UCLA

UCLA Electronic Theses and Dissertations

Title

Modeling Multistate Models with Back Transitions: Statistical Challenges and Applications

Permalink

https://escholarship.org/uc/item/0g055686

Author

Aralis, Hilary Jeanne

Publication Date

2016

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

University of California

Los Angeles

Modeling Multistate Processes with Back Transitions: Statistical Challenges and Applications

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Biostatistics

by

Hilary Jeanne Aralis

2016

© Copyright by

Hilary Jeanne Aralis

2016

Abstract of the Dissertation

Modeling Multistate Processes with Back Transitions: Statistical Challenges and Applications

by

Hilary Jeanne Aralis
Doctor of Philosophy in Biostatistics
University of California, Los Angeles, 2016
Professor Ronald S. Brookmeyer, Chair

Multistate models are widely used in health research to analyze life history processes in which each individual is assumed to occupy one of a finite number of states at any given point in time. Models allowing for back transitions are necessary when considering recurrent events or disease states from which recovery is possible along with subsequent return to illness. The objective of this dissertation is to consider current challenges in the statistical analyses of multistate models arising in public health. Applications to the fields of HIV/AIDS and dementia are considered.

To assess the effect of concurrent or overlapping partnership patterns on the trajectory of the HIV epidemic in a population, it is necessary to estimate both the extent and the magnitude of concurrency. Data are typically available in the form of retrospective sexual history reports. We introduce a joint multistate and point process model in which states are defined as the number of ongoing partnerships an individual is engaged in at a given time. Sexual partnerships starting and ending on the same date are referred to as one-offs and modeled as discrete events. The proposed method treats each individual's continuation in and transition through various numbers of ongoing partnerships as a separate stochastic process and allows the occurrence of one-offs to impact subsequent rates of partnership formation and dissolution. Among a sample of men having sex with men and seeking HIV testing at a Los Angeles clinic, the estimated point prevalence of concurrency was higher among men later diagnosed HIV positive. One-offs were associated with increased rates of subsequent partnership dissolution.

In constructing a disease progression multistate model, panel data consisting of the states occupied by an individual at a series of discrete time points are often used to to estimate transition intensities of the underlying continuous-time process. When transition intensities depend on the time elapsed in the current state and back transitions between states are possible, this intermittent observation process leads to intractability of the likelihood function. We present an iterative stochastic expectation-maximization (SEM) algorithm that relies on a simulation-based approximation to the likelihood function and implement this algorithm using rejection sampling. In a simulation study, we demonstrate the feasibility and performance of the proposed procedure. We then demonstrate application of the algorithm to a study of dementia, the Nun Study, consisting of intermittently-observed elderly subjects in one of four possible states corresponding to intact cognition, impaired cognition, dementia, and death. We show that the proposed SEM algorithm substantially reduces bias in model parameter estimates compared to an alternative naive approach. We then extend the utility of this disease progression model to settings in which healthy individuals have a non-negligible probability of being misclassified into a disease state. The proposed model accomplishes unbiased estimation of the semi-Markov model parameters associated with transition rates and probabilities while simultaneously estimating the true underlying misclassification rate without requiring information from a gold standard. In applying this SEM algorithm addressing misclassification to the Nun Study, findings suggest that the rate of misclassification may be relatively high among this sample and that true back transitions from impaired to intact cognition are somewhat rare.

By describing and addressing statistical challenges in multistate modeling, we demonstrate the utility of and reduce existing barriers to the implementation of such models across a wide range of applications in the field of public health and in situations where imperfect data are available.

The dissertation of Hilary Jeanne Aralis is approved.

Robert E. Weiss

Catherine A. Sugar

Pamina M. Gorbach

Ronald S. Brookmeyer, Committee Chair

University of California, Los Angeles

2016

To my grandmother, who taught me how to live life with undying optimism, determination, and humor.

Table of Contents

List of Figures

List of Tables

Acknowledgments

A version of the material presented in Chapter 3 has been published in *Statistics in Medicine* with contributions from co-authors, Drs. Ron Brookmeyer and Pamina Gorbach [1]. Chapters 4 and 5 contain materials from a manuscript currently submitted to *Lifetime Data Analysis*.

This work was supported in part by the National Institutes of Health (NIH) Training Grant 5-T32-AI007370, Biostatistical Training in AIDS. The material presented in Chapter 3 was made possible by use of data collected as part of the MetroMates project which was funded by the National Institute of Drug Abuse (NIDA) under grant number 1R01DA022116-01A1 awarded to Dr. Pamina Gorbach at the University of California Los Angeles. The material presented in Chapters 5 and 6 was made possible by use of the Nun Study data provided by Dr. Richard Kryscio at the University of Kentucky. The Nun Study was funded by the National Institute on Aging (NIA) under grant number R01AG038651.

I am profoundly grateful for the guidance and support of my advisor, Dr. Ron Brookmeyer. Without his time, energy, and insightfulness, this dissertation would not be possible. I am thankful for his encouragement at every step of the way and for taking a genuine interest in my success and happiness. The research and career advice he provided at our weekly meetings will undoubtedly serve to guide me for many years to come. I would also like to thank Drs. Robert Weiss, Catherine Sugar, and Pamina Gorbach for serving as members of my committee and making the entire process an enjoyable experience. Their exceptional questions and recommendations have contributed greatly to the quality of this dissertation and will doubtless serve to shape the direction of the future work described herein.

I would like to thank my friends, family, and mentors for their unwavering support and encouragement. I would like to express my gratitude to my parents who instilled me with confidence and taught me, through their words and actions, the importance of hard work and dedication. I would like to thank my brothers, who continue to challenge and inspire me. Lastly, I would like to thank Elan, for being a constant source of love, support, and friendship and for the beautiful life we share together.

Vita

2005 Bachelor of Arts in Molecular and Cellular Biology
University of California, Berkeley
Berkeley, California, USA

2008 Master of Science in Statistics
San Diego State University
San Diego, California, USA

Publications

Aralis HJ, Gorbach PM, Brookmeyer R. Measuring concurrency using a joint multistate and point process model for retrospective sexual history data. *Statistics in Medicine* 2016; 35(24):4459-4473.

Lester P, Aralis H, Sinclair M, Kiff C, Lee KH, Mustillo S, Wadsworth SM. The impact of deployment on parental, family and child adjustment in military families. *Child Psychiatry & Human Development* 2016; 21:1-12.

Lester P, Liang LJ, Milburn N, Mogil C, Woodward K, Nash W, Aralis H, Sinclair M, Semaan A, Klosinski L, Beardslee W. Evaluation of a family-centered preventive intervention for military families: parent and child longitudinal outcomes. *Journal of the American Academy of Child & Adolescent Psychiatry* 2016; 55(1):14-24.

Ehrenstein OS, Aralis H, Flores ME, Ritz B. Fast food consumption in pregnancy and subsequent asthma symptoms in young children. *Pediatric Allergy and Immunology* 2015; 26(6):571-577.

Aralis HJ, Macera CA, Rauh MJ, MacGregor A. Traumatic brain injury and PTSD screening efforts evaluated using latent class analysis. *Rehabilitation Psychology* 2014; 59(1):68-78.

Ehrenstein OS, Aralis H, Cockburn M, Ritz B. In utero exposure to toxic air pollutants and risk of autsim. *Epidemiology* 2014; 25(6):851-858.

Macera CA, Aralis HJ, Highfill-McRoy R, Rauh MJ. Posttraumatic stress disorder after combat zone deployment among Navy and Marine Corps men and women. *Journal of Women's Health* 2014; 23(6):499-505.

Lester P, Lee KH, Aralis H, Sinclair M, Mustillo S, Wadsworth SM. Multiple deployments and very young children: how do parents and families influence child social-emotional adjustment? In *The Intergenerational Impact of War.* Military Family Research Institute at Purdue University, 2013; 55-75.

Macera CA, Aralis HJ, Rauh MJ, MacGregor AJ. Do sleep problems mediate the relationship between traumatic brain injury and development of mental health symptoms after deployment? *Sleep* 2013; 36(1):83-90.

Rauh MJ, Aralis HJ, Melcer T, Macera CM, Sessoms P, Bartlett J, Galarneau MR. Effect of traumatic brain injury among U.S. service members with amputation. *Journal of Rehabilitation Research and Development* 2013; 50(2):161-172.

Norris JN, Virre E, Aralis H, Sracic MK, Darren T, Gertsch JH. High altitude headache and acute mountain sickness at moderate elevations in a military population during battalion-level training exercises. *Military Medicine* 2012; 177(8):917-923.

CHAPTER 1

Introduction

1.1 Importance of Multistate Models in Public Health Research

Multistate models are widely used to analyze life history processes in which each individual is assumed to occupy one of a finite number of states at any given point in time. Recent examples include applications to models of disease progression for diseases such as cirrhosis [2], age-related macular degeneration [3], and bipolar disorder [4] and models of patient outcomes following therapeutic cardiovascular intervention [5] and elective general surgery [6]. Multistate models are deemed the preferred analytical approach in many public health applications due to their assumed coherence with underlying biological mechanisms or disease dynamics that posit categorical state occupancy as a function of time. Multistate models can be used to elucidate pathways of association across multiple states which are not directly apparent when modeling survival separately for each state or outcome. As noted by Eulenburg et al. in their recent study of breast cancer endpoints, multistate models allow for the simultaneous analysis of transitions between states and adjustments for intermediate events, making them preferable to traditional Cox proportional hazards models [7]. This point has been further emphasized by Jazic et al. in their recent study promoting the use of multistate illness-death models for cancer research in place of commonly used survival models which rely on composite endpoints such as disease-free progression [8].

Although progressive multistate models in which a trajectory moves along a time axis from an occupied state to a state never previously occupied have a wide range of applications in public health, instances in which non-progressive models are preferable also abound. A non-progressive model is one in which back transitions are possible, meaning that an individual can return to a state previously occupied. Such models are necessary when considering recurrent events leading to periods of hospitalization or infection, for instance, or disease states from which recovery is possible along with subsequent return to illness. In some instances, the formulation of a model that can enable identification of factors significantly impacting rates and probabilities of transition from an unhealthy state back to a healthier state is the primary research objective. For example, numerous studies have attempted to identify protective factors associated with increased rates of disease remission, recovery of functional ability, and return to intact cognition [9, 10, 11].

As an alternative to modeling progression through various states of disease and health, multistate models in which individuals are assumed to occupy states corresponding to potential disease risk factors may be useful in the field of public health research. Multistate models with back transitions are particularly important when modeling episodic behavioral patterns and prolonged periods of exposure that may put individuals at heightened disease risk. For example, multistate models may be used to model patterns of imprisonment, cigarette smoking, or engagement in risky sexual behaviors for a sample of individuals across an extended period of time. Parameter estimates from such models can be used to inform researchers about differences in risk behaviors or exposures between different populations or under different intervention conditions. Multistate model estimates can also be used as input to disease simulation models designed to evaluate the trajectory and viability of a disease within a population exhibiting certain risk patterns. Ultimately, information obtained from these research initiatives can assist public health experts in designing effective interventions to prevent disease and reduce morbidity at the population level.

