov process $X(\cdot)$, the joint distribution of $X_0, X_1, \ldots, X_M$ is in general complex because the time points $v_0, v_1, \ldots, v_M$ do not correspond to transition times between states of the process. If the process is not progressive, the states visited in the sample path are not determined by the knowledge of the states occupied at visit times.

Inferences about the parameters $\lambda_{ij}(\cdot)$ are further complicated by misclassification errors. Rather than $\mathbf{X}$, suppose that the observation for a subject is given by $\mathbf{Y} = (Y_0, Y_1, \ldots, Y_M)$, where $Y_m \in \{1, 2, \ldots, K\}$ denotes X_m subject to misclassification error. We assume that the misclassification error probabilities satisfy the conditional independence assumption

$$P(Y_0, Y_1, \ldots, Y_M | X_0, X_1, \ldots, X_M) = \prod_{i=0}^{M} P(Y_i | X_i). \tag{2.1}$$

That is, conditional upon the true values of $X(\cdot)$ at the visit times, the distribution of Y_m depends only on the value of $X(\cdot)$ at v_m. We denote these error probabilities by $\alpha_{ik} = P(Y_m = k | X_m = i)$.

2.2 *Likelihood function*

The likelihood contribution for an individual can be written as

$$\begin{aligned} L = P(\mathbf{Y} = \mathbf{y}) &= \sum_{\mathbf{x}} P(\mathbf{Y} = \mathbf{y} | \mathbf{X} = \mathbf{x}) P(\mathbf{X} = \mathbf{x}) \\ &= \sum_{\mathbf{x}} P(Y_0 = y_0 | X_0 = x_0) P(Y_1 = y_1 | X_1 = x_1) \cdots P(Y_M = y_M | X_M = x_M) P(\mathbf{X} = \mathbf{x}) \\ &= \sum_{\mathbf{x}} c_{\mathbf{xy}} P(\mathbf{X} = \mathbf{x}), \end{aligned} \tag{2.2}$$

where $x_m, y_m \in \{1, \ldots, K\}$, $c_{\mathbf{xy}} = \prod_{m=0}^{M} \alpha_{x_m, y_m}$, and the summation is over all possible state sequences.

If the transition intensities from at least one of the K states can be assumed to be time homogeneous—that is, $\lambda_{ij}(t) = \lambda_{ij}$ for $j = 1, \ldots, K$ if and only if $i \in \mathcal{C}$, where $\mathcal{C}$ is a nonempty subset of $\{1, \ldots, K\}$—then $P(\mathbf{X} = \mathbf{x})$ in the evaluation of (2.2) is nicely simplified. A consequence of this assumption is that for any times $0 \leqslant t_0 < t_1 < \cdots$ (whether or not these correspond to visit times) and any $x_m \in \mathcal{C}$ for $m = 0, 1, \ldots,$

$$\begin{aligned} &P[X(t_{m+1}) = x_{m+1}, X(t_{m+2}) = x_{m+2}, \ldots | X(t), 0 \leqslant t \leqslant t_m, X(t_m) = x_m] \\ &\quad = P[X(t_{m+1}) = x_{m+1}, X(t_{m+2}) = x_{m+2}, \ldots | X(t_m) = x_m, \phi(t_m)] \\ &\quad = P[X(t_{m+1} - t_m) = x_{m+1}, X(t_{m+2} - t_m) = x_{m+2}, \ldots | X(0) = x_m], \end{aligned} \tag{2.3}$$

where $\phi(t_m)$ denotes the time at which the process last entered state x_m prior to t_m. The semi-Markov nature of $X(\cdot)$, which ensures that the future of the process for times greater than $\phi(t_m)$ depends on the history of the process only through $\phi(t_m)$ and $x_m = X(\phi(t_m))$, combined with the memoryless property of sojourn times in state x_m yields this stationarity property of the process at times when the process is in

Downloaded from https://academic.oup.com/biostatistics/article/8/2/252/230485 by guest on 21 August 2022

a state belonging to $\mathcal{C}$. To illustrate how (2.3) can greatly simplify the evaluation of the finite-dimensional distributions of $X(\cdot)$, suppose that $\mathcal{C} = \{1\}$. Then, for example,

$$\begin{aligned}
&P[X(t_0) = 1, X(t_1) = 3, X(t_2) = 1, X(t_3) = 2, X(t_4) = 1, X(t_5) = 2, X(t_6) = 4, X(t_7) = 4] \\
&\quad = P[X(t_5) = 2, X(t_6) = 4, X(t_7) = 4 | X(t_0) = 1, X(t_1) = 3, \ldots, X(t_4) = 1] \\
&\qquad \times P[X(t_3) = 2, X(t_4) = 1 | X(t_0) = 1, X(t_1) = 3, X(t_2) = 1] \\
&\qquad \times P[X(t_1) = 3, X(t_2) = 1 | X(t_0) = 1] \cdot P[X(t_0) = 1] \\
&\quad = P[X(t_5 - t_4) = 2, X(t_6 - t_4) = 4, X(t_7 - t_4) = 4 | X(0) = 1] \\
&\qquad \times P[X(t_3 - t_2) = 2, X(t_4 - t_2) = 1 | X(0) = 1] \\
&\qquad \times P[X(t_1 - t_0) = 3, X(t_2 - t_0) = 1 | X(0) = 1] \cdot P[X(0) = 1].
\end{aligned}$$

