12-23-2013

Multi-state Models for Natural History of Disease

Amy Laird
University of Washington - Seattle Campus, amylaird@u.washington.edu

Rebecca A. Hubbard
Group Health Research Institute, rhubb@uw.edu

Lurdes Y. T. Inoue
University of Washington - Seattle Campus, linoue@u.washington.edu

Chapter 1

Multi-state Models for Natural History of Disease

A. E. Laird, R. A. Hubbard, L. Y. T. Inoue

1.1 Introduction

A variety of disease processes can be described through transitions between discrete states, such as stable or accelerated disease state in leukemia; development of AIDS defining illnesses in HIV; or diminished lung function in asthma patients. In progressive diseases, subjects traverse disease states in only one direction while in non-progressive diseases it is possible for subjects to experience repeated occurrences of some or all states.

Studies of chronic disease often utilize longitudinal observations of a cohort of subjects to characterize natural history of disease. Subjects may be observed continuously in which case the exact time of transition between states is known. However, in studies of human health, panel observation, in which subjects are observed and disease states assessed only at discrete time-points, is more common. In data arising from panel observation, the exact time of state transitions is unknown.

Multi-state models provide flexible tools for describing characteristics of disease processes and can be estimated under either continuous or panel observation. These models are particularly useful in the context of panel observation because they allow us to estimate the rate of disease progres-

sion even if exact transition times are not observed. If the observations are equally spaced in time then discrete time models, such as the Markov chain, can be used. If the length of time between observations is not constant, continuous time models can be employed. The most commonly used multi-state model is the Markov process model, which assumes that the probability of transition between disease states depends only on the elapsed time between observations. To accommodate more complex multi-state disease processes which may feature time-varying transition probabilities, nonhomogeneous Markov models or semi-Markov models can be used.

In this chapter, we review basic properties of some commonly used multi-state models with a focus on models that are appropriate for panel observation in biomedical/biological applications. We introduce Bayesian estimation methods for these models and demonstrate their use in two longitudinal studies of disease progression.

1.2 Multi-state models: background and notation

In this section we introduce notation and review a few multi-state models commonly used for describing natural history of disease. We will generally use $\{Z(t), t \in [0, \infty)\}$ to denote a stochastic process taking a finite set of states $\mathcal{S} = \{1, 2, \ldots, m\}$. In our applications $Z(t)$ represents the disease state of a patient at time t.

1.2.1 Markov processes.

In a Markov process, at each time point, the state of the process at a future time $t + s$ depends on the history of the process only through the state at present time, t. Formally, the process is *Markov* if for all $s, t \geq 0$ and for every $i, j \in \mathcal{S}$

$$P(Z(t+s) = j | Z(t) = i, Z(u) = z(u), 0 \leq u < s) = P(Z(t+s) = j | Z(t) = i) \doteq p_{ij}(t, t+s).$$

Denote the initial distribution of the process by $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)$ where $\phi_i \doteq P(Z(0) = i), \forall i \in \mathcal{S}$ and such that $\phi_i \geq 0$ for each i and $\sum_{i \in \mathcal{S}} \phi_i = 1$. Conditional on $\boldsymbol{\phi}$, the process can be characterized by the matrix of transition probabilities $\mathbf{P}(t, t+s) = [p_{ij}(t, t+s)]$ for $s, t \geq 0$. Alternatively, the process can be characterized by the matrix $\mathbf{Q}(t) = [q_{ij}(t)]$ of transition intensities, defined for $t \geq 0$ as

$$q_{ij}(t) \doteq \lim_{s \to 0} \frac{p_{ij}(t, t+s)}{s} \quad \text{for } i, j \in \mathcal{S}, j \neq i,$$

and

$$q_{ii}(t) \doteq -\sum_{j \neq i} q_{ij}(t) \quad \text{for each } i \in \mathcal{S}.$$

The $q_{ij}(\cdot)$ are also known as *cause-specific hazard functions* (Prentice et al., 1978). It follows from this definition that $\mathbf{P}(t,t) = \mathbf{I}$.

In the more general case where transition probabilities depend on both the elapsed time, s, and the chronological time, t, the process is a *nonhomogeneous Markov process*. If the transition probabilities depend only on the elapsed time s and not on the chronological time t, then the Markov process is *homogeneous* and $p_{ij}(t, t+s) \equiv p_{ij}(s)$ and $q_{ij}(t) \equiv q_{ij}$. As a consequence of the Markov property and homogeneity, the transition probability matrix $\mathbf{P}(s)$ satisfies the *Chapman-Kolmogorov equation*:

$$p_{ij}(s) = \sum_{k \in \mathcal{S}} p_{ik}(u) p_{kj}(s-u), \;\; 0 < u < s,$$

or equivalently in matrix notation:

$$\mathbf{P}(s) = \mathbf{P}(u)\mathbf{P}(s-u), \;\; 0 < u < s.$$

From the Chapman-Kolmogorov equations we can derive the forward and backward equations:

$$\frac{d}{ds}\mathbf{P}(s) = \mathbf{Q}\mathbf{P}(s) = \mathbf{P}(s)\mathbf{Q},$$

which can be solved to yield

$$\mathbf{P}(s) = \exp(\mathbf{Q}s) \doteq \sum_{n=0}^{\infty} \frac{\mathbf{Q}^n s^n}{n!},$$

where $\mathbf{Q}^0 \equiv \mathbf{I}$. This last equation makes clear that for a homogeneous Markov process, the matrix of transition intensities and the matrix of transition probabilities give equivalent characterizations of the process.

If the domain of the Markov process is the set of nonnegative integers $\mathbb{Z}^* = \{0, 1, 2, \dots\}$ rather than a real interval, then the process $\{Z_t, t \in \mathbb{Z}^*\}$ is called a *Markov chain* and the Markov assumption reduces to

$$P(Z_{t+1} = j | Z_t = i, Z_{t-1} = z_{t-1}, \dots, Z_0 = z_0) = P(Z_{t+1} = j | Z_t = i) \doteq p_{ij}(t), \;\; t \in \mathbb{Z}^*.$$

A Markov chain is uniquely characterized by its initial distribution $\boldsymbol{\phi}$ and transition probability matrix $\mathbf{P}(\cdot)$. Similar to a Markov process, a Markov

chain is *homogeneous* if the transition probabilities do not depend on chronological time so that $p_{ij}(t) \equiv p_{ij}$. For a homogeneous Markov chain p_{ij} gives the probability of making a transition from state i to state j in one step, but we can also consider the probability of being in state j several steps after being in state i. The matrix of n-step transition probabilities, denoted $\mathbf{P}^{(n)}$, is given by $\mathbf{P}^{(n)} = \mathbf{P}^n$, the matrix of one-step transition probabilities raised to the n^{th} power. This follows from the discrete-time version of the Chapman-Kolmogorov equations.

Examination of the transition probability matrix $\mathbf{P}(\cdot)$ yield insights into the behavior of the Markov chain. Considering a homogeneous Markov chain, if $p_{ii} = 1$, then state i is called an *absorbing state* (Chiang, 1980, p. 114; Limnios and Oprişan, 2001, p. 86). A state j is said to be *accessible* from state i if $p_{ij}^n > 0$ for some $n \geq 0$ (Ross, 1996, p. 168). States of a Markov process may be classified in an analogous way by examining the transition intensity matrix, $\mathbf{Q}(\cdot)$.

1.2.2 Markov renewal processes and semi-Markov processes.

To discuss a process for which the Markov assumption is relaxed, we turn to a framework that separates the evolution of the process into its sequence of states and of sojourn times, where a *sojourn time* is the length of time between two consecutive transitions.

Consider a discrete two-dimensional stochastic process called a *J-X process*, $(J - X) = \{(J_n, X_n),\ n \geq 0\}$, where the J-process represents the states visited and the X-process represents the sojourn times in each of those states. Hence $X_n \geq 0$ and $J_n \in \mathcal{S}$, $\mathcal{S} = \{1, 2, \ldots, m\}$, for each $n \geq 0$, and by convention $X_0 = 0$ almost surely. The process begins in state J_0, where it remains for time X_1 before making a transition to state J_1 and, in general, remains in state J_n for time X_{n+1} before making a transition to a state J_{n+1}. The time at which the n^{th} transition occurs is $T_n \doteq \sum_{r=1}^{n} X_r,\ n \geq 1$ and $T_0 = 0$ (see Figure 1.1).

Define the initial distribution of the process as $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)$ where $\phi_i = P(J_0 = i)$ with $\phi_i \geq 0$ for all i and $\sum_{i \in \mathcal{S}} \phi_i = 1$. Moreover, define $N(t) = \sup\{n \geq 0 :\ T_n \leq t\}$, the number of transitions made during $[0, t]$. The semi-Markov assumption is that for all $s \geq 0$,

$$\begin{aligned} P(J_n = j, X_n \leq s | (J_k, X_k),\ k = 0, 1, \ldots, n-1) &= P(J_n = j, X_n \leq s | J_{n-1}, T_{n-1}, n-1) \\ &\doteq {}^{(n-1)}K_{J_{n-1}j}(T_{n-1}, s) \end{aligned}$$

for $n \geq 1$ and $j \in \mathcal{S}$, where for each i and j, ${}^{(n-1)}K_{ij}(\cdot, \cdot)$ is a real-valued

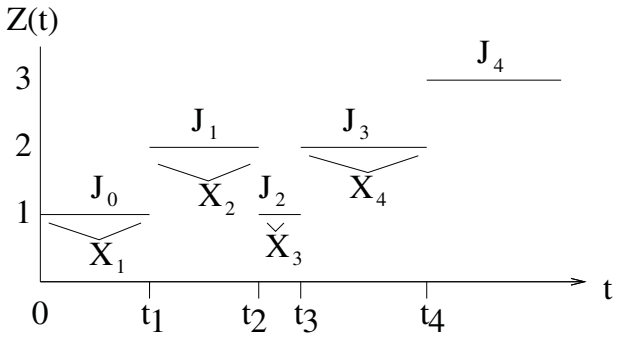

Figure 1.1: Example of a *J-X* process, showing relationships between $\{(J_n, X_n),\ n \geq 0\}$ and associated process $Z(\cdot)$. Under the semi-Markov assumption, the *J-X* process is a Markov renewal process, and $Z(\cdot)$ is the associated semi-Markov process.

function satisfying

$$^{(n-1)}K_{ij}(t, t+s) = 0 \quad \text{for } s \leq 0 \text{ or } t \leq 0$$

and

$$\lim_{s\to\infty} \sum_{j\in\mathcal{S}} {}^{(n-1)}K_{ij}(t, t+s) = 1 \quad \text{for each } i \in \mathcal{S},\ t \geq 0.$$

If this assumption holds, then $\{(J_n, X_n)\}$ is a *Markov renewal process* and the associated process $Z(t) \doteq J_{N(t)}$ is a *completely nonhomogeneous semi-Markov process* (refer to Figure 1.1). We can interpret this assumption as the statement that the future of the process depends on the entire history only through the current state, J_{n-1}, the elapsed chronological time, T_{n-1}, and the number of transitions between states, $n-1$, that the process has made.