In this dissertation, we will address three different methodological issues that were motivated by two applications in the field of public health. For each methodological issue, we describe the formulation of a multistate model with the capacity to answer specific questions of applied interest. We then address statistical challenges that arise in estimating the desired metrics for each proposed model given the available data and discuss the effectiveness of the adopted methodologies. In doing so, we hope to demonstrate utility and reduce existing barriers to the implementation of multistate models across a wide range of applications in the field of public health and in situations where imperfect data are available.

The two public health applications motivating the work presented in this dissertation are HIV/AIDS transmission and disease progression. The first application arises from the study of sexual partnership patterns impacting transmission of HIV within a given community. The applied objective is the accurate measurement of concurrency, defined as overlapping dates of sexual partnership, using retrospective sexual history reports obtained from a cross-sectional sample of individuals. The second applied objective is the modeling of disease progression for diseases in which back transitions from illness to health are possible and data are available in the form of intermittent observations of an individual's disease status. In this application there also exist concerns about misclassification arising from non-negligible false positive rates. Out of these two applications grew three methodological research goals. The first goal is described within Section 1.2, the second within Section 1.3, and the third within Section 1.4.

1.2 Challenges in Modeling of Concurrent Sexual Partnerships

Sexual partnership dynamics known to impact HIV transmission include the number and duration of partnerships, the frequency and type of sexual intercourse engaged in, and the length of time between partnerships [12, 13, 14]. However, the question of whether or not concurrency, defined as overlapping dates of sexual partnership, impacts HIV transmission independent of these other factors remains unanswered [15, 16].

In identifying concurrency, a frequently used operational definition involves classifying consecutive partnerships as having either a negative or a positive partnership gap [17]. For example, assume an individual is sampled from the population of interest and asked to report the number of days elapsed since the start and end of each previous partnership he or she engaged in during some elapsed time interval. For each set of consecutive partnerships, it is then possible to subtract the number of days corresponding to the beginning of the more recent partnership from the number of days corresponding to the end of the previous partnership. If the resulting difference is positive, we classify these partnerships as serially monogamous. If the difference is negative, we classify these partnerships as occurring concurrently. Thus, the question becomes whether or not concurrent partnership patterns result in increased rates of HIV transmission relative to serially monogamous patterns, when holding all other sexual partnership dynamics, such as number and duration of partnerships, fixed.

A number of frequently cited studies have used mathematical models to demonstrate that the risk of HIV transmission is theoretically greater when partnerships are concurrent rather than serially monogamous [18, 19]. However, strong empirical evidence to support the effect of concurrency on HIV transmission has been difficult to obtain resulting in an ongoing debate among experts in the field of HIV transmission research [20, 21, 22].

The statistical analysis of sexual partnership dynamics, and concurrency in particular, is complicated because data is frequently obtained from cross-sectional surveys in which participants record sexual histories over a specified elapsed time interval. Furthermore, sexual histories may include both partnerships that last over a prolonged period of time as well as isolated sexual encounters that occur on a single day. It is important to incorporate both ongoing and single day sexual partnerships when constructing a model for behavioral patterns that can be used to answer questions regarding risk of transmission.

1.3 Challenges in Disease Progression Modeling: Panel Data

In their simplest form, disease progression models consider each individual to be in a state of health, illness, or death at every point in time across a period of observation. In more complex disease progression models, multiple disease states are considered representing either various stages of disease or alternative disease outcomes. Multistate models for disease progression can be estimated using a straightforward maximum likelihood approach when knowledge of the sequence of states visited and the duration of time spent in each state is available for all individuals in a random sample drawn from the population of interest. In practice, however, complete longitudinal data on an individual's duration in and transition through a sequence of states corresponding to health and illness is seldom available. Instead, individuals are typically observed to be in a given state at a series of discrete points in time, often corresponding to routinely scheduled medical visits or survey administration waves. The data that arises from this intermittent observation process is often referred to as panel data [23, 24]. Panel data are typically obtained from conduct of a longitudinal study and may alternatively be referred to simply as longitudinal data.

When panel data are collected, the sequence of states an individual occupies across an observation time interval, often referred to as the individual's path, is typically unknown. Due to the intermittent nature of the observation process, it cannot be assumed that every state an individual occupies will be observed. If the structure of the assumed model is progressive, implying that an individual can occupy a state at most once, the possible sequence of states for a given individual will be limited in number. For instance, the commonly used three-state progressive illness-death model described by Joly et al. allows for at most two different potential paths existing between consecutive observations [25]. However, as noted previously, progressive models are not appropriate in all health research applications because they do not allow for transitions to the state previously occupied, referred to as back transitions. Back transitions are an important phenomenon in many real world multistate processes. Specific examples include recurrent infections or hospitalizations and progression of diseases such as relapse-remitting multiple sclerosis or diabetic retinopathy [26, 27, 28, 29]. For such models, enumeration of all potential paths between consecutive observations is frequently not possible. In fact, when assuming a continuous-time multistate process, there is no limit to the number of back transitions that can occur between any two discrete points in time and thus no limit to the number of potential paths that would need to be considered when estimating the parameters of a multistate model.

1.4 Challenges in Disease Progression Modeling: Misclassification

Many commonly used diagnostic tools have non-negligible false positive rates resulting in the misclassification of healthy individuals into a diseased state. This issue is especially relevant when considering intermittently-observed disease state data. Methods for the estimation of multistate models with back transitions are highly sensitive to measurement error. Such methods rely heavily on the accuracy of relatively minimal data obtained at discrete points in time to effectively impute unobserved state transitions and durations required for parameter estimation. In such instances, misclassification of a healthy individual into a diseased state can substantially alter our perception of the disease trajectory for that individual. For instance, an annually-observed individual who remains in a state of health over the course of ten years but is incorrectly identified as having a disease at year five would contribute misleading information to the likelihood function used in model estimation. The individual would appear to have experienced at least one back transition potentially resulting in artificial inflation of the rate of back transition and reduction of the expected sojourn time for the healthy state.

In an ideal setting, we would correct for the bias arising from misclassification by either directly identifying the observations that were misclassified or using an ancillary estimate of the probability of misclassification given occupancy of a known state. Both approaches typically rely on the availability of a gold standard diagnostic tool. Unfortunately, for many diseases, a gold standard either doesn't exist or is not feasible for use due to barriers such as expense or patient burden [30]. Thus, the challenge in disease progression modeling with misclassification is to accomplish unbiased estimation of the multistate model parameters associated with transition rates and probabilities while simultaneously estimating the true underlying misclassification rate without requiring information from a gold standard.

1.5 Dissertation Roadmap

In this dissertation, we begin by describing the broad overarching continuous-time multistate process framework including the necessary background, definitions, and notation in addition to some practical assumptions that enable modeling of such processes. In the subsequent chapter, we address the first research objective by presenting a joint multistate and point process model for retrospective sexual history data and describe how model parameters can be used in the estimation of two critical measures of concurrency. We apply this joint modeling approach using epidemiological data collected from a sample of men having sex with men and seeking HIV testing at a Los Angeles clinic.

To address the second research objective, we describe the statistical challenges in estimating a semi-Markov model for disease progression with back transitions when data are available in the form of intermittent observations. Following this description, we propose a stochastic expectation-maximization (SEM) approach for estimation of such models and compare the proposed approach to an alternative, naive approach using a simulation study. In the following chapter, we apply the proposed SEM approach to dementia onset modeling using the Nun Study data collected from a sample of elderly members of the School Sisters of Notre Dame who received routine cognitive assessments as part of a longitudinal study on aging and Alzheimer's disease. We then consider an extension to the proposed SEM approach for disease progression modeling that allows for misclassification of a healthy individual in the diseased state and demonstrate the performance of the proposed extension using a simulation study. The SEM method with misclassification is then applied to the Nun Study to estimate the probability of an elderly participant with intact cognition being incorrectly classified as having impaired cognition when completing a routine cognitive assessment. We conclude by reviewing the accomplishments of the research presented herein and discussing promising areas for future research.

CHAPTER 2

Background and Definitions

2.1 Defining a Continuous-Time Multistate Process

We begin by describing a general continuous-time multistate process and then sequentially introduce a set of assumptions that enable the practical application of multistate models to real world data. Multistate models are usually constructed by collecting an independent sample of units (individuals) and assuming that each unit (individual) can be modeled as its own continuous-time multistate process. A multistate process is one which can occupy any one of a finite number of states at realizations occurring across time. At time t, let process $Y(t)$ take on values in state space S, $S = \{0, 1, \ldots, s-1\}$, for a process with s states. The distribution of a multistate process can be fully defined by either H, a $s \times s$ transition matrix with entries $H_{kl}(t_1, t_2)$, or the set of transition intensities, $\lambda_{kl}(t, \mathscr{F}(t))$, where for all $k, l \in \{0, 1, \ldots, s-1\}$,

$$\begin{aligned} &H_{kl}(t_1, t_2) = P\big(Y(t_2) = l \big| Y(t_1) = k; \mathscr{F}(t_1)\big) \qquad \text{for } t_1 < t_2, \text{ and} \\ &\lambda_{kl}(t, \mathscr{F}(t)) = \lim_{\Delta t \to 0} \frac{H_{kl}(t, t+\Delta t)}{\Delta t}. \end{aligned}$$

We define $\mathscr{F}(t)$ as the filtration representing all information about the process up until time t. The filtration, $\mathscr{F}(t)$ can thus be summarized by $\big\{N(t), T_1, \ldots, T_{N(t)}, Y(0), Y(T_1), \ldots, Y(T_{N(t)})\big\}$ where $N(t)$ is the count of transitions occurring up until time t and T_m are the observed times of transition for $m = 1, \ldots, N(t)$. Thus, $H_{kl}(t_1, t_2)$ denotes the probability of being in state l at time t_2 given that the process was in state k at the earlier time t_1 and the history of the process up until time t_1. Transition intensities, $\lambda_{kl}(t, \mathscr{F}(t))$ denote the hazard function associated with the rate of escape or exit from state k to state l for $k \neq l$, while $\lambda_{kk}(t, \mathscr{F}(t))$ is the hazard function associated with the duration of time the process spends in state k prior to a transition occurring, also known as the sojourn time. When data are available from a sample of units, these transition intensities are typically treated as population parameters that may either be assumed constant across units or expressed as a function of explanatory variables.