Thus, instead of having to compute the joint probability of the 8-dimensional vector $[X(t_0), \ldots, X(t_7)]$, the result in (2.3) reduces this to computing the distribution of one 3-dimensional vector and two 2-dimensional vectors. More generally, if we define,

$$m_i = \min\{m | m > i \text{ and } x_m \in \mathcal{C}, \text{ or } m = M\},$$

then for any positive integer L,

$$\begin{aligned}
&P[X(t_0) = x_0, \ldots, X(t_L) = x_L] \\
&\quad = P[X(t_0) = x_0] \prod_{i=1, x_i \in \mathcal{C}}^{L} P\left[X(t_{i+1}) = x_{i+1}, \ldots, X\left(t_{m_i}\right) = x_{m_i} \middle| X(t_0) = x_0, \ldots, X(t_i) = x_i\right] \\
&\quad = P[X(t_0) = x_0] \prod_{i=1, x_i \in \mathcal{C}}^{L} P\left[X(t_{i+1} - t_i) = x_{i+1}, X\left(t_{m_i} - t_i\right) = x_{m_i} \middle| X(0) = x_i\right]. \qquad (2.4)
\end{aligned}$$

By taking $L = M$ and $t_m = \upsilon_m$, the result in (2.4) simplifies the calculation of the likelihood contribution in (2.2) by allowing $P(\mathbf{X} = \mathbf{x})$ to be expressed as a product of one-step probabilities of the form

$$P[X(\Delta_1) = x_1 | X(0) = x_0] \qquad (2.5)$$

and of multistep probabilities of the form

$$P[X(\Delta_1) = x_1, \ldots, X(\Delta_m) = x_m | X(0) = x_0], \qquad (2.6)$$

where $0 < \Delta_1 < \Delta_2 < \cdots$ correspond to differences between the original visit times $\upsilon_0, \ldots, \upsilon_M$.

Calculations of the conditional probabilities in (2.5) and (2.6) can be difficult because of the unknown transitions in the underlying sample paths. However, they also can be simplified by applying the result in (2.4) to the probabilities corresponding to the potential underlying sample paths that yield the observed states in (2.5) and (2.6).

3. A THREE-STATE SEMI-MARKOV MODEL

Consider a three-state semi-Markov model (Figure 1) where states 1 and 2 are transient while state 3 is absorbing and can be entered from either of the other states. As an example, suppose states 1, 2, and 3

Downloaded from https://academic.oup.com/biostatistics/article/8/2/252/230485 by guest on 21 August 2022

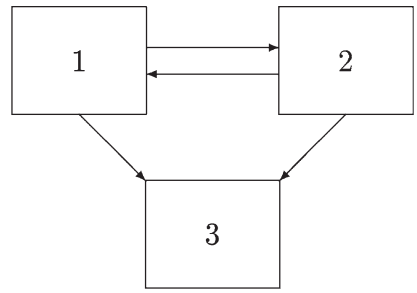

Fig. 1. A three-state semi-Markov process. States 1 and 2 are transient and recurrent and the absorbing state 3 can be entered from either state.

represent the infection-free, currently infected, and clinical disease statuses, respectively. Then a subject who is initially uninfected can become infected for a period of time, after which the infection resolves to the infection-free status (1–2–1), or the infection leads to clinical disease (1–2–3). Alternatively, the subject could develop clinical disease from the infection-free state (1–3). For someone beginning in state 1, every sample path will consist of k visits ($k \geqslant 0$) to state 2, followed by a visit to state 1 or 3 (if $k \geqslant 1$).

Suppose that individuals begin in state 1 and $\mathcal{C} = \{1\}$, so that $\lambda_{1j}(\cdot) = \lambda_{1j}$ for $j = 2, 3$. The remaining transition intensities, $\lambda_{2j}(\cdot)$ for $j = 1, 3$, are arbitrary. Let $f_{ij}(\cdot)$ denote the subdensity function corresponding to $\lambda_{ij}(\cdot)$; that is,

$$f_{ij}(t) = \lambda_{ij}(t) \exp\{-\Lambda_i(t)\},$$

for $t \geqslant 0$, where $\Lambda_i(t) = \sum_{j=1}^{K} \int_0^t \lambda_{ij}(u)\mathrm{d}u$, $i \neq j$. The subdistribution function corresponding to $f_{ij}(\cdot)$ is denoted $F_{ij}(\cdot)$, where $F_{ij}(t) = \int_0^t f_{ij}(u)\mathrm{d}u$.

From (2.3)–(2.6), calculation of the likelihood contribution of a subject requires consideration of at most the conditional probabilities,

$$\begin{aligned} &P[X(\Delta_1) = 1 | X(0) = 1], \\ &P[X(\Delta_1) = 2, X(\Delta_2) = 2, \ldots, X(\Delta_{m-1}) = 2, X(\Delta_m) = j | X(0) = 1], \end{aligned} \tag{3.1}$$

for $j = 1, 2, 3$, where $\Delta_i > 0$ and $0 < \Delta_1 < \Delta_2 < \cdots$ denote the distinct values of $v_j - v_i$ and $v_i < v_j$ are the visit times. When visit times are equally spaced, say every Δ time units, these conditional probabilities simplify to

$$\begin{aligned} &P[X(\Delta) = 1 | X(0) = 1], \\ &P[X(\Delta) = 2, X(2\Delta) = 2, \ldots, X((m-1)\Delta) = 2, X(m\Delta) = j | X(0) = 1], \end{aligned}$$

for $j = 1, 2, 3$ and $m = 1, 2, \ldots$.