This very general formulation of a semi-Markov process includes several important special cases. If the kernel $[^{(n-1)}K_{ij}(\cdot,\cdot)]$ does not depend on the number of transitions, then the process is *nonhomogeneous semi-Markov* (Iosifescu-Manu, 1972), and if additionally the kernel does not depend on the chronological time t, then the process is *homogeneous semi-Markov* (Lévy, 1954a,b; Smith, 1955). We discuss the latter case in more detail. Specifically, if for all $s \geq 0$,

$$\begin{aligned} P(J_n = j, X_n \leq s | (J_k, X_k),\ k = 0, 1, \ldots, n-1) &= P(J_n = j, X_n \leq s | J_{n-1}) \\ &\doteq K_{J_{n-1}j}(s), \end{aligned}$$

where each $K_{ij}(\cdot)$ is a real-valued function satisfying

$$K_{ij}(s) = 0 \quad \text{for} \;\; s \leq 0$$

and

$$\lim_{s\to\infty} \sum_{j\in\mathcal{S}} K_{ij}(s) = 1 \quad \text{for each} \;\; i \in \mathcal{S},$$

then the associated process $Z(\cdot)$ is a homogeneous semi-Markov process. The assumption for such a process is that the future evolution depends on the history only through the current state of the process and the elapsed time in this state. This assumption is much weaker than the homogeneous Markov assumption. From this point forward we consider only homogeneous semi-Markov processes, referred to in many sources as simply semi-Markov processes, and we assume that each element of the *semi-Markov kernel* $K_{ij}(\cdot)$ is absolutely continuous. A semi-Markov process can be uniquely characterized by its initial distribution $\boldsymbol{\phi}$ and the *kernel* $\mathbf{K}$.

Let us further examine the marginal process $\{J_n, n \geq 0\}$ and the process $\{X_n, n \geq 0\}$ conditional on $\{J_n, n \geq 0\}$, which we call the J- and X-processes, respectively. Using the semi-Markov assumption and the Lebesgue Monotone Convergence Theorem (Pyke, 1961a), we can show that the J-process is a homogeneous Markov chain, called the *embedded Markov chain* of the semi-Markov process that is governed by the transition probability matrix defined by $p_{ij} \doteq \lim_{s\to\infty} K_{ij}(s)$ for all $i, j \in \mathcal{S}$.

To discuss the X-process, define, for $s \geq 0$, the following functions:

$$F_{ij}(s) \doteq \begin{cases} \frac{K_{ij}(s)}{p_{ij}}, & p_{ij} > 0; \\ 1_{(s\geq 1)}, & p_{ij} = 0 \end{cases}$$

for each i and j and

$$H_i(s) \doteq \sum_{j\in\mathcal{S}} K_{ij}(s)$$

for each i. We can show that, for $s \geq 0$,

$$\begin{aligned} F_{ij}(s) &= P(X_n \leq s | J_{n-1} = i, J_n = j) \\ H_i(s) &= P(X_n \leq s | J_{n-1} = i). \end{aligned}$$

These are known, respectively, as the conditional and unconditional distributions of the sojourn time in state i. We note that the above definition of

$F_{ij}(\cdot)$ in the case that $p_{ij} = 0$ is arbitrary. Let $f_{ij}(\cdot)$ be the density corresponding to $F_{ij}(\cdot)$ which exists given our assumption that the semi-Markov kernel is absolutely continuous.

We can express each element of the kernel in a natural way as the product of the respective transition probability and the conditional sojourn time distribution:

$$K_{ij}(s) = F_{ij}(s) \cdot p_{ij} \quad \text{for} \quad s \geq 0.$$

We noted previously that the semi-Markov process can be uniquely characterized by $(\boldsymbol{\phi}, \mathbf{K})$, and the above argument makes clear (Janssen and Manca, 2006) that it can also be characterized by $(\boldsymbol{\phi}, \mathbf{P}, \mathbf{F})$. From standard survival analysis we know that under some regularity conditions, the time to failure can be characterized by the cumulative distribution function $F(\cdot)$ or the hazard function $h(\cdot)$. By analogy, if X_n is the sojourn time in state $J_{n-1} = i$ before going to state $J_n = j$, then we can characterize the distribution of X_n by F_{ij} or equivalently by the conditional hazard function, $h_{ij}(\cdot)$, for each $i \neq j \in \mathcal{S}$, defined for $s \geq 0$ as:

$$h_{ij}(s) \doteq \lim_{\Delta s \downarrow 0} \frac{1}{\Delta s} P(s \leq X_n < s + \Delta s | J_{n-1} = i, J_n = j,\ X_n \geq s).$$

Hence we can alternatively characterize the semi-Markov process by $(\boldsymbol{\phi}, \mathbf{P}, \mathbf{h})$.

1.3 Estimation and inference in multi-state models: a review of approaches

We begin this section with a summary of the existing approaches to estimation of continuously observed multi-state processes. Next, we examine available approaches to estimation of processes under panel observation. If the process is observed only at discrete time points, a Markov model has just enough structure so that the transition intensities can still be estimated (Kalbfleisch and Lawless, 1985). However, in a semi-Markov model the transition intensities depend on the elapsed time in the current state, which is unknown under panel observation. Estimation is therefore less tractable and methods for a general process under panel observation do not exist. Thus, we examine methods that have been developed for processes under specific assumptions for the sequence of allowed transitions, defined as state models, and under various other assumptions. As we shall see, most of the existing literature is based on maximum likelihood estimation, however. Thus, we end this section with a discussion of how Bayesian approaches can be implemented for estimating disease natural history.

1.3.1 Methods for continuously observed processes.

When the exact time of transition between states is known, maximum likelihood estimators (MLEs) for transition intensities in the Markov process are given by

$$\hat{q}_{ij} = \frac{N_{ij}}{T_i},$$

where N_{ij} is the number of transitions observed from state i to state j and T_i is the total time spent in state i (Albert, 1962). Moreover, large sample properties for these estimators have been developed by Billingsley (1961) and Albert (1962). Likewise, when a semi-Markov model is used for a continuously observed process, the parameters corresponding to the embedded Markov chain are estimated via the sample proportions

$$\hat{p}_{ij} = \frac{n_{ij}}{n_i},$$

where n_{ij} is the number of observed transitions from state i to j and n_i is the total number of transitions from i to any state (Anderson and Goodman, 1957). In the semi-Markov model, it remains to estimate the distributions of the sojourn times conditional on the sequence of visited states and a variety of approaches may be taken: fully parametric (Weiss and Zelen, 1965), piecewise exponential (Colvert and Boardman, 1976; Ouhbi and Limnios, 1999), or nonparametric (Voelkel and Crowley, 1984; Kaplan and Meier, 1958). Moreover, tests have been developed to examine the Markov or semi-Markov assumption. Chang, Chuang, and Hsiung (2001) consider an illness-death model (see Figure 1.2b) and propose goodness-of-fit statistics for testing the hypotheses that the underlying process is either (1) homogeneous semi-Markov or (2) nonhomogeneous Markov, and derive asymptotic distributions of the statistics.

Frequently there is reason to account for differences among subgroups of the patient population. To carry this out by covariate adjustment, parametric and semiparametric regression approaches have been taken (Therneau and Grambsch, 2000). Specifically, covariates can be incorporated into Markov process models by allowing for covariate dependence of transition intensities. For instance, in a homogeneous Markov model we can specify a proportional hazards type model,

$$q_{jk} = q_{jk0} g(\mathbf{W}'\boldsymbol{\beta}_{jk}),$$

where q_{jk0} is the baseline transition intensity, $\mathbf{W}$ is a vector of covariates, and $g(\cdot)$ is a positive-valued function. Typically, $g(\cdot)$ is taken to be the

exponential function to ensure positivity of the transition intensities. This model has been previously implemented in a number of applications. Kay (1986) took this approach in modeling hepatic cancer, while Jackson et al. (2003) applied it to estimate the effect of age on rates of progression of abdominal aortic aneurysm.

In semi-Markov models, covariates may be included in a regression model via the embedded Markov chain, conditional sojourn distributions, or both. Regression modeling of the embedded Markov chain is often based on the multinomial logistic regression model. Conditional sojourn times are often modeled via the proportional hazards model. Lawless and Fong (1999) give an overview of methods for modeling sojourn times that account for the presence of covariates and possible dependencies among sojourn times within a subject. They also review methods for dealing with various observation schemes, including left truncation of observations as well as selection mechanisms in observational studies. They discuss the use of random effects to deal with unexplained inter-subject or temporal variability. Since random effects are a modeling device and can introduce computational difficulties, the authors suggest that the use of random effects be avoided when the process is incompletely observed.