2.2 Markov and Semi-Markov Properties

To make the multistate framework useful for modeling, simplifying assumptions usually need to be made. The process described previously is general enough to accommodate non-Markovian and time inhomogeneous process. In practice, Markovian assumptions are often made such that the multistate process adheres to either the Markov or the semi-Markov property. The Markov property stipulates that the future of the process at a given point in time depends only on the current state of the process and therefore the history of the process can be ignored. Under the Markov assumption, for all $k, l \in \{0, 1, \ldots, s-1\}$,

$$\begin{aligned} &H_{kl}(t_1, t_2) = P\big(Y(t_2) = l \big| Y(t_1) = k\big) \qquad \text{for } t_1 < t_2, \text{ and} \\ &\lambda_{kl}(t, \mathscr{F}(t)) = \lambda_{kl}(t). \end{aligned}$$

Another assumption that is often made is that the process is time homogeneous meaning that transition intensities are stationary with respect to time such that for all $k, l \in \{0, 1, \ldots, s-1\}$, $\lambda_{kl}(t) = \lambda_{kl}$. The homogenous Markov model therefore implies exponentially distributed sojourn times with constant hazards. A straightforward expression for the probability of transitioning from state k to state l conditional upon being in state k at a given arbitrary time, t, can be expressed as

$$\begin{aligned} P_{kl} &= P(Y(T_{m+1}) = l | Y(T_m) = k) \quad \forall \;\; m = 1, \ldots \\ &= \frac{\lambda_{kl}}{\sum_{m \in S, m \neq k} \lambda_{km}}. \end{aligned}$$

Thus, for a homogeneous Markov process the probability of transitioning from one state to another is proportional to the associated transition intensity.

The semi-Markov property is less restrictive than the Markov property and stipulates that the future of the process at a given point in time may depend on both the current state of the process and the time elapsed in the current state. Thus, the history of the process prior to entrance into the current state can be ignored. Under the semi-Markov assumption, for all $k, l \in \{0, 1, \ldots, s-1\}$,

$$H_{kl}(t_1, t_2) = P\big(Y(t_2) = l \big| Y(t_1) = k, t_1 - T_{N(t_1)}\big) \qquad \text{for } t_1 < t_2, \text{ and}$$
$$\lambda_{kl}(t, \mathscr{F}(t)) = \lambda_{kl}(t, t - T_{N(t)}),$$

where $t - T_{N(t)}$ represents the time elapsed at time t since entry into the current state. Once again, it is possible to make the assumption of time homogeneity implying $\lambda_{kl}(t, t - T_{N(t)}) = \lambda_{kl}(t - T_{N(t)})$. The time homogenous semi-Markov process allows hazards of transition to vary as a function of time elapsed in the current state which gives rise to non-exponentially distributed durations. We can demonstrate the impact of various combinations of assumptions on our expression for the transition intensities using the following diagram.

	Time Inhomogeneous	Time Homogeneous
Non-Markov	$\lambda_{kl}(t, \mathscr{F}(t))$	$\lambda_{kl}(\mathscr{F}(t))$
Semi-Markov	$\lambda_{kl}(t, t - T_{N(t)})$	$\lambda_{kl}(t - T_{N(t)})$
Markov	$\lambda_{kl}(t)$	λ_{kl}

The semi-Markov process cannot be considered a continuous-time Markov process because it does not satisfy the Markov property continuously across time. It does, however, satisfy the criteria for a discrete-time Markov process when only the instants of transition, T_m, are considered. A discrete-time Markov process, sometimes referred to as a Markov chain, differs from a continuous-time Markov process in that the process consists of the sequence of states only and does not include information about the duration of time spent in each state. The Markovian property that the next state visited depends only on the current state occupied is still retained. The discrete-time process associated with a continuous-time semi-Markov process is often referred to as the embedded Markov chain. As we will see, in application it is often preferable to express the transition intensities for a semi-Markov process, $\lambda_{kl}(t - T_{N(t)})$, as the product of two components: the first representing the transition probabilities from the embedded discrete-time Markov chain, P_{kl}, and the second representing the state-specific sojourn time distributions which are functions of $t - T_{N(t)}$.

2.3 Modeling Framework for Markov and Semi-Markov Processes

In developing a framework for modeling Markov and semi-Markov processes we focus on expressing the transition intensities in such a way as to explain the observed heterogeneity across the sample of observed processes. While retaining the concept that the entire history of a Markov or semi-Markov process up until time t can be summarized by information observed at time t, we can modify our assumptions to allow incorporation of explanatory variables into the modeling of transition intensities. We define $\boldsymbol{X}(t)$ as a multivariate explanatory process whose value is known right before time t. Sometimes referred to as a partial Markov property [4], the condition allowing for incorporation of $\boldsymbol{X}(t)$ can be described as

$$\begin{aligned} H_{kl}(t_1, t_2) &= P\big(Y(t_2) = k \big| Y(t_1) = l; \mathscr{F}(t_1)\big) \\ &= P\big(Y(t_2) = k \big| Y(t_1) = l, \boldsymbol{X}(t_1)\big) \quad \text{for } t_1 < t_2, \end{aligned}$$

where $\mathscr{F}(t_1)$ is now the filtration of both Y and $\boldsymbol{X}$ up to time t_1 [31]. Analogously, the condition allowing for incorporation of $\boldsymbol{X}(t)$ assuming a partial semi-Markov property can be described as

$$\begin{aligned} H_{kl}(t_1, t_2) &= P\big(Y(t_2) = k \big| Y(t_1) = l; \mathscr{F}(t_1)\big) \\ &= P\big(Y(t_2) = k \big| Y(t_1) = l, t_1 - T_{N(t_1-)}, \boldsymbol{X}(t_1)\big) \quad \text{for } t_1 < t_2. \end{aligned}$$

By definition, $\boldsymbol{X}(t)$ can include durations, time-fixed explanatory variables, and time-dependent variables allowing great flexibility to the set of potential partial Markov or partial semi-Markov models. For simplicity, we will omit the t notation when referring to $\boldsymbol{X}(t)$ throughout the remainder of this section as it is only relevant when time-varying explanatory variables are included.

In constructing a partial Markov model, λ_{kl} may now depend on the process $\boldsymbol{X}$ for all $k, l \in \{0, 1, \ldots, s-1\}$. A common approach to incorporating explanatory variables when constructing a Markov model involves parametrizing the transition intensity functions using the log linear expression

$$\lambda_{kl}(\boldsymbol{X}) = \exp(\boldsymbol{\beta}_{kl}^T \boldsymbol{X}),$$

where $\boldsymbol{\beta}_{kl}$ denotes the vector of parameters to be estimated for all $k, l \in \{0, 1, \ldots, s-1\}$. This modeling framework allows us to interpret the regression coefficients, $\boldsymbol{\beta}_{kl}$, in terms of hazard ratios.

To construct a partial semi-Markov model, λ_{kl} must be expressed as a function of both $\boldsymbol{X}$ and $t - T_{N(t)}$. As alluded to previously, the transition intensity functions for a semi-Markov model are preferably expressed as the product of the embedded discrete-time Markov chain transition probabilities, P_{kl}, and the state-specific sojourn time distribution functions, which we denote f_{kl}. For all $k, l \in \{0, 1, \ldots, s-1\}$,

$$\lambda_{kl}(t - T_{N(t)}, \boldsymbol{X}) = P_{kl}(\boldsymbol{X}) f_{kl}(t - T_{N(t)}, \boldsymbol{X}).$$

In constructing a semi-Markov model, explanatory variables can be incorporated into either or both of the two components. A typical parameterization for the embedded Markov chain transition probabilities P_{kl} will involve a logit transformation such that

$$\log\left(\frac{P_{kl}(\boldsymbol{X})}{1 - P_{kl}(\boldsymbol{X})}\right) = \exp(\boldsymbol{\beta}_{kl}^T \boldsymbol{X}),$$

subject to the constraint that $\sum_{l=0}^{s-1} P_{kl} = 1$ for all $k \in \{0, 1, \ldots, s-1\}$. This modeling framework is analogous to logistic regression and allows us to interpret regression coefficients, $\boldsymbol{\beta}_{kl}$, in terms of odds ratios. We usually assume that the sojourn time distributions for a semi-Markov model, f_{kl}, belong to a parametric family such as the Weibull or Gamma. To incorporate explanatory variables into the sojourn time distribution functions, it is typical to express the hazard rate for f_{kl} using a Cox proportional regression model [32]. The hazard rate is defined by

$$\alpha_{kl}(t - T_{N(t)}, \boldsymbol{X}) = \alpha_{kl0}(t - T_{N(t)}) \exp(\boldsymbol{\beta}_{kl}^T \boldsymbol{X}),$$

where $\alpha_{kl0}(t - T_{N(t)})$ denotes the baseline hazard which equals the hazard associated with the chosen parametric family. For example, for the Weibull distribution,

$$\alpha_{kl0}(t - T_{N(t)}) = \left(\frac{\nu_{kl}}{\sigma_{kl}}\right)\left(\frac{t - T_{N(t)}}{\sigma_{kl}}\right)^{(\nu_{kl}-1)}$$

for all $k, l \in \{0, 1, \ldots, s-1\}$ and for scale parameters $\sigma_{kl} > 0$ and shape parameters $\nu_{kl} > 0$. This modeling framework allows us to interpret the regression coefficients, $\boldsymbol{\beta}_{kl}$, in terms of hazard ratios. For example, incorporating covariates at the level of f_{kl} for the Weibull distribution gives us

$$\begin{aligned} f_{kl}(t - T_{N(t)}, \boldsymbol{X}) &= \alpha_{kl}(t - T_{N(t)}, \boldsymbol{X}) S_{kl}(t - T_{N(t)}, \boldsymbol{X}) \\ &= \left(\frac{\nu_{kl}}{\sigma_{kl}}\right) \left(\frac{t - T_{N(t)}}{\sigma_{kl}}\right)^{(\nu_{kl}-1)} \exp\left(\boldsymbol{\beta}_{kl}^T \boldsymbol{X}\right) \exp\left[- \left(\frac{t - T_{N(t)}}{\sigma_{kl}}\right)\right]^{\exp(\boldsymbol{\beta}_{kl}^T \boldsymbol{X})}, \end{aligned}$$

where $S_{kl}(t - T_{N(t)}, \boldsymbol{X})$ is the Weibull survival function associated with hazard rate $\alpha_{kl}(t - T_{N(t)}, \boldsymbol{X})$. The expression for f_{kl} can be directly incorporated into the likelihood function to facilitate parameter estimation.

Although we have presented process elements such as transition intensities, transition probabilities, and state-specific sojourn time distribution functions using specific notation within this chapter, the notation associated with these terms will vary in the chapters to come as is necessary to accommodate the different modeling frameworks and applications. The terms used to refer to each element in subsequent chapters will inform their relation to the materials presented in this chapter.

CHAPTER 3

A Joint Model for Concurrent Sexual Partnerships

3.1 Challenges in Measuring Concurrency

Many of the challenges in establishing empirical evidence supporting the impact of concurrent, or overlapping, sexual partnerships on HIV transmission can be traced back to difficulties in measurement. To assess the effect of concurrent partnership patterns on the trajectory of the HIV epidemic in a population, it is necessary to estimate both the extent and the magnitude of concurrency. Specifically, interest lies in estimating the point prevalence of concurrency, referred to more generally as the *concurrent partnership distribution*, and the mean duration of concurrency, referred to here as the *mean concurrent partnership sojourn time* during which a person has k ongoing partnerships. We assume both these metrics can be estimated using sexual history data obtained by independently sampling individuals from a population existing in a stationary state with respect to its partnership patterns. Thus, the concurrent partnership distribution and the mean concurrent partnership sojourn times are estimated for a population at steady state and are not expressed as a function of time.

- *Concurrent partnership distribution* π_k is the probability that an individual member of a population is engaged in k ongoing partnerships at any given point in calendar time for $k \in \{0, 1, 2, \ldots\}$.