Note that for each observed path, there may be infinitely many underlying sample paths. For example, for the observed sequence $\boxed{1} - \boxed{2} - \boxed{1}$, the underlying sample path of the process can be of the form

$$\boxed{1} - 2 - 1 - \boxed{2} - \boxed{1},$$

$$\boxed{1} - \boxed{2} - 1 - 2 - \boxed{1},$$

$$\boxed{1} - 2 - 1 - \boxed{2} - 1 - 2 - 1 - 2 - \boxed{1},$$

and so on, where the unboxed states denote the unobservable state changes of the process between visit times. We now consider the probability elements for the one- and multistep conditional probabilities in (3.1) in detail.

Downloaded from https://academic.oup.com/biostatistics/article/8/2/252/230485 by guest on 21 August 2022

3.1 *Conditional probabilities*

Let $P_{1j}(k,t)$ denote the conditional probability that the process is in state j at time t_0+t after k visits to state 2 in the interval (t_0, t_0+t), given that the process is in state 1 at time t_0, for $j=1,2,3$ and $k=0,1,\ldots$. Due to stationarity in (2.3), we may set $t_0=0$. Note that $P_{12}(k,t)=0$ for $k=0$, since there must be at least one visit to state 2. The case of $j=1$ for $k=0,1,2$ is depicted in Figure 2, where u represents the unknown time of transition from state 1 to 2 and z denotes that from state 2 to 1. In this notation,

$$P(X_1=1|X_0=1)=\sum_{k=0}^{\infty} P_{11}(k,t_1). \tag{3.2}$$

The sum will be finite if sojourn times from state 2 have a guarantee time; that is, support bounded away from zero. For example, if $\lambda_{2j}(t)=0$ for $0\leqslant t\leqslant G$, then the upper limit in the sum is the greatest integer less than or equal to t/G.

For the calculation of $P_{11}(k,t)$, we have $P_{11}(0,t)=\exp\{-\Lambda_1(t)\}$ and

$$P_{11}(1,t)=\int_u^t\left[\int_0^t f_{12}(u)\,f_{21}(z-u)\mathrm{d}u\right]\exp(-\Lambda_1(t-z))\mathrm{d}z.$$

For $k>1$, the probability of k transitions to state 2 is given by a convolution of probability functions $P_{11}^*(1,x)$ and $P_{11}(k-1,t-x)$, where

$$P_{11}^*(1,x)=\int_0^x f_{12}(u)\,f_{21}(x-u)\mathrm{d}u$$

differs from $P_{11}(k,x)$ in that its corresponding underlying process ends at the transition time (marked by state 1^* in Figure 2). Thus,

$$P_{11}(k,t)=\int_0^t P_{11}^*(1,x)P_{11}(k-1,t-x)\mathrm{d}x.$$

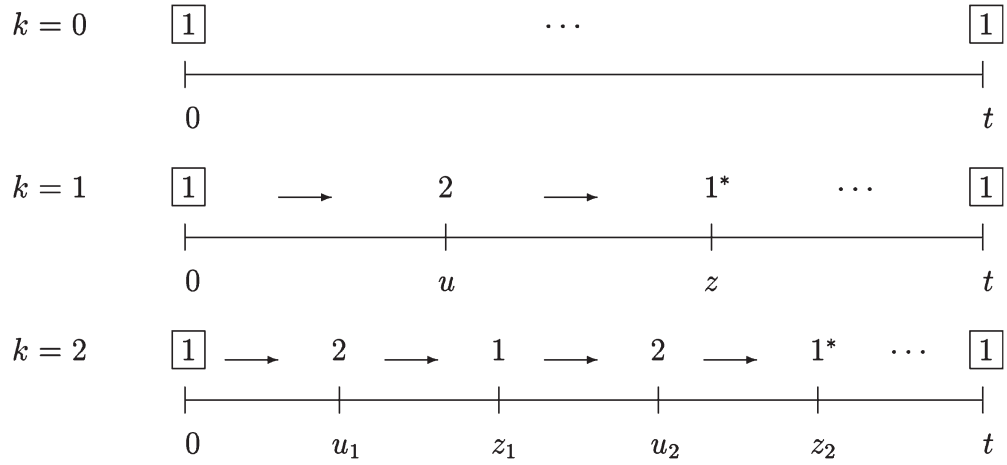

Fig. 2. Depiction of sample paths for $P_{11}(k,t)$, $k=1,2,3$. The boxed states represent the states occupied by the process at the visit time, the dots between the states indicate that the process has remained in the same state since the previous transition, and the arrows indicate the unobservable transitions. The asterisk next to 1 indicates the path segment that ends at transition to state 1.