There are several classes of methods that apply to certain state models. Some methods are built on the assumption that the embedded Markov chain of the process is *ergodic*, which implies in particular that an absorbing state such as death cannot exist (Ross, 1996). By contrast, other methods assume that the underlying process is progressive. Voelkel and Crowley (1984) approach semi-Markov processes in a counting processes framework and show that, under some assumptions, a progressive semi-Markov process can be transformed via a random function of the chronological time into the *multiplicative intensity model* introduced by Aalen (1978). Voelkel and Crowley then consider a particular progressive state model and establish asymptotic properties of the estimator of the probability of being in one of the states.

Although a number of methods have addressed censoring, most have focused on right-censoring in the final state or left-censoring in the initial state (e.g. Lagakos, Sommer, and Zelen, 1978). Extending these methods to a panel observation scheme has been elusive.

Finally, the Bayesian approach has been utilized to estimate Markov models (Converse et al., 2012; Price et al., 2011; Kneib and Hennerfeind, 2008; Pan et al., 2007; Sweeting et al., 2005) in biomedical and biological applications. We have not identified Bayesian applications using standard semi-Markov models, however, for modeling disease processes.

1.3.2 Panel data: Markov models.

In a seminal paper, Kalbfleisch and Lawless (1985) proposed a method to estimate the instantaneous transition probabilities of a general multi-state process under panel observation assuming the process is Markov. Using the fact that the transition probability and transition intensity matrices are related via $\mathbf{P}(s) = \exp(\mathbf{Q}s) \doteq \sum_{r=0}^{\infty} \frac{\mathbf{Q}^r s^r}{r!}$ for $s \geq 0$ for a homogeneous Markov process, the authors proposed an efficient scoring procedure to estimate $\mathbf{Q}$ via maximum likelihood. Specifically, if subjects are observed at times $t_0, t_1, \ldots, t_m$, and if $\mathbf{Q}$ depends on $\boldsymbol{\theta}$ then the likelihood of $\boldsymbol{\theta}$ is given by

$$L(\boldsymbol{\theta}) = \prod_{l=1}^{m} \prod_{i,j \in \mathcal{S}} p_{ij}(t_l - t_{l-1})^{n_{ijl}}$$

where n_{ijl} is the number of subjects who are observed in state i at t_{l-1} and state j at t_l. The closed-form expression of $\mathbf{P}(s)$ as well as $\frac{\partial}{\partial \theta_u}\mathbf{P}(s)$ enables the application of a scoring rule involving only first derivatives to carry out inference about $\boldsymbol{\theta}$. The algorithm was extended to allow the transition rates to depend on covariates (Kalbfleisch and Lawless, 1985). Hubbard (2007) extended the method to nonhomogeneous Markov models. The method relies on a transformation of chronological time, $h(t)$, that yields an operational timescale on which the process is homogeneous with matrix of transition intensities $\boldsymbol{Q}_0$. Then

$$\mathbf{P}(t_1, t_2) = \mathbf{P}(h(t_2) - h(t_1)) = e^{\mathbf{Q_0}(h(t_2) - h(t_1))}.$$

At its core these methods rely on the Markov assumption, and hence sojourn times in each state are modeled as exponential, where the exponential parameters may depend on various factors. However, many disease processes are observed to progress in a way that exhibits non-constant hazard (Weiss and Zelen, 1965; Kang and Lagakos, 2007). Hence, methods that do not rely on the Markov assumption are needed.

1.3.3 Panel data: semi-Markov models for progressive processes.

Under intermittent observation of subjects, the sequence of disease states and corresponding sojourn times are not necessarily known. The inherent missing information in panel data makes estimation of semi-Markov processes more challenging.

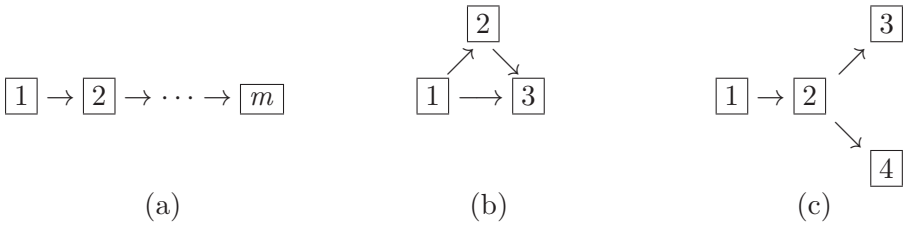

Figure 1.2: Examples of types of state models for which estimation for panel data is simplified: (a) simple progressive, (b) illness-death model, (c) progressive with competing risks.

These difficulties can be overcome in some state models. In a simple progressive model with m states (see Figure 1.2a), which gives rise to chain-of-events data, events are assumed to occur in a prescribed sequence. This implies that the probability matrix of the embedded Markov chain is degenerate with

$$p_{ij} = \begin{cases} 1, & \text{for } j = i+1,\ i = 1, \ldots, m-1; \text{ and } i = j = m; \\ 0, & \text{otherwise.} \end{cases}$$

Thus, for a simple progressive model, it remains only to estimate the sojourn time distribution in each of these states. Assuming that each conditional sojourn time distribution is absolutely continuous leads to a likelihood with convolution products of the conditional sojourn densities $f_{i,i+1}(\cdot;\boldsymbol{\theta_i})$ and survival distributions $S_i(\cdot;\boldsymbol{\theta_i})$ for $i = 1, \ldots, m-1$. Specifically, expressing the data as in Kalbfleisch and Lawless (1985), or equivalently in the "sufficient" form through $\mathbf{t}$, a vector of length $2(m-1)$, we represent successive pair of components for observation times preceding and following a time of transition. For example, considering $m = 3$, the components of $\mathbf{t}$ represent:

t_1 : last observed time in state 1,
t_2 : first observed time in state 2,
t_3 : last observed time in state 2, and
t_4 : first observed time in state 3,

where t_2, t_3, and t_4 may or may not be defined, depending on how each subject was censored. With this notation, the likelihood of the parameters given panel observations on N subjects is given by

$$L(\boldsymbol{\theta_1}, \boldsymbol{\theta_2} | \mathbf{t_1}, \ldots, \mathbf{t_N}) = \prod_{i=1}^{N} L_1^{(1-\delta_i)\cdot(1-\epsilon_i)} \cdot L_2^{\delta_i\cdot(1-\epsilon_i)} \cdot L_3^{(1-\delta_i)\cdot\epsilon_i} \cdot L_4^{\delta_i\cdot\epsilon_i},$$

where the likelihood contributions are given by

$$\begin{aligned} L_1 &= \int_{t_3}^{t_4}\int_{t_1}^{t_2} f_{12}(u_1)\cdot f_{23}(u_2-u_1)du_1du_2 \\ L_2 &= \int_{t_1}^{t_4}\int_{t_1}^{u_2} f_{12}(u_1)\cdot f_{23}(u_2-u_1)du_1du_2 \\ L_3 &= \int_{t_3}^{\infty}\int_{t_1}^{t_2} f_{12}(u_1)\cdot S_{23}(u_2-u_1)du_1du_2 \\ L_4 &= \int_{t_1}^{\infty} S_{12}(u)du, \end{aligned}$$

and δ_i and ϵ_i are indicators that subject i was not observed in states 2 and 3 respectively. That is, L_1 is the contribution of a subject who was observed in all three states; L_2 represents a subject observed in states 1 and 3 only; and L_3 and L_4 represent subjects who were right-censored in state 2 and state 1 respectively.

Considering the case of a three-state simple progressive model, De Gruttola and Lagakos (1989) proposed a nonparametric approach to estimate the conditional distributions of the sojourn times in states 1 and 2. Their approach, an extension of the self-consistency algorithm of Turnbull (1976) for univariate survival data, involves modeling the two sojourn time distributions as discrete random variables. Assuming the process enters state 1 at time zero, they let Y_1 and $Z = Y_1 + Y_2$ denote the transition times into states 2 and 3 respectively, and (Y_L, Y_R, Z_L, Z_R) be the "sufficient data" for a single realization of the process, i.e. the observation times immediately preceding and following the two transitions. This notation is similar to $\mathbf{t}$ introduced above. The authors choose locations of the mass points of Y_1 and Y_2, $0 \le y_{11} < \cdots < y_{1r}$ and $0 \le y_{21} < \cdots < y_{2s}$ respectively, and note that the observation (Y_L, Y_R, Z_L, Z_R) uniquely determines a set of "admissible values" of (y_{1j}, y_{2k}). To recast the data in this format they define α_{jk} as the indicator that (y_{1j}, y_{2k}) is an admissible value of (Y_1, Y_2). With $w_{1j} = P(Y_1 = y_{1j})$ and $w_{2k} = P(Y_2 = y_{2k})$, the likelihood is given by

$$L(\mathbf{w_1}, \mathbf{w_2}) = \prod_{i=1}^{N}\left(\sum_{j=1}^{r}\sum_{k=1}^{s}\alpha_{jk}^{i} w_{1j} w_{2k}\right),$$

where $\mathbf{w_1} = (w_{11}, \ldots, w_{1r})'$ and $\mathbf{w_2} = (w_{21}, \ldots, w_{2s})'$. It can be shown that the self-consistent estimate is equivalent to that using the *EM algorithm* (Dempster, Laird, and Rubin, 1977).

In parallel with De Gruttola and Lagakos (1989), Frydman (1992) extended the algorithm of Turnbull (1976) to a three-state simple progressive model, but assumed the underlying process was nonhomogeneous Markov rather than homogeneous semi-Markov. Frydman additionally assumed that entry into the third state was either observed exactly or right-censored. She applied her method to the HIV application in De Gruttola and Lagakos (1989) and obtained similar results except in the inference regarding the sojourn time in the infected state before developing AIDS symptoms.