- *Mean concurrent partnership sojourn time* ρ_k for an individual engaged in k ongoing partnerships is the the mean duration of time the individual will remain in k ongoing partnerships before experiencing the next partnership formation or dissolution for $k \in \{0, 1, 2, \ldots\}$.

Estimation of these two population concurrency metrics enables researchers to draw inferences about specific populations from which data were collected and to ultimately compare patterns of concurrency across different populations or subpopulations. Further, empirical estimates of these population concurrency metrics could be used as input when constructing infectious disease mathematical models, such as agent-based and social network models, which would allow researchers to examine the viability and trajectory of the HIV epidemic over time and under variable conditions [33, 34, 35, 36].

An optimal study design for estimating these population concurrency metrics would involve recruitment of a cohort of individuals prior to sexual debut followed by ongoing collection of partnership information from each participant throughout the duration of his or her life. Unfortunately, such designs are prohibitively expensive and typically infeasible due to implementation obstacles. Instead, partnership data is typically collected retrospectively in the form of sexual history information obtained using an approach known as the calendar method [37]. Following this approach, researchers administer a cross-sectional survey to a sample of individuals from the target population asking respondents to identify each sexual partnership, either ongoing or concluded, that occurred in part or in full during the previous year or other pre-specified elapsed interval of time. For each identified partnership, respondents are then asked to provide the first and last dates of sexual intercourse.

A drawback of the calendar method is that careful consideration needs to be taken when attempting to appropriately analyze data obtained using this technique. Traditionally, attempts to analyze retrospective sexual history data have used the partnership as the unit of observation and have thus ignored heterogeneity across individuals and time. For example, the mean partnership duration is usually calculated by averaging across partnership durations reported by all individuals across all time points, assuming partnerships to be independent and identically distributed [38]. In some instances, these partnership-level analyses have also inadequately addressed right censoring and length-biased sampling [3]. Another shortcoming of current analytical approaches is the tendency to use only a snapshot of the available information. For example, to obtain an estimate of the concurrent partnership distribution, a specific time point is selected, such as one month prior to the survey date, and the observed distribution at that time is taken as the estimated distribution thereby discarding a large portion of the available data [39, 40]. In 2010, a UNAIDS working group developed guidelines for measuring concurrency and recommended that the point prevalence at six months prior to the interview be used as an indicator of concurrency within a population [41]. Following the dissemination of these guidelines, numerous articles were published questioning the validity of the proposed indicator citing issues of recall bias and demonstrating the variability in point prevalence estimates across differing points in time [40, 42].

Another concern rarely addressed when analyzing retrospective sexual history data is the handling of individual partnerships reported as having the same first and last dates of sexual intercourse. These partnerships are usually assumed to represent one-time sexual encounters and will be referred to as *one-offs*. Here we distinguish one-offs from other partnerships which we will refer to throughout this paper as *ongoing partnerships*. The high rate of one-offs reported among many of the populations targeted for prevention and treatment efforts necessitates the consideration of these events in the statistical analysis stage. Regardless of the per-one-off transmission probabilities, the cumulative effect of relatively high rates of one-offs on HIV transmission within a community may be substantial. Another advantage of explicitly modeling these one-off events is to account for the potential impact of one-off events engaged in by an individual on the likelihood of a subsequent partnership formation or dissolution event.

Based on the challenges described above, we aim to develop a modeling framework that meets four criteria. The proposed model should:

1. Treat individuals as the independent units of observation rather than partnerships which may exhibit dependence when engaged in by the same individual at the same or different points in time.

2. Allow estimation of population metrics of interest for measuring concurrency: the *concurrent partnership distribution* and the *mean concurrent partnership sojourn time* for an individual engaged in k ongoing partnerships.

3. Be flexible enough to incorporate explanatory variables to identify and characterize factors affecting concurrency.

4. Account for one-offs and allow the occurrence of one-offs to potentially impact the subsequent formation and dissolution of ongoing partnerships.

3.2 A Joint Model for Retrospective Sexual History Data

To address the stated modeling objectives, a joint multistate and point process model is proposed. As described previously, a multistate model is a model for a continuous-time stochastic process which may, at any time, occupy one of a number of discrete states [43, 44]. Typically, multistate models are fit to longitudinal observations of a categorical variable. For sexual history data, each individual's continuation in and transition through differing states, where state is defined as the number of ongoing partnerships an individual is engaged in at a given point in time, can be modeled using a multistate modeling approach. In this manner, each individual's partnership patterns over time are treated as a single stochastic process. Figure 3.1(a) depicts data for an individual who reported the first and last dates of sexual intercourse for three partnerships occurring within the past year. Figure 3.1(b) demonstrates the way in which the reported partnership information can be translated into process data appropriate for use in fitting a multistate model. The depicted individual begins the year interval in a state of zero ongoing partnerships. He then experiences three partnership formation events followed by two dissolution events and ends the year in a state of one ongoing partnership.

The multistate component of the joint model addresses one of our modeling objectives by treating each individual's sexual history as its own stochastic process. Partnerships engaged in by the same individual at the same and different points in time are inherently linked by the modeling of transition intensities associated with partnership formation and dissolution. However, the point process component of the joint model is necessary to accommodate one-offs. By proposing a joint model, the state occupied by the multistate process for an individual at a given time can influence the rate of occurrence of one-offs. Additionally,

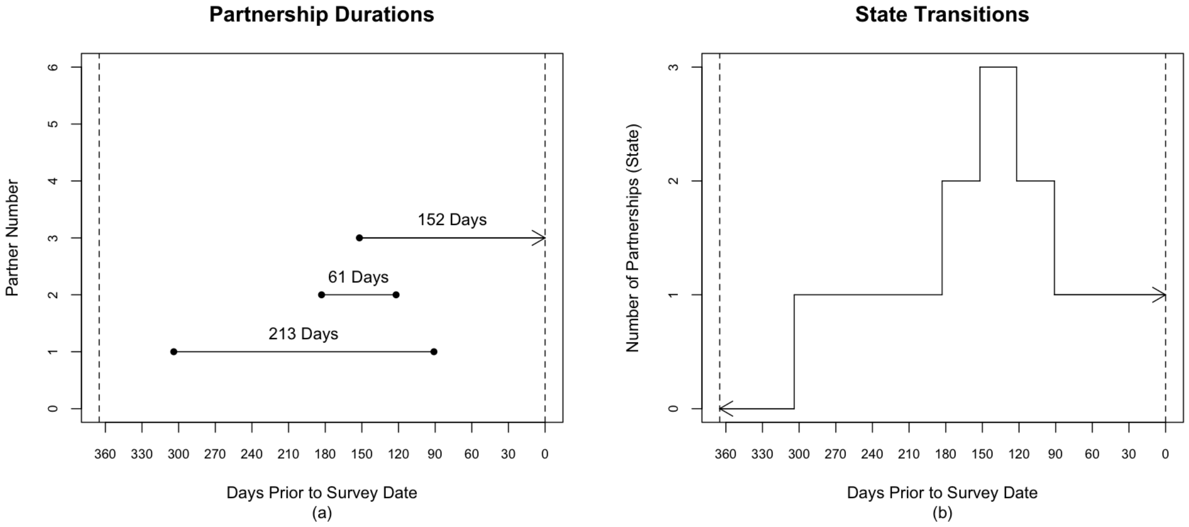


Figure 3.1: Translation of retrospective sexual history data provided by a single respondent into multistate modeling information

the joint nature of the model allows the occurrence of one-offs to affect the subsequent intensity of transition from one state to another. Joint modeling of a multistate process and a discrete event process has been recently demonstrated using medical record data with random informative observation times [45].

Let $Y(t)$ denote the number of ongoing partnerships an individual is engaged in at calendar time t such that $Y(t)$ takes values in $\{0, 1, 2, \ldots\}$ for all t where t corresponds to external or calendar time. We let $Y(t)$ represent the count of ongoing partnerships, which excludes the occurrence of one-offs that are alternatively modeled by the count process component of the joint model. Jumps in $Y(t)$ thus correspond to partnership formation or dissolution events. Assume multiple partnership formation or dissolution events cannot occur at the exact same point in time such that $Y(t)$ can only jump to adjacent states resulting in a birth-death-type process. If we assume $Y(t)$ is a time homogeneous continuous-time Markov multistate process, $Y(t)$ can be fully characterized by specification of either the transition probabilities from state k to state l

$$p_{kl}(\Delta t) = P\left(Y(t + \Delta t) = l | Y(t) = k\right)$$

for all $t \geq 0$ and $k, l \in \{0, 1, 2, \ldots\}$, or by the transition intensities,

$$\alpha_{kl} = \lim_{\Delta t \to 0} \frac{p_{kl}(\Delta t)}{\Delta t}$$

for $k, l \in \{0, 1, 2, \ldots\}$, which represent the instantaneous probability of transition to state l given occupation of state k. Under the Markov assumption, these transition intensities are assumed constant with respect to time yielding exponentially distributed sojourn times. In order for the occurrence of one-offs to influence the subsequent intensity of transition we must relax this assumption by allowing transition intensities to vary within occupancy of a state. We adopt a phase-type Markov model with intensities that fluctuate in response to each one-off event that occurs during occupancy of a state [46]. To implement this, we choose to parametrically model transition intensities by

$$\alpha_{kl}(t) = \beta_{kl0}\text{exp}(\beta_{kl1}N(t) + \boldsymbol{\beta}_{kl2}^T\boldsymbol{X}(t)), \tag{3.1}$$

where $N(t)$ represents the count of one-off events at calendar time t having occurred since the last transition in $Y(t)$ and $\boldsymbol{X}(t)$ denotes a vector of explanatory variables. For an individual who enters a state of k ongoing partnerships, $N(t)$ counts the occurrence of one-offs since entry into the current state. Each time an ongoing partnership is formed or dissolved an individual transitions to a new state and the counting process for one-offs, $N(t)$, starts over at zero. To make the definition of $N(t)$ precise, let $\nu(t)$ represent the cumulative count of one-offs having occurred at time t such that $\nu(t)$ represents a true counting process. Let $N(t) = \nu(t) - \nu(s(t))$ where $s(t) = \max_{0 \leq x < t}(\{x : Y(x) \neq Y(t)\}, 0)$ such that $s(t)$ represents the calendar time at which the last partnership formation or dissolution occurred prior to time t. $N(t)$ takes values in $\{0, 1, 2, \ldots\}$ for all t. Assume no two one-offs can occur at the exact same time such that $N(t)$ is a counting process within each state occupied. In general, the α_{kl} transition intensities may depend on the history of the process, $\mathcal{H}$, which includes the trajectories associated with $N(t)$ and $\boldsymbol{X}(t)$ over time ranging from 0 to t. Here we consider the special case in which the history of the process can be ignored and $\alpha_{kl}(t|\mathcal{H})$ can be reduced to $\alpha_{kl}(t)$ and expressed as a function of $\boldsymbol{X}(t)$ and $N(t)$ observed at time t. Thus, the intensity of transition at time t may depend log-linearly on the number of one-offs having occurred, baseline characteristics of the individual, and time-dependent characteristics of the