Downloaded from https://academic.oup.com/biostatistics/article/8/2/252/230485 by guest on 21 August 2022

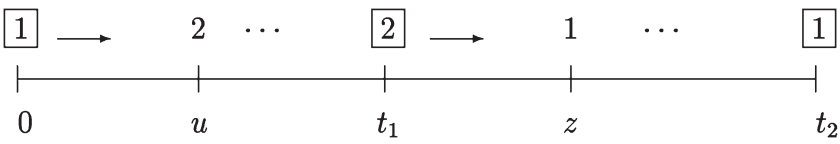

Fig. 3. Depiction of a sample path for $P_{121}(1, 0, t_1, t_2)$. The sample path observed to be in states 1, 2, and 1 at visit times 0, t_1, and t_2, respectively, with one transition to state 2 in the visit interval $(0, t_1]$ is shown.

We can extend the definition of $P^*_{11}(1, x)$ to $P^*_{11}(k_1, x)$, and generalize the convolution function to

$$P_{11}(k, t) = \int_0^t P^*_{11}(k_1, x) P_{11}(k_2, t - x)\mathrm{d}x,$$

where $k_1 + k_2 = k$, for $k_1, k_2 \geqslant 1$.

Next, let $P_{13}(k, t)$ denote the probabilities contributing to the calculation of $P[X(t) = 3 \mid X(0) = 1]$. Since state 3 can be reached from state 1 or state 2, $P_{13}(k, t)$ can be expressed as the sum of $P^{(1)}_{13}(k, t)$, which denotes that state 1 immediately precedes state 3, and $P^{(2)}_{13}(k, t)$, which denotes the other possibility that transition to state 3 was made from state 2. Then (see Appendix as supplementary material available at *Biostatistics* online)

$$P[X_1 = 3|X_0 = 1] = \sum_k \{P^{(1)}_{13}(k, t) + P^{(2)}_{13}(k, t)\}.$$

Having computed $P[X(t_0 + t) = 1|X(t_0) = 1]$ and $P[X(t_0 + t) = 3|X(t_0) = 1]$, $P[X(t_0 + t) = 2|X(t_0) = 1]$ is obtained as the complement of their sum.

The ideas in one-step conditional probability calculations can easily be extended to obtain the multistep conditional probabilities in (2.6). For example, consider $P[X(t_2) = 1, X(t_1) = 2 \mid X(0) = 1]$. Define $P_{121}(k_1, k_2, t_1, t_2)$ to be the conditional probability that $X(t_1) = 2$ and $X(t_2) = 1$, with k_1 visits to state 2 in $(0, t_1]$ and k_2 visits to state 2 in $(t_1, t_2]$, given that $X(0) = 1$. Note that $P_{121}(k_1, k_2, t_1, t_2)$ differs from $P_{11}(k_1 + k_2, t)$ in that state 2 is known to be occupied in the process at time t_1 in $P_{121}(k_1, k_2, t_1, t_2)$. To illustrate, a sample path corresponding to $P_{121}(1, 0, t_1, t_2)$ is depicted in Figure 3. Details are presented in the Appendix as supplementary material available at *Biostatistics* online.

3.2 *Estimation and model assessment*

The expressions developed above can be combined to obtain an expression for $P(\mathbf{X} = \mathbf{x})$ and then the likelihood contribution for an individual from (2.2). The overall likelihood is obtained as the product of the contributions of individual subjects and will be a function of $\{\lambda_{12}, \lambda_{13}, \lambda_{21}(\cdot), \lambda_{23}(\cdot)\}$. If parametric forms are assumed for $\lambda_{21}(\cdot)$ and $\lambda_{23}(\cdot)$, the likelihood function will depend on a finite-dimensional parameter vector and standard numerical methods can be used to obtain the maximum likelihood estimator. Under the standard regularity conditions, the maximum likelihood estimators will be consistent and asymptotically normal.

To examine the adequacy of model fit, the visit patterns can be partitioned into several categories, and the observed frequencies of the categories can be compared with the estimated model-based frequencies. An overall χ^2 statistic would then have an approximate chi-square distribution with the degrees of freedom equal to the number of independent cells minus the number of model parameters.

Downloaded from https://academic.oup.com/biostatistics/article/8/2/252/230485 by guest on 21 August 2022

4. ILLUSTRATIONS

4.1 *HPV application*

Data, model assumptions, and methods. We illustrate the proposed methods with the placebo data from a completed pilot clinical trial of a candidate vaccine for HPV type 16 (Koutsky *and others*, 2002). There were six scheduled visits at baseline (day 0), months 7, 12, 18, 24, and 30, at which times the presence or absence of HPV type 16 (denoted HPV16) and CIN 2/3 (denoted CIN) was assessed. We considered the 699 placebo subjects, initially HPV 16 negative and free of CIN 2/3, who provided the data on the first five (388) or full six visits (311). To facilitate the computations, we treated month 7 visit as if it occurred at month 6, so that all visits would be equally spaced. The setting can be described by the four-state process depicted in Figure 4, where state 3 (state 4) corresponds to diagnosis of CIN following a scheduled visit where HPV16 was not (was) detected. This model can be viewed as an extension of the three-state model in Figure 1 in which state 3 is divided into two states based on the presence of HPV16 when CIN was diagnosed. In practice, these two CIN states would be interpreted as having been caused by HPV16 (state 4) or another type of HPV (state 3). Overall, 10 subjects were diagnosed with CIN during the study period, five of whom were HPV16 positive at the time of diagnosis. The majority of subjects were observed to be in state 1 at all six visit times (260/311) or at all five times (344/388).