Although the method of De Gruttola and Lagakos avoids imposing distributional assumptions on the sojourn times in each state, it has several drawbacks. First, it implicitly assumes that each subject was observed at least once in every state. Thus, a subject who was observed in only states 1 and 3 would need to be discarded. This leads to biased estimation where the magnitude of the bias increases with the interval between observations. Second, the method involves discretizing the two sojourn times. This means that decisions must be made about the location of the mass points of Y_1 and Y_2. Though certain guidelines may be used to avoid lack of identifiability and loss of information, the implementation of the method requires each subject with at least one admissible value of $(\mathbf{y_1}, \mathbf{y_2})$.

A slightly more general state model that arises in many applications is the *illness-death* model shown in Figure 1.2b. This model is useful for studying an incurable, potentially fatal disease. Beginning in a state of good health, subjects at risk may progress to illness, or may die from another cause or they may die from the illness. When the third state represents death, it may be assumed that transitions to this state are observed exactly. Alternatively, the three states may represent stages of disease that do not necessarily occur in a prescribed sequence. For example, Chang, Chuang, and Hsiung (2001) consider modeling the risk of breast cancer over time among women who may or may not have had benign breast disease. Though they assumed the transition times were known exactly, in this application each of these two transitions may be subject to interval censoring. Developing methods for this model is more challenging than for a simple three-state progressive process because a subject with observed trajectory $1, 1, 3$, for example, may or may not have visited state 2.

Methods for estimation and inference exist for more general progressive state models, but are often tailored to an application and impose assumptions, specific to the application, which simplify estimation. Common assumptions are that some of the transition times are known exactly or that subjects are observed at least once in each state they visit. We examine several of these methods here.

Frequently in applications there is a need to consider multiple absorbing states, or competing risks; an example of a state model accounting for this feature is shown in Figure 1.2c. Foucher, Giral, Soulillou, and Daures (2007) considered a slightly more complicated five-state progressive disease process with competing risks to model patients' natural history following kidney transplantation. The authors modeled sojourn times in each state as exponentiated Weibull, a parametric form that allows the conditional hazard function to be nonmonotonic. However, the assumption that subjects were observed in each visited state implied that estimation of the embedded Markov chain was trivial. Additionally, they did not account for the interval censoring of intermediate states.

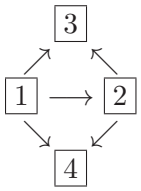

Figure 1.3: State model of Foucher et al. (2010). State numbers indicate (1) baseline value of creatinine clearance (CL); (2) decreased CL; (3) return to dialysis; (4) death.

If the number of states in the progressive process is small, all possible trajectories can be accounted for when deriving the likelihood. Foucher, Giral, Soulillou, and Daures (2010) carry out this procedure for the state model shown in Figure 1.3, where states 1-4 are defined by the patient's baseline creatinine clearance (CL); decreased CL; return to dialysis; and death with a functional transplant. They assume that the times of each patient's entry into state 1 and entry into state 3 or 4 (if applicable) are known exactly. The time of entry into state 2 is interval-censored, and a patient who is not observed in state 2 may or may not have entered state 2. Similar to Foucher et al. (2007), the authors impose a exponentiated Weibull form on the sojourn times, and build the likelihood from each of the four possible trajectories through the state space, using convolution products to deal with censoring of state 2. They obtain maximum likelihood estimates of the exponentiated Weibull parameters and of the transition probabilities of the embedded Markov chain. The authors additionally incorporate covariates and derive a goodness-of-fit statistic to test homogeneity of the semi-Markov process. Their method for modeling progressive disease could

be generalized somewhat: it could be adapted for other fairly simple state diagrams and, as they note, it could be modified to handle interval-censored absorbing states. However, numerical maximization of the likelihood is quite computationally expensive when the likelihood contributions involve more than two interval-censored times.

1.3.4 Panel data: semi-Markov models for more general processes.

Several authors have developed methods for modeling non-progressive processes without using the Markov assumption, but have imposed other strong assumptions. Specifically, Kang and Lagakos (2007) propose a method to model a nonprogressive homogeneous semi-Markov process subject to interval censoring as well as misclassification of states. Their model, however, assumes that transition intensities from at least one state were duration-independent. This assumption implies the existence of a set of states for which the Markov assumption applies and allows for great simplification of the likelihood function.

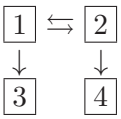

Figure 1.4: State model of Kang and Lagakos (2007) investigating cervical intraepithelial neoplasia (CIN) and human papillomavirus (HPV) infection. State 3 represents CIN diagnosis after a visit in which the patient was not infected with HPV (state 1), while state 4 represents a CIN diagnosis was HPV infected (state 2).

Some authors utilized semi-Markov models for nonprogressive processes with just two states as shown in Figure 1.5a. Since the embedded Markov chain has deterministic transition probabilities $p_{ij} = I_{\{i=j\}},\ i, j \in \{1, 2\}$, the task at hand is to estimate the sojourn time distributions. Mitchell, Hudgens, King, Cu-Uvin, Lo, Rompalo, Sobel, and Smith (2011) proposed an approach to estimate the duration of the HPV infection given panel observations of infection status assuming that: (1) all subjects are initially in state 1, (2) the Markov assumption is satisfied when the process is in state 1, and (3) subjects are observed at prespecified, equally-spaced, common visit times (e.g. every six months). Given these assumptions, the sojourn time in state 1 is a geometric random variable with point masses at the scheduled

visit times, say $1, 2, \ldots, n_t$, where n_t is the number of possible observed time points after study entry. Let p_{12} denote the associated parameter, so that the probability of spending t units of time in state 1 before transitioning to state 2 is given by $p_{12} \cdot (1-p_{12})^{t-1}$ for $t = 1, 2, \ldots$. The sojourn time in state 2 is a discrete random variable with point masses at these common scheduled visit times, so that the probability of spending time t in state 2 before transitioning to state 1 is $p_{21}(t)$ for $t = 1, 2, \ldots, n_t$ with $\sum_{t=1}^{n_t} p_{21}(t) = 1$. Letting $\mathbf{p} \doteq \{p_{12}; p_{21}(1), \ldots, p_{21}(n_t)\}$ and imposing the above restriction on $p_{21}(1), \ldots, p_{21}(n_t)$, each individual's likelihood contribution is

$$\pi_{y_0,\ldots,y_{n_t}}(\mathbf{p}) \quad = \quad p_{j_0 j_1}(x_1) \cdots p_{j_{m-1} j_m}(x_m) \cdot S_{j_m}(x_m+),$$

where $j_0, j_1, \ldots$ is the sequence of visited states, $x_0 = 0, x_1, x_2, \ldots$ is the sequence of corresponding sojourn times, $y_0, y_1, \ldots$ is the sequence of observed states at each time point, m is the number of states visited by visit n_t, x_m+ is the right-censored time spent in the final state, $S_{J_n}(\cdot)$ is the survival function in state J_n, and $p_{J_{n-1}J_n}(t) = P(J_n = j, X_n = t | J_{n-1} = i)$.

The authors allowed for isolated missing visits assuming they occurred at random (MAR assumption), but excluded subjects with two or more consecutive missing visits. Under the MAR assumption, the likelihood for N subjects is given by

$$L(\mathbf{p}) \quad = \quad \prod_{i=1}^{N} \sum_{y_{n_t} \in \{0,1\}} \cdots \sum_{y_0 \in \{0,1\}} \alpha^i_{y_0,\ldots,y_{n_t}} \cdot \pi_{y_0,\ldots,y_{n_t}}(\mathbf{p}),$$

where $\alpha^i_{y_0,\ldots,y_{n_t}}$ is the indicator that $\{y_0, \ldots, y_{n_t}\}$ is an "admissible" observation in the sense of De Gruttola and Lagakos (1989), given the possibility of missing observations at scheduled visit times. Mitchell et al. maximized the log likelihood over the parameter space via a quasi-Newton algorithm.

Mitchell et al. extended this method to make a distinction between subjects who have not been infected since time zero and those who have been infected and cleared the infection while on the study, to allow for the possibility that previous infection influences the rate by which subjects transition into the infected state. Specifically, they considered the three-state model shown in Figure 1.5b. They assumed that subjects were in state 1^* at time zero, and additionally that both states 1 and 2 were Markov. They extended this method to relax the assumption that state 2 was Markov, but noted that relaxing this assumption for state 1^* would be challenging because times in the first observed state are subject to left censoring.

Although Mitchell et al. cited neither Turnbull (1976) nor De Gruttola and Lagakos (1989), their primary approach is a self-consistency algorithm,

(a) (b)

Figure 1.5: State models considered by Mitchell et al. (2011) in a study of duration of HPV infection. In the primary method, states (1) and (2) represent the uninfected and infected states respectively, as shown in (a). In the extension, the uninfected state was split into never infected (1*) and previously infected (1), as shown in (b).

and is very similar to the method of De Gruttola and Lagakos. The present method has advantages such as not imposing distributional assumptions on the sojourn time in the infected state, but has stringent assumptions on the observational scheme: since it models the process as a discrete-time semi-Markov chain, the method in its present form requires that subjects are observed at a common set of evenly-spaced visits. As a result of modeling time discretely, this method is subject to some of the same issues as De Gruttola and Lagakos (1989) is. Mitchell et al. compared their method with that of Kang and Lagakos (2007), and noted that their own method imposes no parametric assumptions or guarantee times on the sojourn time in the infected state. However, the discrete nature of the method, itself, imposes a guarantee time on this sojourn time since there is no point mass at time zero.

Alternatively, some authors estimate semi-Markov processes by embedding a (latent) Markov process. This artifact simplifies the likelihood evaluation. Crespi, Cumberland, and Blower (2005) considered modeling herpes simplex virus type 2 (HSV-2), which is characterized by recurrent lesions. However, the number of lesions is not observable; only the presence or absence of lesions, known as the viral shedding status, can be ascertained. The viral shedding status itself is often asymptomatic and is therefore observed only at clinic visits, giving rise to panel data.