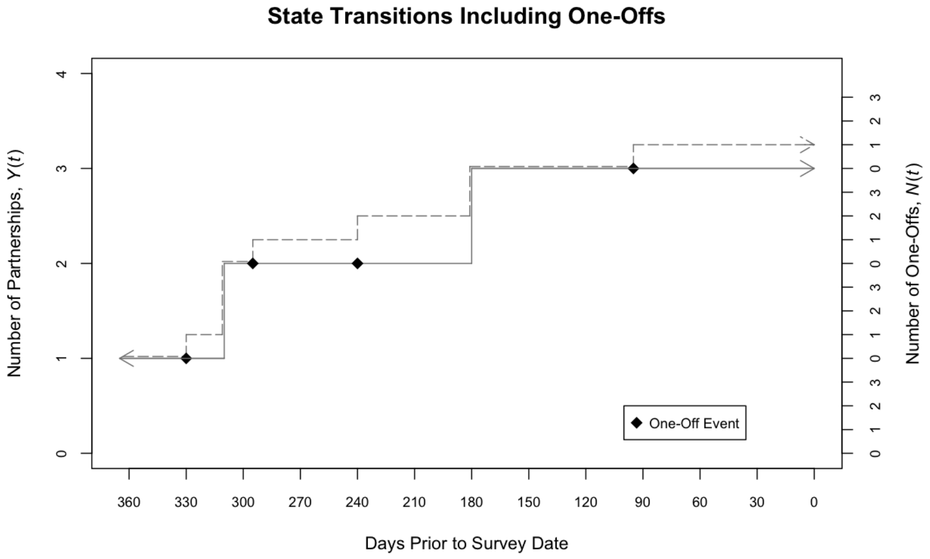


Figure 3.2: Sample sexual partnership pattern for a single respondent demonstrating both components of the joint model: the multistate process for the number of ongoing partnerships, and the point process for the number of one-offs

individual or his ongoing partners. Time-dependent variables are restricted, however, in that they are only allowed to vary along with the phase-type intensities which vary only at the instant a partnership is formed or dissolved or a one-off occurs. For example, $\boldsymbol{X}(t)$ could contain an indicator for engagement in a main partnership or occurrence of a one-off with a commercial sex worker but could not contain information on an individual's CD4 count as measured at arbitrary time points.

We choose to model $N(t)$ as a Markov-modulated Poisson process which allows variation in the rates of one-offs over time according to the number of ongoing partnerships an individual is engaged in. Markov-modulated Poisson processes are doubly stochastic in that the Poisson process rate varies according to a continuous-time Markov chain [47]. In our proposed model, the continuous-time Markov chain regulating the rates of one-offs corresponds to the multistate process $Y(t)$ representing the number of ongoing partnerships an individual is engaged in at time t. The interaction of both components of the proposed joint model are visually depicted in Figure 3.2. At each day prior to the survey date corresponding to time t, an individual can be said to be in a state of $Y(t)$ ongoing partnerships and to have experienced $N(t)$ one-offs which can be read from the left and right vertical axes of Figure 3.2, respectively. The dashed line which increments along with each partnership or one-off event depicts the times at which the transition intensities for partnership formation or dissolution may vary. A Markov-modulated Poisson process can be fully characterized through specification of the intensity function $\lambda(t)$ which represents the infinitesimal rate at which events are expected to occur around time t. Thus, in modeling $N(t)$ it suffices to model $\lambda(t)$. We parametrically model intensity functions for individuals in a state of k ongoing partnerships by

$$\lambda_k(t) = \gamma_{k0}\text{exp}(\boldsymbol{\gamma}_{k1}^T \boldsymbol{X}(t)), \tag{3.2}$$

where $\boldsymbol{X}(t)$ is a vector of potentially time-dependent explanatory variables with the same restrictions as described previously. $\boldsymbol{X}(t)$ does not need to consist of the same explanatory variables across the two components of the model and careful consideration of the joint nature of the model should be taken prior to selecting covariates for inclusion in both components of the model. In choosing to consider the two components of the joint model together as a two-dimensional vector, $(Y(t), N(t))$, the overarching modeling framework can alternatively be described as a bivariate continuous-time Markov process in which the Markov property holds for states defined through specification of both $Y(t)$ and $N(t)$.

3.3 Estimation of Joint Model Parameters

We fit the joint model described in section 3.2 and specifically given by equations (3.1) and (3.2) using maximum likelihood estimation. Let $\boldsymbol{\beta}_{kl} = \{\beta_{kl0}, \beta_{kl1}, \beta_{kl2}\}$ denote the vector of regression parameters expressed in equation (3.1). In theory, transition intensity regression parameters $\boldsymbol{\beta}_{kl}$ can be estimated for each $k, l \in \{0, 1, 2, \ldots\}$. In practice, we choose to specify a limited number of unique transition intensities. For modeling of sexual history data, we will estimate parameters $\boldsymbol{\beta}_{kl}$ associated with transition from a state of k ongoing partnerships to a state of l ongoing partnerships for three distinct types of transitions:

$$\begin{array}{ll} \text{Formation of a monogamous partnership} & (k=0, l=1) \\ \text{Formation of a concurrent partnership} & (k>0, l=k+1) \\ \text{Dissolution of any ongoing partnership} & (k>0, l=k-1). \end{array}$$

To clarify use of the term monogamous, here we define a *monogamous partnership* as any partnership that is the sole ongoing partnership that an individual has reported being engaged in at a given point in time. Thus, engagement in a monogamous partnership does not preclude the occurrence of one-offs or the formation of additional concurrent partnerships at a future point in time. To simplify notation, let $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$, and $\boldsymbol{\beta}_2$ denote the parameters to be estimated for the intensity of formation of a monogamous partnership, formation of a concurrent partnership, and dissolution of a partnership, respectively. Additionally, we choose to estimate $\boldsymbol{\gamma}_k = \{\gamma_{k0}, \gamma_{k1}\}$ from equation (3.2) separately for states of no ongoing partnership ($k = 0$), one ongoing partnership ($k = 1$) and concurrent partnerships ($k \geq 2$). For simplicity, we will denote these parameters as $\boldsymbol{\gamma}_0$, $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$, respectively. Therefore, the full likelihood will be maximized to obtain parameter estimates $\hat{\boldsymbol{\theta}} = \{\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\gamma}}_0, \hat{\boldsymbol{\gamma}}_1, \hat{\boldsymbol{\gamma}}_2\}$.

In constructing the likelihood we must calculate the probability of survival across intervals of time during which the transition probabilities remain constant. The term *event* will be used to signify any one of the following: a partnership formation, a partnership dissolution, or the occurrence of a one-off. We assume that multiple events cannot occur at the same instant in time. Due to the Markov modeling approach, inter-event times are exponentially distributed. For an individual in a state of k ongoing partnerships who experiences an event at time t_1, the probability of that individual remaining in state k until time t_2 without experiencing another event is

$$S_k(t_1, t_2) = \begin{cases} \exp(-\{(\alpha_{01}(t_1) + \lambda_0(t_1))(t_2 - t_1)\}) & \text{if } k = 0 \\ \exp(-\{(\alpha_{k,k+1}(t_1) + \alpha_{k,k-1}(t_1) + \lambda_k(t_1))(t_2 - t_1)\}) & \text{if } k \neq 0, \end{cases}$$

where α_{01}, $\alpha_{k,k+1}$, and $\alpha_{k,k-1}$ denote the transition intensities associated with the formation of a monogamous partnership, formation of a concurrent partnership, and dissolution of any ongoing partnership, respectively. λ_k denotes the rate of one-offs for an individual in

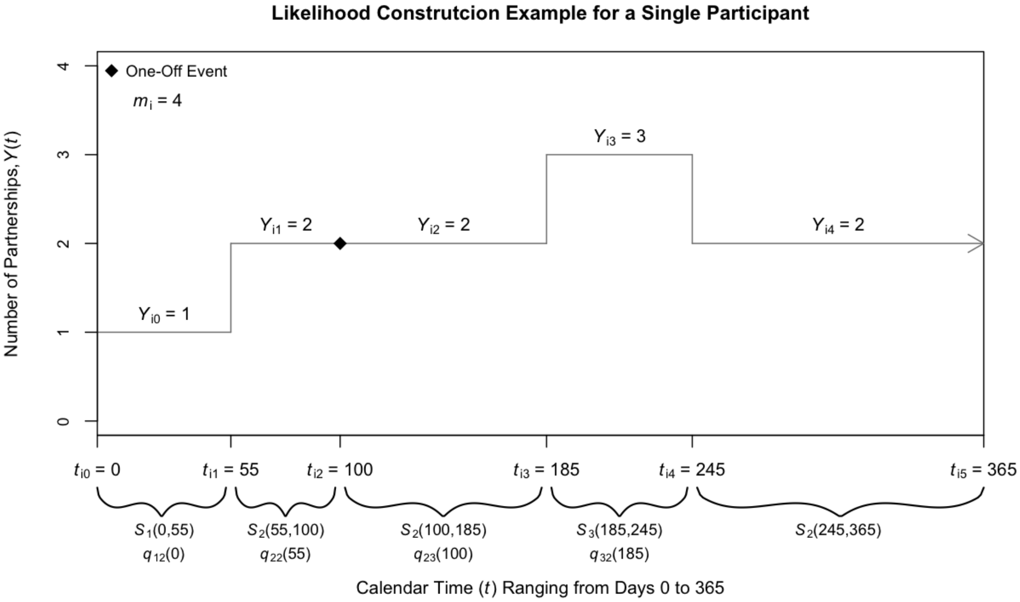


Figure 3.3: Demonstration of approach taken to calculate the likelihood contribution for an individual i with $m_i = 4$ transitions occurring during the year interval. Expressions for all S and q shown beneath the plot are multiplied together to yield the total contribution for individual i.

a state of k ongoing partnerships. To construct the likelihood, we must also calculate the instantaneous probability of an event occurring at time t. Given an individual in a state of k ongoing partnerships experienced an event at time t_1 and remained in state k until time t_2, the probability of a specific event occurring at time $t_2 > t_1$ is assumed constant and equal to

$$q_{kl}(t_1) = \begin{cases} \alpha_{01}(t_1) & \text{if } k = 0, l = 1 \\ \alpha_{k,k+1}(t_1) & \text{if } k > 0, l = k + 1 \\ \alpha_{k,k-1}(t_1) & \text{if } k > 0, l = k - 1 \\ \lambda_k(t_1) & \text{if } k \geq 0, l = k, \end{cases}$$

where $q_{kk}(t_1)$ indicates no change in the number of ongoing partnerships and is used to denote the occurrence of a one-off. Let $i = 1, \ldots, n$ index each respondent in an independent sample of size n. For each individual i, let m_i indicate the total number of events experienced, either ongoing partnership events or one-off events, over the course of the year interval. Let $T_i = \{t_{i0}, t_{i1}, \ldots, t_{im_i}, t_{i(m_i+1)}\}$ be the set of event times for individual i such that t_{i0} indicates the time at which the year interval begins, t_{i1} the time when the first event occurs, t_{im_i} the time when the last event occurs and $t_{i(m_i+1)}$ the time at which the year interval ends. Event times are ordered such that $t_{i0} < t_{i1} < \ldots < t_{i(m_i+1)}$. Similarly, let $Y_i = \{y_{i0}, y_{i1}, \ldots, y_{im_i}\}$ be the sequence of states for individual i such that y_{i0} and y_{im_i} indicate the numbers of ongoing partnerships individual i is engaged in at the start and end of the year interval, respectively. Importantly, adjacent elements of Y_i need not differ, for example, y_{i2} would equal y_{i3} in the instance that the third event experienced by individual i was a one-off. An example demonstrating use of this notation for a single individual is depicted in Figure 3.3.