We considered $\mathcal{C} = \{1\}$ and modeled the transition intensities for state 2 as step functions, incorporating the biological minimum time required for clearance of infection or progression to disease; that is,

$$\lambda_{2j}(t) = \begin{cases} 0, & \text{if } t < G_{2j}, \\ \lambda_{2j}, & \text{if } t \geqslant G_{2j}, \end{cases}$$

for $j = 1, 4$. Hence, $\lambda_{2j}(t)$ is dependent on time due to the corresponding guarantee time, G_{2j}. We considered guarantee times of $G_{21} = 5$ and $G_{24} = 6$ (in months) for the HPV study. We assumed that the diagnostic tests for HPV and CIN have perfect sensitivity but may have imperfect specificity, and reestimated model parameters assuming several assumed specificities.

Results. Estimates of $\boldsymbol{\lambda}$ assuming various misclassification errors are presented in Table 1. The maximized likelihood decreases sharply as either specificity declines below one and indicates that the assumption of no misclassification errors provides the best model fit. The very small estimates of λ_{13} in the presence of imperfect CIN diagnostic test would suggest that the observed prevalence of women in state 3 is almost fully explainable by misclassification, implying that HPV infections of types other than 16 do not lead to CIN 2/3 during the 30-month period of follow-up. However, this would be inconsistent with the literature which suggests that various HPV types are responsible for CIN outcomes. (cf. Liaw *and others*, 1999). Therefore, the model assuming no misclassification errors seems more reasonable for the data.

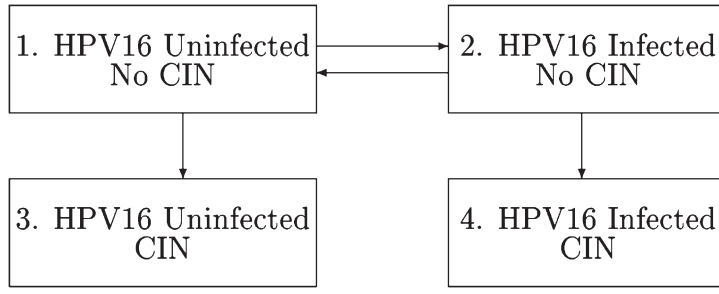

Fig. 4. A four-state semi-Markov process. A process describing the natural history of HPV16 is presented.

Downloaded from https://academic.oup.com/biostatistics/article/8/2/252/230485 by guest on 21 August 2022

Table 1. *Estimates of $\boldsymbol{\lambda}$ under various specificities. The specificity for HPV detection is denoted by P(PCR−| HPV−), where HPV detection is assessed by a polymerase chain reaction (PCR) assay, and the specificity for CIN detection is denoted by P(Dx−|CIN−) to indicate correct negative diagnosis (Dx−)*

P(Dx−|CIN−)	$\boldsymbol{\lambda}$	P(PCR−|HPV−)		
		1	0.9975	0.9950
1	λ_{12}	0.0051	0.0047	0.0045
	λ_{13}	0.0003	0.0005	0.0005
	λ_{21}	0.0882	0.0792	0.0724
	λ_{24}	0.0031	0.0031	0.0031
Log-likelihood		−517.470	−568.651	−569.353
0.9975	λ_{12}	0.0053	0.0049	0.0047
	λ_{13}	$<1.0 \times 10^{-10}$	$<1.0 \times 10^{-10}$	$<1.0 \times 10^{-10}$
	λ_{21}	0.0881	0.0794	0.0728
	λ_{24}	0.0023	0.0023	0.0023
Log-likelihood		−557.970	−560.712	−564.054
0.9950	λ_{12}	0.0053	0.0049	0.0047
	λ_{13}	$<1.0 \times 10^{-10}$	$<1.0 \times 10^{-10}$	2.8×10^{-10}
	λ_{21}	0.0880	0.0793	0.0727
	λ_{24}	0.0015	0.0015	0.0015
Log-likelihood		−560.934	−563.675	−567.017

We found the inverse sample information to be numerically unstable and relied on the bootstrap method using 100 bootstrapped samples. The model assuming known $G_{21} = 5$, $G_{24} = 6$, and no misclassification errors gives the following estimates of $\boldsymbol{\lambda}$ and standard errors, based on four unknown parameters in the model: $\widehat{\lambda_{12}} = 0.0051(0.00050)$, $\widehat{\lambda_{13}} = 0.0003(0.00013)$, $\widehat{\lambda_{21}} = 0.0882(0.014)$, and $\widehat{\lambda_{24}} = 0.0031(0.0037)$. Following the guarantee times, the estimated risk of clearance is about 28 times higher than the risk of progression to CIN after 6 months. The conditional probabilities of HPV16 infection (entering state 2) and non-HPV16-related CIN development (entering state 3) are 0.944 and 0.056, given that a woman leaves the uninfected state (state 1). Once a woman is infected with HPV16, the conditional probabilities of clearing the HPV16 infection, $P(\sigma_{n+1} = 1|\sigma_n = 2)$, and progressing to CIN, $P(\sigma_{n+1} = 4|\sigma_n = 2)$, are 0.969 and 0.031, respectively. The mean times to clearance and progression to HPV, conditional on the next state, are about 16 and 17 months, respectively, consistent with commonly used definitions of "persistent infection" (Ho *and others*, 1998; Koutsky *and others*, 2002; Moscicki *and others*, 1998). The estimated cumulative probability that a woman develops CIN based on the model estimates (Figure 5) shows a steady rise over time, with about one-third of the outcomes attributed to HPV16, also consistent with the epidemiological literature (Liaw *and others*, 1999; Herrero *and others*, 2000). HPV16 prevalence (Figure 5) stabilizes to 7–8% after about 3 years.