The latent number of lesions at a point in time can be considered a *birth-death process*, a Markov process in which the states, $0, 1, 2, \ldots$ represent the size of the population at each point in time (see Figure 1.6a). The authors

assume that the process has transition intensities

$$Q_{ij}(t) = \begin{cases} \lambda, & j = i+1,\ i = 0, 1, 2, \ldots; \\ \mu, & j = i-1,\ i = 1, 2, 3, \ldots; \\ 0, & \text{otherwise,} \end{cases}$$

for all $t \geq 0$, for some $\lambda, \mu > 0$, where λ represents the rate at which lesions are formed, and μ represents the rate at which they are cleared. Implicit in this model is the assumption that lesions form independently of each other. If states $1, 2, 3, \ldots$ of this homogeneous Markov process are collapsed into a single state, denoted 1+, the corresponding process, denoted $Z(\cdot)$, is semi-Markov, as the sojourn time in state 1+ now depends on the elapsed time in this state (Figure 1.6b). While the distribution of the sojourn time in the non-shedding state is exponential with rate λ, the sojourn time in the shedding state does not follow a familiar distribution. The observed viral shedding status at each point in time is an induced semi-Markov model.

(a) $\boxed{0} \leftrightarrows \boxed{1} \leftrightarrows \boxed{2} \leftrightarrows \cdots$ hidden Markov process, $W(\cdot)$

(b) $\boxed{0} \leftrightarrows \boxed{1+}$ semi-Markov process, $Z(\cdot)$

Figure 1.6: State models considered by Crespi et al. (2005). The unobservable number of recurrences at time t is modeled as a birth-death process (a), while the viral shedding status is modeled as the corresponding semi-Markov process (b) defined by collapsing states $1, 2, 3, \ldots$ of the birth-death process.

Crespi et al. express the panel data likelihood via a hidden Markov model approach and carry out inference on the parameters in a Bayesian framework. Moreover, they use a random effects model to accommodate heterogeneity across individuals, and express the mean of each random effect as a function of covariates. Posterior distributions of λ and μ allow for inference on both the hidden Markov process $W(\cdot)$ and the semi-Markov process $Z(\cdot)$. Titman and Sharples (2010) also consider a hidden Markov process which accounts for a classification error in the assessment of the state at each observation time.

1.3.5 Bayesian approach to multi-state models

Our review of the literature indicated that while methods for estimating Markov models under both continuous and intermittent observation are well-established, that is not the case for semi-Markov models, especially with

panel data. Further, because most methods are tied to particular applications, they do not generalize to more complex state models. In particular, methods for more general state models often require additional modeling assumptions to simplify the likelihood evaluation or to allow for parameter identifiability. Finally, the vast majority of methods for multi-state models are frequentist.

The Bayesian approach offers a viable alternative to estimating multi-state models. At the core of the Bayesian approach one expresses the uncertainty about all unknowns with prior distributions. Thus, Bayesian estimation of the homogeneous Markov processes proceeds by assuming priors on the initial state distribution $\boldsymbol{\phi}$ of the process (here a natural choice is the Dirichlet distribution for its conjugacy property) as well as on the transition intensities q_{ij}. Oftentimes, we express the transition intensities as dependent on covariates via a proportional hazards formulation as discussed in Section 1.3.1 in which case we assume priors on the regression coefficients $\boldsymbol{\beta}_{jk}$ instead. A similar approach can be used for nonhomogeneous Markov processes as we illustrate in Section 1.4.1.

Likewise, Bayesian estimation of continuously observed semi-Markov processes requires priors for the initial state distribution, but also on the transition probabilities of the embedded Markov chain and on the parameters governing the sojourn time distributions. Alternatively, when the interest lies in estimating the effect of a given set of risk factors on the disease state transitions or sojourn times, then one assigns priors on the corresponding regression parameters. When the semi-Markov process is, however, observed intermittently we are faced with the additional challenge that the full trajectory of the disease process or the durations in each state are unknown. Bayesian estimation proceeds by modeling the unobserved states or durations in the state as latent. We illustrate this method in Section 1.4.2.

Generally the posterior distributions of the model parameters are not available in closed form, but estimation can be accomplished using Markov chain Monte Carlo (MCMC) methods. While some computational challenges may arise, for example, in devising algorithms that allow for good mixing, the Bayesian approach offers several advantages in multi-state modeling. It allows us to incorporate available knowledge about disease processes via expert information, for example, which gives us the potential to address identifiability issues that arise given the sparsity of information in panel data. Furthermore, it provides exact inference that does not rely on asymptotic results. Although the Bayesian approach does not eliminate methodological issues arising with panel observation, it can enhance our ability to estimate models which are intractable under classical approaches.

1.4 Bayesian multi-state models: Applications

In this section, we illustrate the use of the Bayesian approach to multi-state modeling in the context of two applications to specific disease processes.

1.4.1 Nonhomogeneous Markov models: Modeling delirium in stem cell transplant recipients.

Delirium is an acute neuropsychiatric condition associated with rapid onset of a change in levels of consciousness, attention, cognition, and perception. Cancer patients are at particularly high risk for delirium due to the presence of a variety of risk factors in this population including age, emotional distress, cognitive impairment caused by chemotherapeutic agents, and organ failure (Fann and Sullivan, 2003). Prevalence of delirium among cancer patients has been estimated at 25-40% and as high as 85% among terminally ill cancer patients (Fann, 2000). The extremely high prevalence of delirium in this patient population makes it of particular interest.

Researchers studied incidence and progression of delirium in a cohort of 90 patients treated at the Fred Hutchinson Cancer Research Center (FHCRC) in Seattle, Washington, between 1997 and 1999. Subjects were assessed for delirium using the Delirium Rating Scale (DRS) (Trzepacz et al., 1988) on average every 2.5 days during the first thirty days following transplant. Existing work has detailed incidence of delirium, risk factors for delirium incidence and severity, and association of long term health outcomes with delirium incidence in this cohort (Fann et al., 2002, 2005, 2007). In several previous studies of this cohort, delirium was treated as a dichotomous outcome and the incidence rate for delirium was assumed constant across the observation period. However, delirium is characterized by a complex, fluctuating course that may be more fully understood by using a multi-state model that allows for both clinical and sub-clinical delirium states. The state model for the delirium disease process is presented in Figure 1.7. To investigate temporal variability in onset and progression of delirium we modeled the disease via a three state nonhomogeneous model using the time-transformation method of Hubbard et al. (2008).

Parameters of the three-state process described above can be estimated using maximum likelihood or Bayesian methods. An analysis of these data using maximum likelihood methods identified a statistically significant acceleration of the disease process over time (Hubbard et al., 2008). However, if we are additionally interested in allowing for subject-specific variation in the rate of acceleration of the process, maximum likelihood methods become

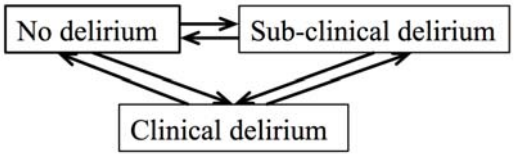

Figure 1.7: State model for delirium process.

intractable. Subject-specific variation in the rate of progression of the disease is of interest in this context because subjects included in the study are extremely heterogeneous. Variation exists in demographics, treatments, and cancer-types in these patients. Many unmeasured factors may influence the course of delirium in individual patients, and patients may possess intrinsic variations in disease susceptibility and recovery. We can account for variation in the rate of evolution of the disease process by introducing random effects into the time-transformed nonhomogeneous Markov process. We investigated subject-specific differences in the nonhomogeneity of the delirium process by applying a random effects model that allows for between-subject variability in the time transformation parameter. Specifically, the likelihood for this model takes the form

$$L^{(m)}(\boldsymbol{Q}_0, \boldsymbol{\theta}) = \prod_{i=1}^{m} \left\{ P(X(u_{i1}) = x_{i1}) \prod_{j=2}^{n} \left\{ e^{\boldsymbol{Q}_0 (h(u_{ij};\boldsymbol{\theta}_i) - h(u_{ij-1};\boldsymbol{\theta}_i))} \right\}_{x_{ij-1} x_{ij}} \right\}, \tag{1.1}$$

where $h(u; \boldsymbol{\theta}_i)$ is a function that transforms the time-scale of the process from the observed time scale, on which the process in nonhomogeneous, to an operational time-scale on which the process is assumed homogeneous. To allow for subject-specific variation in the rate of evolution of the process, we introduce a subject-specific time-transformation parameter, $\boldsymbol{\theta}_i$, which allows for between-subject variability in the rate of evolution of the process.

Bayesian estimation is straightforward by introducing priors for the transition intensities and time-transformation parameters. In our delirium application, we assumed independent log normal priors for the elements of $\boldsymbol{Q}_0$ and $\boldsymbol{\theta}$,

$$\begin{aligned} \pi(\log q_{0ab}) &\sim \text{Normal}(\mu_q, \sigma_q^2), \quad a \neq b \\ \pi(\log \theta_{ik}) &\sim \text{Normal}(\mu, \sigma^2). \end{aligned}$$

In order to limit the dimensionality of the posterior density of model parameters, we used a time transformation function with a single parame-

ter, $h(u;\theta) = u\theta^u$. We placed a normal prior on μ and inverse gamma prior on σ^2. Hyperparameter values for prior densities were selected according to expert information on the likelihood of observing each kind of transition. Estimation was carried out using a hybrid Gibbs/Metropolis-Hastings MCMC simulation.

Prior and posterior densities for transition intensities are presented in Figure 1.8. The posterior median for μ and σ^2 are very close to prior means due to the strong prior distributions placed on these parameters since only a modest amount of information about individual level time transformation parameters is available due to the relatively small number of observations per subject (between 7 and 18). Example time transformation curves for three subjects with largest and three subjects with smallest posterior median time transformation parameters are presented in Figure 1.9. Subject-specific time transformation parameters indicate slowing of the disease process for some subjects and more rapidly evolving disease processes for others.