The likelihood can then be expressed as the product over all individuals and all events

$$L = \prod_{i=1}^{n} \Big[\prod_{j=0}^{m_i} S_{y_{ij}}(t_{ij}, t_{i(j+1)}) \Big] \Big[\prod_{j=0}^{m_i - 1} q_{y_{ij} y_{i(j+1)}}(t_{ij}) \Big].$$

Following maximization of the log likelihood function using numerical optimization techniques, the covariance matrix for the parameter estimates can be obtained by inverting the negative Hessian. The square root of the diagonal elements of this covariance matrix are asymptotically equal to the standard errors for the corresponding parameter estimates.

3.4 Estimation of Concurrency Metrics

As a result of the bivariate Markov model specification, which implies constant event intensities, the concurrency metric estimators can be expressed in terms of the model parameter estimates. Let C_k be an integer valued random variable representing the number of one-offs that occur from the moment an individual enters a state of k ongoing partnerships until the individual leaves that state by either forming or dissolving an ongoing partnership. Let H_k be the random variable indicating the concurrent partnership sojourn time, that is the duration of time an individual remains in a state of k partnerships prior to an ongoing partnership formation or dissolution event. For an individual in a state of k ongoing partnerships who has experienced r one-offs since the last partnership formation or dissolution event, let μ_{kr} denote the mean duration of time until the next event (partnership formation, dissolution, or one-off). Thus, μ_{kr} is the mean inter-event time after entry into a state of k ongoing partnerships and after a total of r one-offs since entry into the current state. Inter-event times are exponentially distributed because the event occurrence intensities are constant given k and r. Therefore, for a fixed $C_k = c$, H_k will be equal to the amount of time spent in state k with a cumulative total of exactly 0 one-offs, plus the amount of time spent in state k with a cumulative total of exactly 1 one-off, summing all the way up to c one-offs. For an individual in a state of k ongoing partnerships who has experienced r one-offs, let Δ_{kr} represent the probability that the next event (formation, dissolution, or one-off) that occurs is either a partnership formation or dissolution event resulting in escape from the state of k ongoing partnerships. Therefore, $P(C_k = 0) = \Delta_{k0}$ and $P(C_k = 1) = \Delta_{k1}(1 - \Delta_{k0})$ which is equal to the probability of the first event being a one-off and the second event being the formation or dissolution of an ongoing partnership. The mean concurrent partnership sojourn times for all $k \in \{0, 1, 2, \ldots\}$ can then be derived as follows using iterative expectation,

$$\begin{aligned}
\rho_k &= E(H_k) \\
&= E_{C_k}(E(H_k|C_k = c)) \\
&= E_{C_k}\Big(\sum_{r=0}^{c} \mu_{kr}\Big) \\
&= \sum_{c=0}^{\infty}\Big(\Big(\sum_{r=0}^{c} \mu_{kr}\Big)P(C_k = c)\Big) \\
&= \mu_{k0}\Delta_{k0} + \sum_{c=1}^{\infty}\Big(\Big(\sum_{r=0}^{c} \mu_{kr}\Big)\Delta_{kc}\prod_{s=0}^{c-1}(1-\Delta_{ks})\Big)
\end{aligned}$$

$$\hat{\rho}_k = \hat{\mu}_{k0}\hat{\Delta}_{k0} + \sum_{c=1}^{\infty}\Big(\Big(\sum_{r=0}^{c} \hat{\mu}_{kr}\Big)\hat{\Delta}_{kc}\prod_{l=0}^{c-1}(1-\hat{\Delta}_{kl})\Big). \tag{3.3}$$

To calculate $\hat{\rho}_k$, we first define

$$\begin{aligned}
\hat{\alpha}_{klr} &= E\big[\hat{\alpha}_{kl}(t)|Y(t) = k, N(t) = r\big] \\
&= \hat{\beta}_{kl0}\exp\Big(\hat{\beta}_{kl1}r + \hat{\boldsymbol{\beta}}_{kl2}^{T}E\big[\boldsymbol{X}(t)|Y(t) = k, N(t) = r\big]\Big)
\end{aligned} \tag{3.4}$$

$$\begin{aligned}
\hat{\lambda}_{kr} &= E\big[\hat{\lambda}_k(t)|Y(t) = k, N(t) = r\big] \\
&= \hat{\gamma}_{k0}\exp\big(\hat{\boldsymbol{\gamma}}_{k1}^{T}E\big[\boldsymbol{X}(t)|Y(t) = k, N(t) = r\big]\big).
\end{aligned} \tag{3.5}$$

Depending on the variables comprising $\boldsymbol{X}(t)$, $E\big[\boldsymbol{X}(t)|Y(t) = k, N(t) = r\big]$ can usually be estimated given the available data. For instance, if $X(t)$ is an indicator for engagement in a main partnership, $E\big[X(t)|Y(t) = k, N(t) = r\big]$ would simply equal the probability of being in at least one main partnership for an individual who experienced r one-offs since transition into a state of k ongoing partnerships. As a result of the homogeneous Markov assumption which implies independent and exponentially distributed inter-event times,

$$\hat{\mu}_{kr} = \big[\hat{\alpha}_{k(k+1)r} + \hat{\alpha}_{k(k-1)r} + \hat{\lambda}_{kr}\big]^{-1}$$

$$\hat{\Delta}_{kr} = \frac{\hat{\alpha}_{k(k+1)r} + \hat{\alpha}_{k(k-1)r}}{\hat{\alpha}_{k(k+1)r} + \hat{\alpha}_{k(k-1)r} + \hat{\lambda}_{kr}}$$

for all $k \in \{1, 2, \ldots\}$. For $k = 0$, $\hat{\mu}_{kr} = \big[\hat{\alpha}_{01r} + \hat{\lambda}_{0r}\big]^{-1}$ and $\hat{\Delta}_{kr} = \big[\hat{\alpha}_{01r}\big]\big[\hat{\alpha}_{01r} + \hat{\lambda}_{0r}\big]^{-1}$. For practical purposes, in estimating the mean concurrent partnership sojourn times, infinite sums can typically be truncated such that the sum extends only until $P(C_k = c)$ becomes negligible.

Assuming stationarity of the bivariate continuous-time Markov process, an estimator for the concurrent partnership distribution can be derived by solving the equilibrium equation $\boldsymbol{\pi Q} = \mathbf{0}$ for $\boldsymbol{\pi}$ where $\boldsymbol{Q}$ is the infinitesimal generator matrix for the two-dimensional process [48]. We will assume maximal values for possible counts of ongoing partnerships and one-offs to obtain a finite dimensional $\boldsymbol{Q}$ matrix and to make the calculations tractable. The resulting approximation is thus accurate up to arbitrary numerical error stemming from truncation of the state space. Allowing a maximum of K ongoing partnerships and R one-offs, such that $\boldsymbol{\pi} = \{\pi_{00}, \pi_{10}, \pi_{20}, .., \pi_{K0}, \pi_{01}, \pi_{11}, \ldots, \pi_{KR}\}$ where π_{kr} denotes the probability that, at any given point in time, an individual is engaged in k ongoing partnerships after the occurrence of r one-offs since the last formation or dissolution event. For $r \in \{0, 1, \ldots, R\}$, let

$$\mathbf{\Lambda}_r = diag\left(\hat{\lambda}_{0r}, \hat{\lambda}_{1r}, \hat{\lambda}_{2r}, \ldots, \hat{\lambda}_{Kr}\right),$$

$$\mathbf{\Gamma}_r = diag\left(-\hat{\alpha}_{01r}, -\hat{\alpha}_{10r} - \hat{\alpha}_{12r}, -\hat{\alpha}_{21r} - \hat{\alpha}_{23r}, \ldots, -\hat{\alpha}_{(K-1)(K-2)r} - \hat{\alpha}_{(K-1)Kr}, -\hat{\alpha}_{K(K-1)r}\right),$$

$$\boldsymbol{A}_r = \begin{bmatrix} 0 & \hat{\alpha}_{01r} & 0 & 0 & \cdots & 0 \\ \hat{\alpha}_{10r} & 0 & \hat{\alpha}_{12r} & 0 & \cdots & 0 \\ 0 & \hat{\alpha}_{21r} & 0 & \hat{\alpha}_{23r} & \cdots & 0 \\ 0 & 0 & \hat{\alpha}_{32r} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & \hat{\alpha}_{K(K-1)r} & 0 \end{bmatrix},$$

where $\mathbf{\Lambda}_r$, $\mathbf{\Gamma}_r$, and $\boldsymbol{A}_r$ are $(K+1)$-dimensional square matrices. Then, $\boldsymbol{Q}$ is a $(K+1)(R+1)$-dimensional square matrix that can be defined using the above notation and a series of block matrices,

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{A}_0 + \boldsymbol{\Gamma}_0 - \boldsymbol{\Lambda}_0 & \boldsymbol{\Lambda}_0 & 0 & 0 & \cdots & \\ \boldsymbol{A}_1 & \boldsymbol{\Gamma}_1 - \boldsymbol{\Lambda}_1 & \boldsymbol{\Lambda}_1 & 0 & \cdots & 0 \\ \boldsymbol{A}_2 & 0 & \boldsymbol{\Gamma}_2 - \boldsymbol{\Lambda}_2 & \boldsymbol{\Lambda}_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ \boldsymbol{A}_R & 0 & 0 & 0 & 0 & \boldsymbol{\Gamma}_R \end{bmatrix}.$$

The structure of $\boldsymbol{Q}$ is such that the $\boldsymbol{A}_0 + \boldsymbol{\Gamma}_0 - \boldsymbol{\Lambda}_0$ block yields the transition rates between states $(k, 0)$ and $(l, 0)$ and the first $\boldsymbol{\Lambda}_0$ block yields the transition rates between states $(k, 0)$ and $(k, 1)$ [45, 49]. The rest of the generator matrix is structured similarly. There is insufficient information to solve the set of balance equations resulting from $\boldsymbol{\pi Q} = 0$ and we must therefore incorporate our knowledge that $\sum_{i=0}^{K}\sum_{j=0}^{R}\pi_{ij} = 1$. After solving for $\boldsymbol{\pi}$, we obtain the concurrent partnership distribution by summing across numbers of one-offs such that $\hat{\pi}_k = \sum_{j=0}^{R}\pi_{kj}$ for all $k \in \{0, 1, 2, \ldots, K\}$.