Model fit (Table 2) was assessed by comparing the observed frequencies with the expected frequencies of the selected visit patterns. The chosen model (Model 1) appears to fit the data adequately ($p = 0.65$, 8d.f.), while the goodness-of-fit results for the same model but allowing a specificity of 0.99 for the HPV16 assay (Model 2) indicate poorer fit ($p = 0.04$).

Fitting the standard Markov model (Kalbfleisch and Lawless, 1985) yielded constant transition intensities (denoted by γ_{ij}) of $\widehat{\gamma_{12}} = 0.0062$ (0.00067), $\widehat{\gamma_{13}} = 0.0003$ (0.00013), $\widehat{\gamma_{21}} = 0.0551$ (0.0089), and $\widehat{\gamma_{24}} = 0.0050$ (0.0022). The estimated transition intensities from state 1 were very similar to those from the semi-Markov methods. The resulting cumulative incidence and prevalence curves were also similar to

Downloaded from https://academic.oup.com/biostatistics/article/8/2/252/230485 by guest on 21 August 2022

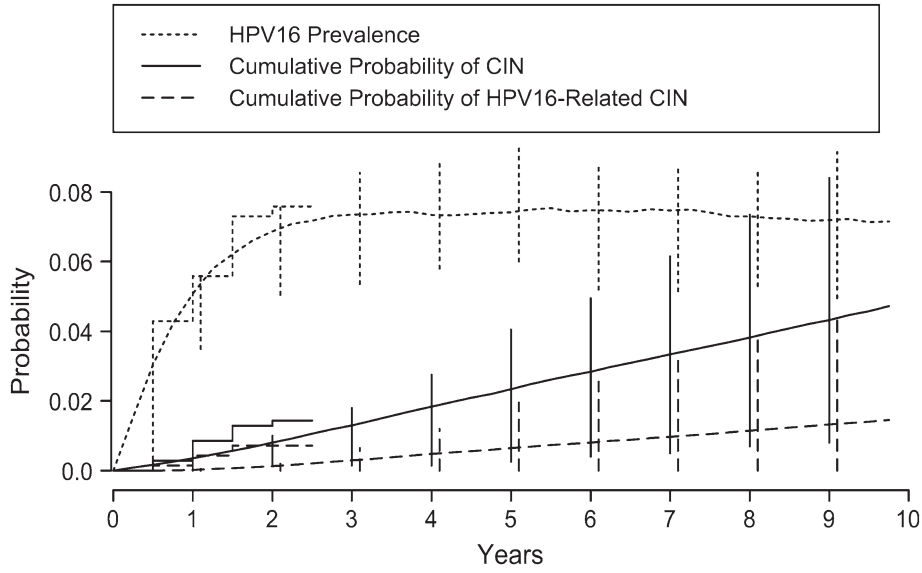

Fig. 5. Model-based estimation of cumulative incidence of CIN 2/3 and prevalence of HPV16. The dotted, solid, and dashed lines represent the prevalence of HPV16, the cumulative probability of overall CIN 2/3, and the cumulative probability of CIN 2/3 caused by HPV16, respectively. The step functions portray the empirical estimates from the 30-month data, and the smooth curves represent the model-based estimates over a period of 10 years. The vertical lines are the pointwise 95% confidence intervals for the predicted probabilities.

Table 2. *Model fit. The dots refer to any remaining visit patterns*

Visit pattern	Observed frequencies	Model 1	Model 2
$[1]-[2]-[1]\cdots$	7	5.6982	3.1728
$[1]-[2]-[2]\cdots$	22	14.9149	13.7062
$[1]-[2]-[3]\cdots$	0	0.0040	0.0039
$[1]-[2]-[4]\cdots$	0	0.1462	0.1894
$[1]-[1]-[2]-[1]\cdots$	3	5.5192	3.0857
$[1]-[1]-[2]-[2]\cdots$	14	14.4465	13.3298
$[1]-[1]-[2]-[3]\cdots$	0	0.0039	0.0038
$[1]-[1]-[2]-[4]\cdots$	0	0.1462	0.1842
$[1]-[1]-[1]-[2]-[1]\cdots$	7	5.3459	3.0009
$[1]-[1]-[1]-[2]-[2]\cdots$	15	13.9929	12.9637
$[1]-[1]-[1]-[2]-[3]\cdots$	0	0.0039	0.0037
$[1]-[1]-[1]-[2]-[4]\cdots$	0	0.1372	0.1791
Other	631	638.6411	649.1717
$(\text{Obs}-\text{Exp})^2/\text{Exp}$		5.9393	16.3933

Downloaded from https://academic.oup.com/biostatistics/article/8/2/252/230485 by guest on 21 August 2022