1.4.2 Semi-Markov models: Modeling HIV infection and progression to AIDS-related symptoms.

Development of the method of De Gruttola and Lagakos (1989) was motivated by a retrospective study of a cohort of 262 patients at the Hôpital Kremlin Bicêtre and Hôpital Cœur des Yvelines in France. These patients had type A or B hemophilia and received periodic blood transfusions that were later found to be HIV contaminated. Blood samples taken at various times allowed for intermittent retrospective assessment of HIV infection status.

De Gruttola and Lagakos chose a simple progressive three-state model for this situation, as depicted in Figure 1.10. They split the patients into two groups defined by the amount of blood product they had received, which we will refer to as the "heavily treated" and "lightly treated" groups. For each of these two groups, De Gruttola and Lagakos carried out separate estimation of the time to infection and the time to progression to AIDS. Of the 262 patients in the cohort, 197 were infected with HIV at the end of follow-up, 43 of whom had progressed to AIDS. All HIV infections were believed to have been caused by receiving contaminated blood. De Gruttola and Lagakos divided the chronological time axis into 6-month intervals, with $Y = 1$ denoting July 1, 1978. Each interval was given a point mass, and the point masses were indexed by $1, 2, 3, \ldots$. De Gruttola and Lagakos chose a simple progressive three-state model for this situation, as depicted in Figure 1.10.

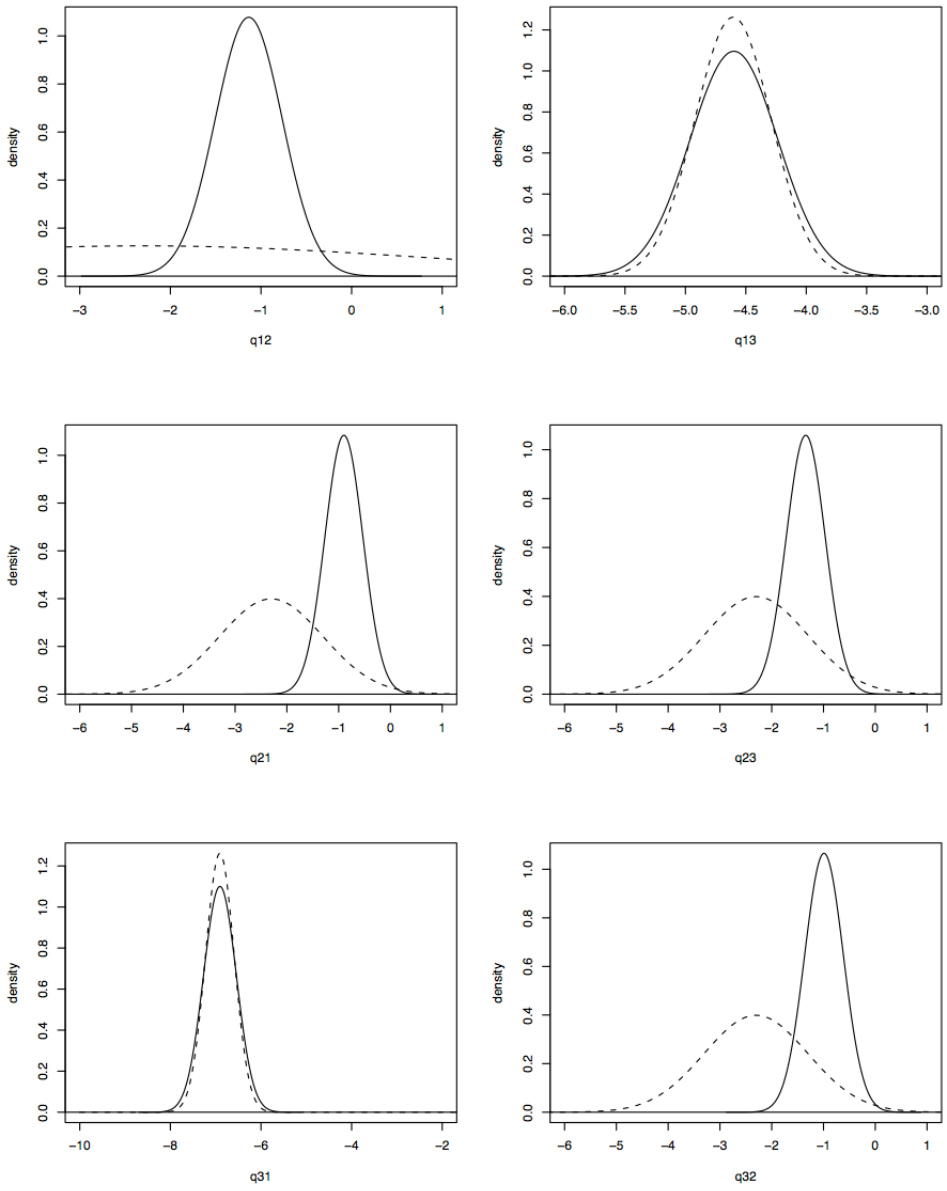

Figure 1.8: Prior (dashed) and posterior (solid) densities for transitions intensities in time transformed random effects model for delirium data.

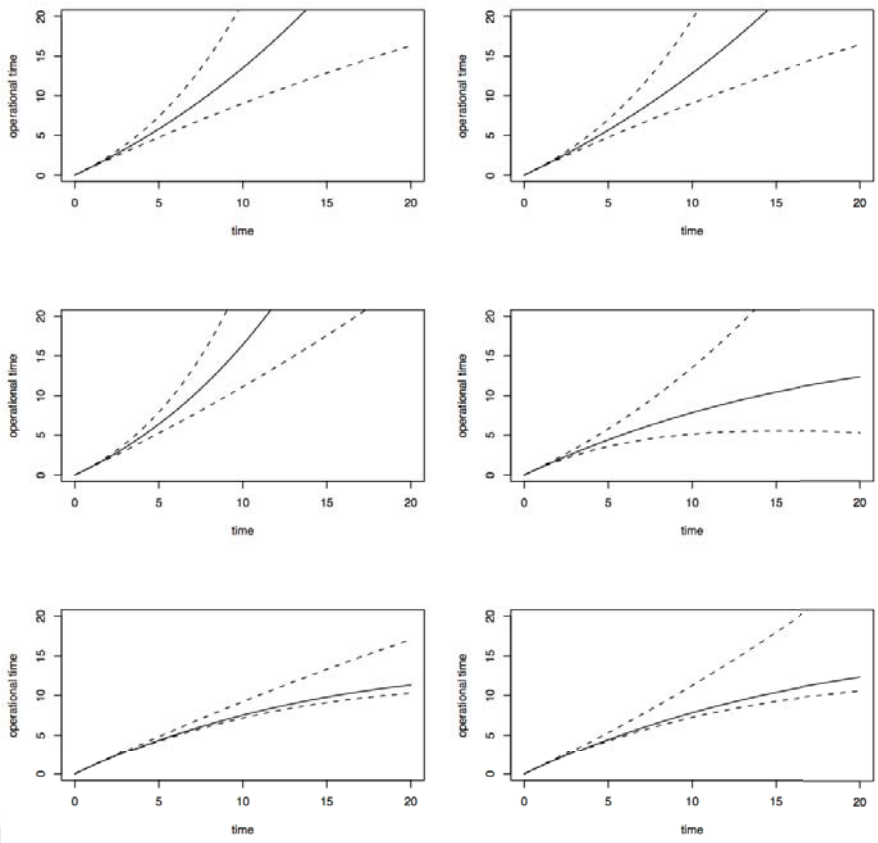

Figure 1.9: Estimated time transformation functions and 95% credible bands (dashed lines) for six stem cell transplantation patients with most extreme posterior median time transformation parameters from random effects time transformation model.

Figure 1.10: State model of De Gruttola and Lagakos (1989).

We will use the data from De Gruttola and Lagakos and consider the estimation of the simple progressive three-state model under the Bayesian approach. The intermittent observation scheme creates a "missing" data problem since the sojourn times in each state, as well as the sequence of states, may not be observed. We can treat the true, unobservable, sojourn times as *latent data* and use data augmentation procedures (Tanner, 1991) to assist inference about the parameters of interest. Assume, as in De Gruttola and Lagakos, that all subjects start in state 1 at time zero. Let the random variable $X_i \geq 0$ denote the true unobservable sojourn time in state i before proceeding to state $i+1$, where $i = 1, 2$. Let $\mathbf{X} = (X_1, \ldots, X_2)$ and assume an absolutely continuous parametric form for the sojourn time in each state: $X_i \sim f_i$ for $i = 1, 2$, and assume that the densities f_1 and f_2 collectively depend on the vector of parameters $\boldsymbol{\theta}$.

We observe the process periodically, giving rise to panel observations $\mathbf{Z} = (Z_0 = 1, Z_1, \ldots, Z_n)$, with $Z_i \in \{1, 2, 3\}$ for each i, corresponding to observation times $\mathbf{s} = (s_0 = 0, s_1, \ldots, s_n)$. Note that in the simple progressive model, $Z_0 \leq \cdots \leq Z_n$. Note also that $\mathbf{Z}$ contains redundant information, and can be expressed equivalently as the vector $\mathbf{t} = (t_1, \ldots, t_4)$, where

$$\begin{aligned} t_1 &= \max\{s_k : Z_k = 1\} \\ t_2 &= \min\{s_k : Z_k = 2\} \\ t_3 &= \max\{s_k : Z_k = 2\} \\ t_4 &= \min\{s_k : Z_k = 3\}, \end{aligned}$$

if each of these exists. With the above notation, the posterior and predictive equations in the data augmentation algorithm (Tanner, 1991) become

$$p(\boldsymbol{\theta}|\mathbf{t}) = \int_{\mathbf{X}} p(\boldsymbol{\theta}|\mathbf{t}, \mathbf{X}) p(\mathbf{X}|\mathbf{t}) d\mathbf{X} \tag{1.2}$$

$$p(\mathbf{X}|\mathbf{t}) = \int_{\boldsymbol{\theta}} p(\mathbf{X}|\mathbf{t}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{t}) d\boldsymbol{\theta}. \tag{1.3}$$

From (1.2)–(1.3), the goal is to obtain $p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{t})$ and $p(\mathbf{X}|\mathbf{t}, \boldsymbol{\theta})$. The particular form of each of these expressions depends on the choice of the model for the sojourn time distributions in each state. We complete the model by specifying priors for parameters $\boldsymbol{\theta}$.