Estimation of standard errors for all $\hat{\rho}_k$ and $\hat{\pi}_k$ can be completed using a nonparametric bootstrap approach for multistate processes [50, 51]. For an observed sample of size n, the approach entails sampling with replacement a total of n individuals and using all of each sampled individual's sexual history data to calculate $\hat{\rho}_k$ and $\hat{\pi}_k$ for $k \in \{0, 1, \ldots\}$ as described above. This entails fitting the proposed model, obtaining the parameter estimates and then using the formulas presented in this section to calculate the concurrency metrics of interest. This resampling process is repeated until g bootstrap samples have been drawn and estimates computed where g is usually large. The variances of $\hat{\rho}_k$ and $\hat{\pi}_k$ can then be estimated as the empirical variances of the g replicates of $\hat{\rho}_k$ and $\hat{\pi}_k$.

3.5 Application to the MetroMates Study

The retrospective sexual history data that motivated the development of the proposed model came from a National Institute of Drug Abuse (NIDA)-funded research study officially titled *Transmission Behavior in Partnerships of Newly HIV Infected Southern Californians* and commonly referred to as the MetroMates study (PI: Dr. Pamina Gorbach). Between February 2009 and May 2012, MSM seeking testing for HIV through the Sexual Health Program at the Los Angeles LGBT Center were recruited to participate in the MetroMates study involving a baseline interview, testing for HIV and other sexually transmitted infections, and a year of follow-up interviews. Criteria for enrollment included: male, at least 18 years of age, report of sex with a male partner in the past 12 months, and a new HIV test. Demographic, behavioral, and other data were collected using Audio Computer-Assisted Self-Interview (ACASI). Data were collected at the respondent level and respondents could elect to provide information for up to six named partners with whom they reported having sexual intercourse within the past year. Using the calendar method, respondents reported the lengths of time since first and last intercourse in days, weeks, months, or years creating variation in precision. To distinguish between partnerships that were ongoing versus dissolved at the time of the survey, responses to an item asking how likely it is that a respondent will have sex with the partner again were used. Responses of ”extremely unlikely” and ”very unlikely” were assumed to indicate a terminated partnership.

Data were collected for 326 participants in the MetroMates study. Among these participants, 1,050 partnerships were reported. Invalid partnerships consisting of 64 partnerships with missing first or last dates of intercourse, 39 partnerships with a last date of intercourse preceding the first date of intercourse, and 47 partnerships with last dates of intercourse prior to the year interval were excluded. Following these exclusions, data were available for 295 male participants with at least one valid partnership. Participants ranged in age from 19 to 62 years (mean = 30.03, standard deviation = 7.85). The MetroMates study protocol called for oversampling of HIV positive men. Among the 295 respondents, 196 received a positive HIV diagnosis and the remaining 99 were HIV negative at the time of the survey. The MetroMates study also selectively enrolled men whose new HIV diagnosis suggested a recent or acute infection. As described by Gorbach et al. [52], during the initial phase of enrollment only men with a recent diagnosis were recruited. To complete enrollment, men with any new diagnosis, including chronically infected men, were recruited. Of the 295 men included in our sample, 74% reported one or more one-off during the year interval for a total of 534 one-offs. Of the 896 partnerships reported, 60% were one-offs, 7% were of duration

30 days or less, 24% were of duration 31-365 days, and 9% were reported as lasting longer than 365 days.

The Markov nature of the proposed model assumes exponentially distributed sojourn times conditional on the number of ongoing partnerships and the number of one-offs having occurred since the last partnership event. To assess the appropriateness of this assumption for the MetroMates sample, we performed graphical diagnostics [53]. Specifically, we plotted the Nelson-Aalen estimated cumulative hazard rate versus time for each condition defined by HIV status, the number of ongoing partnerships, and number of one-offs. For conditions with a sufficient sample size, we assessed the linearity of the plotted curve. Across most conditions, the assumption of exponentially distributed sojourn times appeared valid.

In applying the modeling approach described previously to the MetroMates data, we fit a number of models including explanatory variables such as respondent age and HIV status. In the model selection stage, explanatory variables were incorporated into either or both the multistate and point process components of the model. The model we selected for presentation included the number of one-offs and an indicator for HIV status as covariates in the multistate portion of the model. One-off event rates were estimated separately for individuals in no ongoing partnerships, one ongoing partnership, and concurrent partnerships. The log likelihood function was constructed using code written in R version 3.2.0 and available in the supplementary materials. Minimization of the negative log likelihood function was accomplished using the general-purpose optimization function *optim* available in the base R stats package. The Nelder-Mead direct search method was specified and differing sets of initial values were used to verify the results obtained. To enable calculation of standard errors, a numerical approximation to the Hessian matrix was generated using the R *numDeriv* package. Parameter estimates for the model fit to the MetroMates data are displayed in Table 3.1. Relative to HIV negative men, HIV positive men were estimated to have higher hazard of forming a monogamous partnership, higher hazard of forming a concurrent partnership, and lower hazard of partnership dissolution during the previous year, although these results were not statistically significant.

Table 3.1: Parameter estimates for the joint multistate and Poisson process model fit to the MetroMates data

Event Type	**Parameter**	**Description**	**Estimate**	**Standard Error**	***P* Value**	**Hazard Ratio (95% CI)**
Formations	β_0	Baseline Hazard	0.0019	0.0003		
from State 0	β_1	Count of One-offs	-0.1812	0.1511	0.23	0.83 (0.62, 1.12)
	β_2	HIV Status	0.1520	0.1923	0.43	1.16 (0.80, 1.70)
Formations	β_0	Baseline Hazard	0.0026	0.0004		
from State ≥ 0	β_1	Count of One-offs	-0.1529	0.2839	0.59	0.86 (0.49, 1.50)
	β_2	HIV Status	0.0704	0.1982	0.72	1.07 (0.73, 1.58)
Dissolutions	β_0	Baseline Hazard	0.0053	0.0006		
	β_1	Count of One-offs	0.4457	0.1097	< 0.01	1.56 (1.26, 1.94)
	β_2	HIV Status	-0.1615	0.1403	0.25	0.85 (0.65, 1.12)
One-off Events	γ_0	0 partnerships	0.0048	0.0003	< 0.01	
	γ_1	1 partnership	0.0022	0.0003	< 0.01	
	γ_2	≥ 2 partnerships	0.0038	0.0006	< 0.01	

The number of one-offs was significantly associated with rates of subsequent partnership dissolution. Following the occurrence of each additional one-off, an individual was estimated to experience a 56% increase in the hazard of dissolution of an ongoing partnership.

Using the analytic expressions derived previously and the parameter estimates in Table 3.1, the population concurrency metrics were estimated. The concurrent partnership distribution and the mean concurrent partnership sojourn times were estimated separately for HIV positive and negative individuals in this sample (Table 3.2). Standard errors for the concurrency metrics were calculated based on $g = 1000$ bootstrap samples. The concurrent partnership distribution was calculated across states ranging from 0-7 ongoing partnerships and 0-4 one-offs, as these ranges encompassed the majority of the observed data. States of ≥ 2 ongoing partnerships were combined for presentation in Table 3.2. At any given point in time, approximately 18% of the HIV positive sample would be expected to be engaged in concurrent partnerships as compared to 10% of the HIV negative sample. Sixteen percent of the HIV negative sample was estimated to be engaged in a monogamous partnership at any given point in time relative to 19% of the HIV positive sample.

The mean concurrent partnership sojourn times for states of 0, 1, and 2 or more ongoing partnerships are displayed in Table 3.2. Regardless of HIV status, the mean length of time an individual was expected to remain engaged in a state of 2 or more partnerships prior to forming or dissolving a partnership was approximately 4 months. The mean duration of time spent in a state of one ongoing partnership was also approximately 4 months and did not appear to differ substantially according to HIV status. HIV negative individuals were estimated to remain in a state of no ongoing partnerships for an average duration of 15.5 months, as compared to approximately 14.5 months among HIV positive individuals.

The mean numbers of one-offs per year for individuals engaged in no ongoing partnerships, one ongoing partnership, or concurrent partnerships were obtained by taking the inverse of each element of $\hat{\gamma}$. Not surprisingly, individuals in a single monogamous partnership had the lowest estimated rate of one-offs per year (0.81). On average, men engaged in concurrent partnerships experienced an estimated 1.40 one-offs per year and men engaged in no partnerships experienced 1.75 one-offs per year.

Table 3.2: Population partnership metric estimates based on parameter estimates obtained from the joint model fit to the MetroMates data

	Number of Ongoing Partnerships	**Concurrent Partnership Distribution**		**Mean Concurrent Partnership Sojourn Time**		**Mean Number of One-Offs per Year**	
	k	$\hat{\pi}_k$	SE	$\hat{\rho}_k$	SE	Estimate	SE
HIV -	**0**	0.7384	0.1237	466 days	80 days	1.75	0.10
	1	0.1573	0.0524	119 days	18 days	0.81	0.10
	$\geq$ **2**	0.1043	0.1168	114 days	18 days	1.40	0.20
HIV +	**0**	0.6325	0.0800	441 days	64 days	1.75	0.10
	1	0.1880	0.0414	129 days	12 days	0.81	0.10
	$\geq$ **2**	0.1795	0.0565	123 days	12 days	1.40	0.20

3.6 Discussion

We have described a novel approach for the joint modeling of sexual partnership patterns using retrospective sexual history data containing one-off sexual encounters. The proposed model can be applied to answer pertinent questions in the field of HIV transmission research. Implementation of this approach was demonstrated using epidemiological data collected from a sample of MSM seeking HIV testing at a Los Angeles clinic. Despite the limitations associated with retrospective sexual history survey data, we were able to estimate several important population concurrency metrics using a technique that accounted for different sources of variation and fully utilized the available data.

The joint multistate and point process model addresses all of the modeling objectives outlined previously. The proposed method accounts for dependence among partnerships engaged in by the same person at the same or different points in time by translating the data collected at the partnership-level into individual-level trajectories and modeling these trajectories as independent stochastic processes. Another advantage of the joint model is the explicit modeling of rates of partnership formation and dissolution. Many of the agent-based and other mathematical models constructed to examine the impact of concurrency on HIV transmission have relied on simple empirical estimates of the mean partnership duration or concurrent partnership distribution as input [36, 33, 54]. Our proposed method provides improved estimates of these quantities but also provides formation and dissolution rates that are perhaps more useful in creating a dynamic mathematical model involving forward simulation of concurrent partnership patterns over time. As shown in Figure 3.4, state transition intensities and one-off rates estimated based on the MetroMates data can be easily used to generate simulated sexual partnership trajectories at the individual level. Rates of partnership and one-off events that are dynamic with respect to time could be useful in adapting current network models such that the probabilities of a partnership formation or dissolution between two individuals in a network are variable and more accurately reflect the sexual partnership patterns observed in a a population. The proposed joint model is also flexible enough to allow for the incorporation of explanatory variables to further account for heterogeneity between individuals. Lastly, we have developed a joint model that includes the random occurrence of one-offs. This important extension enables examination of the relative importance of one-offs in driving the spread of HIV within a population and also allows for one-offs to impact HIV transmission indirectly, by affecting rates of subsequent partnership formation and dissolution. Among populations such as the MSM surveyed in the MetroMates study, the high reported rate of one-offs makes this a valuable feature of the proposed model. The joint modeling approach also distinguishes between one-offs and short-term partnerships which may be important for determining factors impacting HIV transmission. One-offs could potentially have a higher probability of transmission for a given sexual encounter due to differences in the type of sex occurring during a one-off. For example, one-offs may be more frequently associated with drug use leading to longer duration of sex or more vigorous sex which could in turn enhance infectiousness. The proposed model could enable identification of such differences in the risk of transmission.