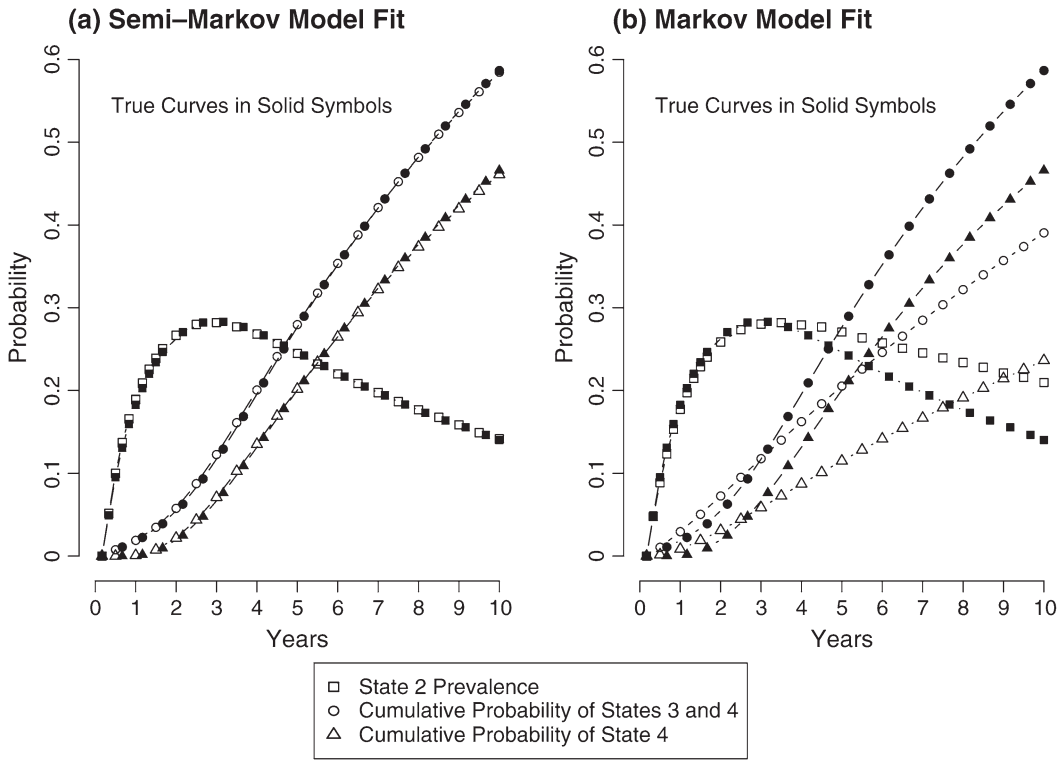

Fig. 6. (a) Semi-Markov and (b) Markov model fits for the simulated data set. The solid symbols represent the true curves and the unfilled symbols the estimates.

those from the semi-Markov model over the period of observation. The corresponding goodness-of-fit χ^2 statistic was 6.93 ($p = 0.54$) for the Markov model, somewhat worse than that of Model 1.

4.2 *A simulated data example*

To further illustrate the value of the methods, data were simulated from a four-state semi-Markov process of Figure 4 with $\mathcal{C} = \{1\}$ and $\lambda_{2j}(t) = a_{2j}b_{2j}(t - G_{2j})^{b_{2j}-1}$, $j = 1, 4$ (Weibull hazard function). For the simulation, $\lambda_{12} = 0.025$, $\lambda_{13} = 0.002$, $a_{21} = 0.2$, $a_{24} = 0.001$, $b_{21} = 0.5$, $b_{24} = 2$, $G_{21} = 4$, $G_{24} = 5$, visits are every 6 months for 30 months and $N = 5000$, a sample size typical of a moderately sized Phase III vaccine trial. The simulated data produced 231 and 233 observations in states 3 and 4, respectively, by the end of the observation period, with 2252 subjects observed to be in state 1 at each visit, and yielded $\widehat{\lambda_{12}} = 0.0263$, $\widehat{\lambda_{13}} = 0.00206$, $\widehat{a_{21}} = 0.209$, $\widehat{a_{24}} = 0.00158$, $\widehat{b_{21}} = 0.510$, and $\widehat{b_{24}} = 1.873$, assuming that G_{21} and G_{24} are known. Figure 6(a) shows that the estimated cumulative CIN incidence and HPV prevalence are very close to the true probabilities. However, when a Markov model is incorrectly assumed, the biases in such estimates increase with time (Figure 6(b)).

5. DISCUSSION

Panel data pose challenges for semi-Markov processes because missing information about an individual's process between the observations can contribute to the probability of being in the current state. Hence, despite the wider applicability of semi-Markov processes compared to the Markov processes, research in

Downloaded from https://academic.oup.com/biostatistics/article/8/2/252/230485 by guest on 21 August 2022

this area has been slim. We showed that when the transition intensities from at least one of the states of the process are time homogeneous, the expression for the joint probability in the likelihood function is tractable. In our data applications, we considered a process that begins in a state in $\mathcal{C}$, so that it is not necessary to take account of the duration of time that individuals have been in this state prior to the start of the study. If $X(0) \notin \mathcal{C}$, the proposed methods would still apply if either the subjects had just entered this state or the durations of time in this state prior to entrance into the study were known. Otherwise, the methods would need to be modified to account for the duration of time in the initial state prior to the start of the study. Methods similar to those in the work by Satten and Sternberg (1999) might be useful for this purpose.