We apply the proposed data augmentation approach to this dataset and compare the results to those obtained by the method of De Gruttola and Lagakos. Specifically, we consider the exponential and Weibull models for the sojourn times in the HIV-uninfected and -infected states, and carry out inference about the corresponding parameters separately for the heavily and lightly treated patients. We assume independent noninformative uniform priors for the rate parameters (under both exponential and Weibull models) and noninformative uniform priors for the shape parameters (Weibull model).

Results are presented for heavily and lightly treated subjects in Figure 1.11 in the form of estimated cumulative distribution functions so that they may be compared with those obtained originally by De Gruttola and Lagakos. Figure 1.11 shows that the Weibull model, unlike the exponential, has enough flexibility to accommodate the shape of the hazard function in each state, as it produced an estimated cumulative distribution function similar to that of the method of De Gruttola and Lagakos (1989).

1.5 Discussion

Natural history of disease can be modeled using a variety of approaches that fall under the general framework of multi-state models, including Markov processes, nonhomogeneous Markov processes, semi-Markov processes, and hidden Markov processes. Simple approaches like the Markov process facilitate estimation but make use of strong assumptions that may be unrealistic for certain diseases. More complex models such as semi-Markov processes may more accurately describe the disease process, but at the cost of substantially complicating estimation, especially when data arise from panel observation. In this chapter, we have discussed a few methods for estimating disease processes. In particular, we illustrated that the Bayesian approach can be used to estimate some of these more complex processes which would be intractable under the maximum likelihood framework.

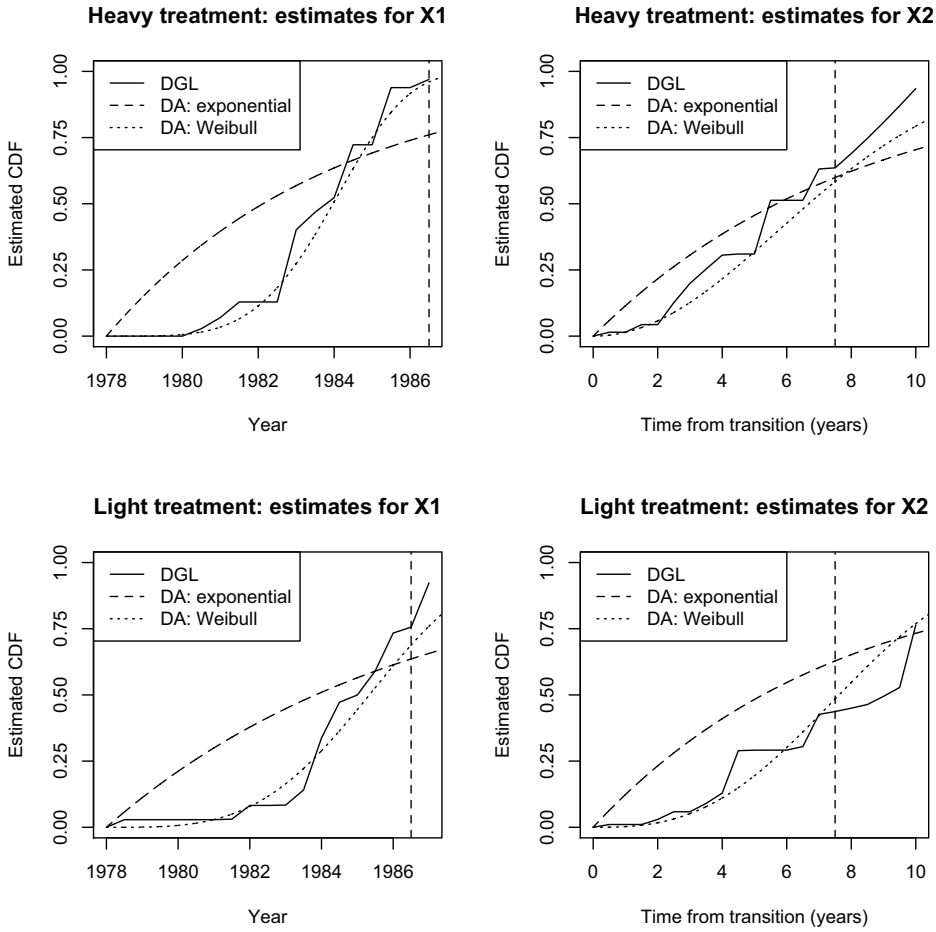

Figure 1.11: Estimated cumulative distribution functions (CDFs) of the sojourn times in the uninfected and infected states based on the proposed data augmentation (DA) method. Results are shown for heavily treated (upper panels) and lightly treated subjects (lower panels), based on exponential and Weibull models of the sojourn times in each state. Results from the method of De Gruttola and Lagakos are shown for reference.

Bibliography

Who case definitions of hiv for surveillance and revised clinical staging and immunological classification of hiv-related disease in adults and children. Technical report, World Health Organization, 2007.

O. Aalen. Nonparametric inference for a family of counting processes. *Annals of Statistics*, 6(4):701–726, 1978.

A. Albert. Estimating the infinitesimal generator of a continuous time, finite state Markov process. *Annals of Mathematical Statistics*, 33(2):727–753, 1962.

E.A. Alvarez. Smoothed nonparametric estimation in window censored semi-Markov processes. *Journal of Statistical Planning and Inference*, 131: 209–229, 2005.

T.W. Anderson and L.A. Goodman. Statistical inference about Markov chains. *Annals Of Mathematical Statistics*, 28(1):89–110, 1957.

P. Billingsley. *Statistical Inference for Markov Processes*. University of Chicago Press, Chicago, 1961.

B.J.N. Blight. Estimation from a censored sample for the exponential family. *Biometrika*, 57(2):389–395, 1970.

B.P. Carlin and S. Chib. Bayesian model choice via markov chain monte carlo methods. *Journal of the Royal Statistical Society, Series B*, 57(3): 473–484, 1995.

AIDS Education & Training Centers National Resource Center. Hiv classification: Cdc and who staging systems. 2012.

I.-S. Chang, Y.C. Chuang, and C.A. Hsiung. Goodness-of-fit tests for semi-Markov and Markov survival models with one intermediate state. *Scandinavian Journal of Statistics*, 2001.

C.L. Chiang. *An Introduction to Stochastic Processes and their Applications*. R.E. Krieger, New York, 1980.

R.E. Colvert and T.J. Boardman. Estimation in the piece-wise constant hazard rate model. *Communications in Statistical Theory and Methods*, A5(11):1013–1029, 1976.

D. Commenges, P. Joly, L. Letenneur, and J.F. Dartigues. Incidence and mortality of alzheimer's disease or dementia using an illness-death model. *Statistics in Medicine*, 23:199–210, 2004.

Sarah J. Converse, J. Andrew Royle, and Richard P. Urbanek. Bayesian analysis of multi-state data with individual covariates for estimating genetic effects on demography. *JOURNAL OF ORNITHOLOGY*, 152(2): S561–S572, FEB 2012. ISSN 0021-8375. doi: 10.1007/s10336-011-0695-0. 9th EURING Analytical Meeting, Ist Superiore Protezione Ricerca Ambientale (ISPRA), Pescara, ITALY, SEP 14-20, 2009.

C.M. Crespi, W.G. Cumberland, and S. Blower. A queueing model for chronic recurrent conditions under panel observation. *Biometrics*, 61: 193–198, 2005.

Getachew A. Dagne and James Snyder. Bayesian hierarchical duration model for repeated events: an application to behavioral observations. *JOURNAL OF APPLIED STATISTICS*, 36(11):1267–1279, 2009. ISSN 0266-4763. doi: 10.1080/02664760802587032.

V. De Gruttola and S.W. Lagakos. Analysis of doubly-censored survival data. *Biometrics*, 45:1–11, 1989.

C. Del Rio and J.W. Curran. *Mandell, Douglas, and Bennetts Principles and Practice of Infectious Diseases*, chapter 121. Saunders Elsevier, 7 edition, 2009.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–22, 1977.

Mansson F. Kvist A. Isberg P.-E. Nowroozalizadeh S. Biague A.J. da Silva Z.J. Jansson M. Fenyo E.M. Norrgren H. Esbjornsson, J. and P. Medstrand. Inhibition of hiv-1 disease progression by contemporaneous hiv-2 infection. *New England Journal of Medicine*, 367(3):224–232, 2012.

J. R. Fann. The epidemiology of delirium: A review of studies and methodological issues. *Seminars in Clinical Neuropsychiatry*, 5(2):64–74, 2000.

J. R. Fann and A. K. Sullivan. Delirium in the course of cancer treatment. *Seminars in Clinical Neuropsychiatry*, 8(4):217–28, 2003.

J. R. Fann, S. Roth-Roemer, B. E. Burington, W. J. Katon, and K. L. Syrjala. Delirium in patients undergoing hematopoietic stem cell transplantation: Incidence and pretransplantation risk factors. *Cancer*, 95(9): 1971–1981, 2002.

J. R. Fann, C. M. Alfano, B. E. Burington, S. Roth-Roemer, W. J. Katon, and K. L. Syrjala. Clinical presentation of delirium in patients undergoing hematopoietic stem cell transplantation: Delirium and distress symptoms and time course. *Cancer*, 103(4):810–820, 2005.

J. R. Fann, C. M. Alfano, S. Roth-Roemer, W. J. Katon, and K. L. Syrjala. Impact of delirium on cognition, distress, and health-related quality of life after hematopoietic stem-cell transplantation. *Journal of Clinical Oncology*, 25(10):1223–1231, 2007.

Y. Foucher, M. Giral, J.-P. Soulillou, and J.-P. Daures. A semi-Markov model for multistate and interval-censored data with multiple terminal events. Application in renal transplantation. *Statistics in Medicine*, 26: 5381–5393, 2007.

Y. Foucher, M. Giral, J.-P. Soulillou, and J.-P. Daures. A flexible semi-Markov model for interval-censored data and goodness-of-fit testing. *Statistical Methods in Medical Research*, 19:127–145, 2010.

H. Frydman. A nonparametric estimation procedure for a periodically observed three-state Markov process, with application to AIDS. *Journal of the Royal Statistical Society, Series B*, 54(3):853–866, 1992.

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian data analysis*. Chapman & Hall/CRC, Boca Raton, 1995.

S.J. Godsill. On the relationship between mcmc model uncertainty methods. *J. Comp. Graph. Stat.*, 10(2):230–248, 2001.

G.S. Gottlieb, P.S. Sow, S.E. Hawes, I. Ndoye, M. Redman, A.M. Coll-Seck, M.A. Faye-Niang, A. Diop, J.M. Kuypers, C.W. Critchlow, R. Respess, J.I. Mullins, and N.B. Kiviat. Equal plasma viral loads predict a similar

rate of cd4+ t cell decline in human immunodeficiency virus (hiv) type 1- and hiv-2–infected individuals from senegal, west africa. *Journal of Infectious Diseases*, 185:905–914, 2002.

P.J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. 82(4):711–732, 1995.

Lemeshow S. Hosmer, W.H. and S. May. *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*. 2008.

R.A. Hubbard. *Modeling a non-homogeneous Markov process via time transformation*. PhD thesis, University of Washington, 2007.

RA Hubbard, LYT Inoue, and JR Fann. Modeling a non-homogeneous Markov process via time transformation. *Biometrics*, 64(3):843 – 850, 2008.

A. Iosifescu-Manu. Non-homogeneous semi-Markov processes. *Studii si Cercetari Matematice*, 24:529–33, 1972.

C. H. Jackson, L. D. Sharples, S. G. Thompson, S. W. Duffy, and E. Couto. Multistate Markov models for disease progression with classification error. *Journal of the Royal Statistical Society Series D-the Statistician*, 52(2): 193–209, 2003.

Grant A.D. Whitworth J. Smith P.G. Jaffar, S. and H. Whittle. The natural history of hiv-1 and hiv-2 infections in adults in africa: a literature review. *Bull World Health Organ*, 82(6):462–469, 2004.

J. Janssen, editor. *Semi-Markov Models: Theory and Applications*. Plenum Press, New York, 1986.

J. Janssen and R. Manca. *Applied Semi-Markov Processes*. Springer, 2006.

J.D. Kalbfleisch and J.F. Lawless. The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, 80(392):863–871, 1985.

J.D. Kalbfleisch and J.F. Lawless. Regression models for right truncated data with applications to aids incubation times and reporting bias. *Statistica Sinica*, 1:19–32, 1991.

M. Kang and S.W. Lagakos. Statistical methods for panel data from a semi-Markov process, with application to HPV. *Biostatistics*, 8(2):863–871, 2007.

E.L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Society*, 53:457–581, 1958.

R. Kay. A Markov model for analysing cancer markers and disease states in survival studies. *Biometrics*, 42(4):855–865, 1986.

M.Y. Kim, V. De Gruttola, and S.W. Lagakos. Analyzing doubly censored data with covariates, with application to aids. *Biometrics*, 49:13–22, 1993.

J.P. Klein and M.L. Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer, 1997.

Thomas Kneib and Andrea Hennerfeind. Bayesian semiparametric multistate models. *STATISTICAL MODELLING*, 8(2):169–198, JUL 2008. ISSN 1471-082X. doi: 10.1177/1471082X0800800203.

S.W. Lagakos, C.J. Sommer, and M. Zelen. Semi-Markov models for partially censored data. *Biometrika*, 65(2):311–317, 1978.

C.D. Lai, M. Xie, and D.N.P. Murthy. A modified weibull distribution. *IEEE Transactions on Reliability*, 52(1):33–37, 2003.

J.F. Lawless and Y.T. Fong. State duration models in clinical and observational studies. *Statistics in Medicine*, 18:2365–2376, 1999.

J.F. Lawless and P. Yan. Some statistical methods for followup studies of disease with intermittent monitoring. In *Multiple comparisons, selection, and applications in biometry*. Marcel Dekker, 1993.

P. Lévy. Systèmes semi-Markoviens à au plus une infinité d'états possibles. *Proc. Int. Congr. Math.*, 2:294, 1954a.

P. Lévy. Processus semi-Markoviens. *Proc. Int. Congr. Math.*, 3:416–426, 1954b.

N. Limnios and G. Oprişan. *Semi-Markov Processes and Reliability*. Birkhäuser, Boston, 2001.

L. Meira-Machado. Inference for non-Markov multi-state models: an overview. *Revstat Statistical Journal*, 9(1):83+, Mar 2011.

L. Meira-Machado, J. de Uña Álvarez, C. Cadarso-Suárez, and P.K. Andersen. Multi-state models for the analysis of time-to-event data. *Statistical Methods In Medical Research*, 18(2):195–222, Apr 2009.

C.E. Mitchell, M.G. Hudgens, C.C. King, S. Cu-Uvin, Y. Lo, A. Rompalo, J. Sobel, and J.S. Smith. Discrete-time semi-Markov modeling of human papillomavirus persistence. *Statistics In Medicine*, 30(17):2160–2170, Jul 30 2011.

G.S. Mudholkar and D.K. Srivastava. Exponentiated weibull family for analyzing bathtub failure-rate data. *IEEE Transactions on Reliability*, 42 (2):299–302, 1993.

B. Ouhbi and N. Limnios. Nonparametric estimation for semi-Markov processes based on its hazard rate functions. *Statistical Inference for Stochastic Processes*, 2:151–73, 1999.

Shin-Liang Pan, Hui-Min Wu, Amy Ming-Fang Yen, and Tony Hsiu-Hsi Chen. A Markov regression random-effects model for remission of functional disability in patients following a first stroke: A Bayesian approach. *STATISTICS IN MEDICINE*, 26(29):5335–5353, DEC 20 2007. ISSN 0277-6715. doi: 10.1002/sim.2999.

R.L. Prentice, J.D. Kalbfleisch, A.V. Peterson Jr., N. Flournoy, V.T. Farewell, and N.E. Breslow. The analysis of failure times in the presence of competing risks. *Biometrics*, 34(4):541–554, 1978.

Malcolm J. Price, Nicky J. Welton, and A. E. Ades. Parameterization of treatment effects for meta-analysis in multi-state Markov models. *STATISTICS IN MEDICINE*, 30(2):140–151, JAN 30 2011. ISSN 0277-6715. doi: 10.1002/sim.4059.

R. Pyke. Markov renewal processes: definitions and preliminary properties. *Annals of Mathematical Statistics*, 32:1231–42, 1961a.

R. Pyke. Markov renewal rprocesses with finitely many states. *Annals of Mathematical Statistics*, 32:1243–59, 1961b.

S.M. Ross. *Stochastic Processes*. Wiley & Sons, Berkeley, 2nd edition, 1996.

D.B. Rubin. *Multiple imputation for nonresponse in surveys*. 1987.

G.A. Satten and M.R. Sternberg. Fitting semi-Markov models to interval-censored data with unknown initiation times. *Biometrics*, 55(2):507–513, 1999.

W.L. Smith. Regenerative stochastic processes. *Proceedings of the Royal Society of London, Series A*, 232:6–31, 1955.

T.R. Sterling and R.E. Chaisson. *Mandell, Douglas, and Bennetts Principles and Practice of Infectious Diseases*, chapter 121. Saunders Elsevier, 7 edition, 2009.

M.R. Sternberg and G.A. Satten. Discrete-time nonparametric estimation for semi-Markov models of chain-of-events data subject to interval censoring and truncation. *Biometrics*, 55(2):514–522, 1999.

MJ Sweeting, D De Angelis, and OO Aalen. Bayesian back-calculation using a multi-state model with application to HIV. *STATISTICS IN MEDICINE*, 24(24):3991–4007, DEC 30 2005. ISSN 0277-6715. doi: 10.1002/sim.2432. 25th Annual Conference of the International-Society-for-Clinical-Biostatistics, Leiden, NETHERLANDS, AUG 15-19, 2004.

M.A. Tanner. *Tools for statistical inference: observed data and data augmentation methods*. Springer-Verlag, Heidelberg, 1991.

M.A. Tanner and W.H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82 (398):528–540, 1987.

T.M. Therneau and P.M. Grambsch. *Modeling survival data: extending the Cox model*. Springer, New York, 2000.

A.C. Titman and L.D. Sharples. Semi-Markov models with phase-type sojourn distributions. *Biometrics*, 66:742–752, 2010.

P.T. Trzepacz, R.W. Baker, and J. Greenhouse. A symptom rating scale for delirium. *Psychiatry Research*, 23(1):89–97, 1988.

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

J.G. Voelkel and J. Crowley. Nonparametric inference for a class of semi-Markov processes with censored observations. *Annals of Statistics*, 12(1): 142–160, 1984.

Brookmeyer R. Wang, M.-C. and N. Jewell. Statistical models for prevalent cohort data. *Biometrics*, 49:1–11, 1993.

J.H. Ware and D.L. DeMets. Reanalysis of some baboon descent data. *Biometrics*, 32(2):459–163, 1976.

G.H. Weiss and M. Zelen. A semi-Markov model for clinical trials. *Journal of Applied Probability*, 2:269–285, 1965.

Egboga A. Todd J. Corrah T.-Wilkins A. Demba E. Morgan G. Rolfe M. Berry N. Whittle, H. and R. Tedder. Clinical and laboratory predictors of survival in gambian patients with symptomatic hiv-1 or hiv-2 infection. *AIDS*, 6:685–689, 1992.