In the HPV data example, the data came from a proof-of-principle study, thus smaller than a Phase III vaccine trial, and the number of CIN 2/3 events was small (10 events). Several Phase III HPV vaccine trials are currently underway, and are approximately five times the size of the pilot trial, and will be more suitable for the methods developed here, as the simulated data example illustrates. Without data on sexual activities and HPV types other than type 16, we assumed $\mathcal{C} = \{1\}$, and exponential distributions with guarantee times were chosen for state 2 for model simplicity. Whereas the guarantee times limited the number of potential paths in our example, one can also limit the number of transitions that can occur in a given time interval when enumerating the potential paths. The clinical estimates for guarantee times were not clear, but we found that varying these times (from 4 to 8 months) did not change the likelihood parameter estimates substantially in our data.

The methods in the paper can be extended to allow fixed covariates in the model, for instance, to compare two treatments in a clinical trial. A one-sample model can be fit separately for each distinct covariate vector, and the overall likelihood is obtained by multiplying the likelihoods from the homogeneous groups.

Acknowledgments

We are grateful to Joseph Heyse and Lisa Chiacchierini for their comments and to Merck Research Laboratories for providing the data used in Section 4. This research was supported by the Howard Hughes Medical Institute as part of the first author's doctoral thesis and by Grants U01 AI38855 (Kang) and AI24643 (Lagakos) from the National Institute of Allergy and Infectious Diseases. *Conflicts of Interest:* None declared.

References

Andersen, P. (1988). Multistate models in survival analysis: a study of nephropathy and mortality in diabetes. *Statistics in Medicine* **7**, 661–70.

Andersen, P. and Borgan, O. (1985). Counting process models for life history data: a review. *Scandinavian Journal of Statistics* **12**, 155–8.

Andersen, P. and Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research* **11**, 91–115.

Commenges, D. (1999). Multi-state models in epidemiology. *Lifetime Data Analysis* **5**, 315–27.

Commenges, D. (2002). Inference for multi-state models from interval-censored data. *Statistical Methods in Medical Research* **11**, 167–82.

Cox, D. and Miller, H. (1977). *The Theory of Stochastic Processes*. London: Chapman and Hall.

Gentleman, R., Lawless, J., Lindsey, J. and Yan, P. (1994). Multi-state Markov models for analysing incomplete disease history data with illustrations for HIV disease. *Statistics in Medicine* **13**, 805–21.

Downloaded from https://academic.oup.com/biostatistics/article/8/2/252/230485 by guest on 21 August 2022

HERRERO, R., HILDESHEIM, A., BRATTI, C., SHERMAN, M. E., HUTCHINSON, M., MORALES, J., BALMACEDA, I., GREENBERG, M., ALFARO, M., BURK, R. *and others* (2000). Population-based study of human papillomavirus infection and cervical neoplasia in rural Costa Rica. *Journal of the National Cancer Institute* **92**, 464–74.

HO, G. Y. F., BIERMAN, R., BEARDSLEY, L., CHANG, C. J. AND BURK, R. D. (1998). Natural history of cervicovaginal papillomavirus infection in young women. *The New England Journal of Medicine* **338**, 423–8.

HOUGAARD, P. (1999). Multi-state models: a review. *Lifetime Data Analysis* **5**, 239–64.

KALBFLEISCH, J. AND LAWLESS, J. (1985). The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association* **80**, 863–71.

KAY, R. (1986). A Markov model for analysing cancer markers and disease states in survival studies. *Biometrics* **42**, 855–65.

KOUTSKY, L., AULT, K., WHEELER, C., BROWN, D., BARR, E., ALVAREZ, F., CHIACCHIERINI, L. AND JANSEN, K. (2002). A controlled trial of a human papillomavirus type 16 vaccine. *The New England Journal of Medicine* **347**, 1645–51.

LAGAKOS, S., SOMMER, C. AND ZELEN, M. (1978). Semi-Markov models for partially censored data. *Biometrika* **65**, 311–8.

LIAW, K. L., GLASS, A. G., MANOS, M. M., GREER, C. E., SCOTT, D. R., SHERMAN, M., BURK, R. D., KURMAN, R. J., WACHOLDER, S., RUSH, B. B. *and others* (1999). Detection of human papillomavirus DNA in cytologically normal women and subsequent cervical squamous intraepithelial lesions. *Journal of the National Cancer Institute* **91**, 954–60.

MOSCICKI, A. B., SHIBOSKI, S., BROERING, J., POWELL, K., CLAYTON, L., JAY, N., DARRAGH, T. M., BRESCIA, R., KANOWITZ, S., MILLER, S. B. *and others* (1998). The natural history of human papillomavirus infection as measured by repeated DNA testing in adolescent and young women. *The Journal of Pediatrics* **132**, 277–84.

SATTEN, G. AND STERNBERG, M. (1999). Fitting semi-Markov models to interval-censored data with unknown initiation times. *Biometrics* **55**, 507–13.

STERNBERG, M. AND SATTEN, G. (1999). Discrete-time nonparametric estimation for semi-Markov models of chain-of-events data subject to interval censoring and truncation. *Biometrics* **55**, 514–22.

STOLER, M. (2000). Human papillomaviruses and cervical neoplasia: a model for carcinogenesis. *International Journal of Gynecological Pathology* **19**, 16–28.

[*Received October 9, 2005; revised May 18, 2006; accepted for publication May 31, 2006*]

Downloaded from https://academic.oup.com/biostatistics/article/8/2/252/230485 by guest on 21 August 2022