Participants in this study do not represent a random sample of all MSM living within Los Angeles nor do they represent all MSM living within the community served by the Los Angeles LGBT Center. This sample was obtained by recruiting individuals who sought HIV testing and the recent sexual activity they reported on would be expected to include behaviors that influenced their decision to seek testing. Further, the study protocol called for the oversampling of HIV positive individuals and, in particular, recently-infected HIV positive individuals [55]. Thus, the generalizability of results presented in this study is limited. We assume that the removal of invalid partnerships resulting in the exclusion of 31 respondents did not significantly bias our results although we have limited means of assessing this assumption. Sixty-one percent of the 31 excluded individuals were HIV positive as compared to 66% of individuals included in the analyzed data. In removing invalid partnerships, we further acknowledge that the partnership rate estimates presented here could be biased downward if the removed partnerships represented actual partnerships occurring during the year interval.

Several sources of uncertainty were present in our analysis of the MetroMates data. Since respondents were only allowed to report on a maximum of six sexual partnerships

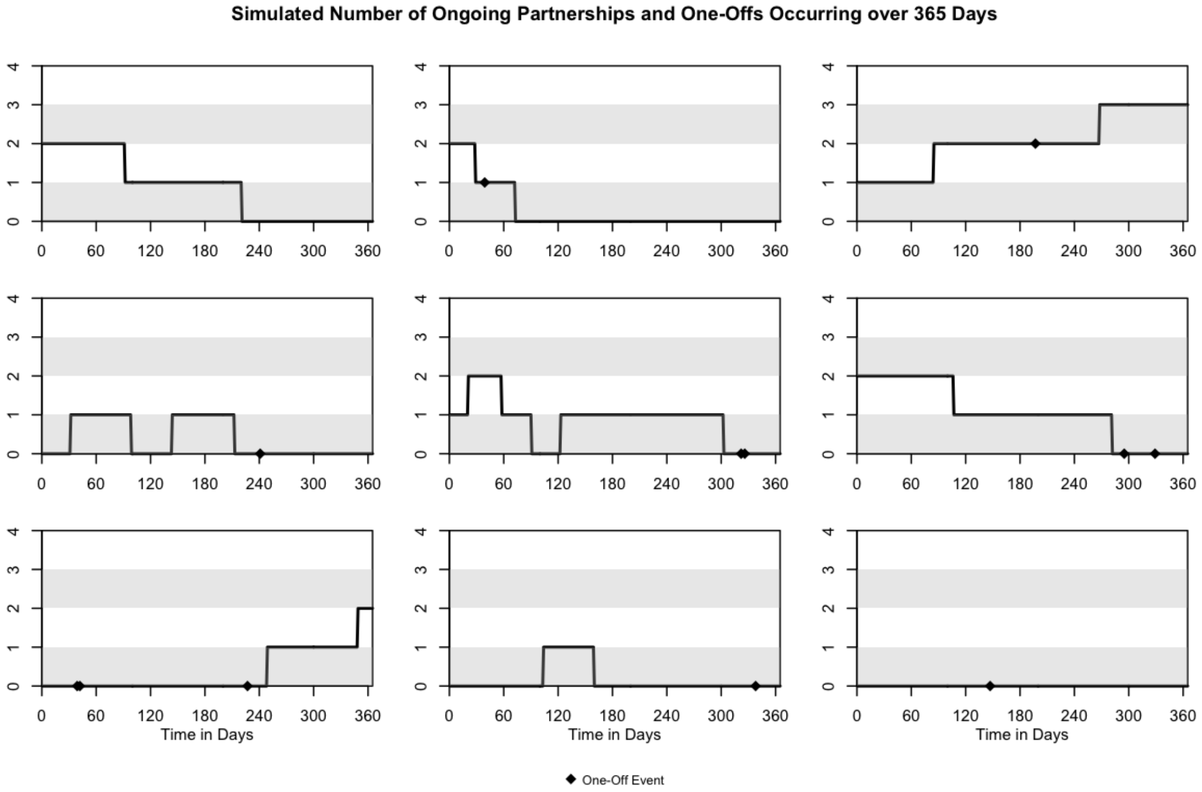


Figure 3.4: Simulated trajectories displaying the number of ongoing partnerships and one-off occurrences over 365 days for 9 individuals based on parameter estimates obtained from the model fit to the MetroMates data. Individuals in the top two rows were assumed to be HIV positive and those in the bottom row were assumed to be HIV negative.

that occurred in part or in full during the previous year, individuals engaging in larger numbers of partnerships across the year interval may have provided incomplete partnership data that could potentially bias the estimates presented. Of the 295 participants included in the final sample, 60 (20.3%) reported on 6 partnerships. Methods to address this issue in future studies need further development but could include alternative questionnaire designs or consideration of subject-specific time intervals of observation during the analysis stage. Additionally, since respondents were allowed to choose the unit of measurement with which they reported time since first and last dates of sexual intercourse, dates used in analyses were often approximated. Future studies are required to explore the potential impact of this source of uncertainty, especially in the context of multistate models with bidirectional transitions. Similarly, this issue of coarseness in the reporting of dates introduces uncertainty surrounding the distinction between one-offs and ongoing partnerships of short duration which is an area for future work. Lastly, due to the questionnaire instructions, respondents were not asked to report partnerships occurring prior to the year interval and therefore the number of one-offs an individual had engaged in at the start of the year interval was unknown. In analyzing the MetroMates data, we assumed zero one offs having occurred since the last formation or dissolution event, which could potentially bias our results. Future studies may choose to consider attempting to capture or impute this missing data.

In considering these results, it is important to recall the sampling approach with regard to HIV status. The HIV positive sample received their positive diagnosis at the time of the survey. Thus, the sexual behaviors these individuals were reporting on occurred prior to their knowledge of their HIV status. The sexual patterns attributed to HIV positive men within this sample should not be assumed to reflect the behaviors an HIV positive man aware of his status would engage in. Additionally, some of the behaviors reported on by recently infected HIV positive individuals within this sample may have occurred prior to the individual's acquisition of HIV. Although the retrospective reporting of the data relative to the date of diagnosis limits some of the conclusions that can be drawn, the timing of calendar method data collection may be advantageous when attempting to answer questions about the association between concurrency and acquisition of HIV. If a significant association between concurrency and subsequent diagnosis of acute HIV infection had been identified, it would not directly support the hypothesis that concurrency impacts HIV transmission at the population-level. In theory, an individual who engages in concurrent partnerships does not put him or herself at greater risk than if he or she had engaged in the same numbers and types of risky behaviors with the same individuals but in a serially monogamous setting. Therefore, we would expect an increase in the rate of transmission among individuals engaging in concurrent partnerships but not necessarily an increase in the rate of acquisition. It is, however, reasonable to consider that individuals engaging in concurrent partnerships are (1) also engaging in more total partnerships and engaging in risky behaviors at a greater rate relative to individuals in monogamous partnerships, and (2) more likely to be engaging in concurrent partnerships with individuals who themselves are engaging in concurrent partnerships. Both of which could explain an association between increased point prevalence of concurrency and subsequent diagnosis with HIV among samples reporting retrospective sexual history data at the time of screening.

We have demonstrated implementation of this joint modeling approach using model specifications that were selected to be appropriate for use with the MetroMates data and to reduce complexity in this initial presentation of the proposed model. Future applications of this model for analysis of sexual history data could select a different set of covariates, including the addition of other time-varying explanatory variables such as partnership-level characteristics. Although the assumption of stationarity is critical for calculation of the concurrency metrics as described herein, inclusion of covariates such as respondent age or calendar date at the time of interview is possible. In this instance, the concurrency metrics can be calculated for categorical age or date strata as done for HIV status in the present application, or calculations can be completed after taking the expected value of these covariates as shown in equations (3.4) and (3.5). Although the presented model for the MetroMates data did not include any explanatory variables significantly associated with the rate of one-offs, the parametric formulation of the point process rate function can easily accommodate inclusion of these variables. A simple modification to the proposed model would allow for a different definition of $N(t)$. For instance, one might elect to let $N(t)$ reflect the count of one-offs occurring only during some specific time interval prior to time t, for instance, one month, such that the impact of one-offs on subsequent events is limited in duration. The proposed model is also general enough to allow for selection of different counting processes in instances when the assumptions surrounding the Poisson process are not valid. For instance, when it is not acceptable to assume that the variance of the counts of one-offs over any given interval of time equals the mean, an alternative counting process distribution, such as the negative binomial, may be more appropriate. Another consideration is the use of zero-inflated count models in instances in which time intervals during which no one-offs occur are observed in excess. The model specified herein also assumes a bivariate continuous-time Markov structure. This framework requires that transition intensities are constant within subintervals of time defined by the occurrence of one-offs, allowing transition intensities from one state to another to differ across the interval of time spent in a given state. Advantages of this framework are the flexibility to allow one-offs to affect subsequent intensities and ease of construction of the likelihood. Alternative non-Markovian models that do not rely on the phase-type intensities assumption are possible although the derived concurrency metric estimators would not be directly applicable.

Future applications of the proposed model to sexual history data may use the general joint multistate and point process framework presented here and alternatively adapt it to meet their needs. Researchers investigating sexual partnership dynamics impacting HIV transmission should consider analyzing sexual history data using a modeling approach such as the one proposed here, that jointly models both ongoing and one-off sexual partnerships and treats the individual, rather than the partnership, as the independent unit of observation.

CHAPTER 4

A Semi-Markov Model for Disease Progression

4.1 Challenges in Estimating Semi-Markov Models with Back Transitions

Methods for analyzing panel data under a Markov multistate model with time-homogenous transition intensities have been developed and are widely used in application [23, 56]. These methods rely on the simplifying assumption that a Markov model observed at pre-specified time points forms a discrete-time Markov process. This result does not pertain, however, to semi-Markov models for which the transition intensities depend on time elapsed in the current state. In many applications, semi-Markov models are preferable or even necessary to effectively model transition intensities over time. For example, in the study of human papillomavirus (HPV), a semi-Markov assumption is needed to account for the strong association between infection duration and progression to cervical abnormality [57]. The use of a semi-Markov process to model functional status over time in aging research has also been advocated by Cai et al. who cite the earlier finding that the likelihood of functional improvement is higher when loss is more recent [58, 59].

The challenges and the importance of developing approaches for the estimation of intermittently-observed semi-Markov multistate models with back transitions has not been overlooked in the literature. In their recent numerical study examining the loss of information due to intermittent observation, Lawless and Rad state that the effects of intermittent observation on the efficiency of transition intensity and transition probability estimates are much more severe for models allowing back transitions [26]. Wei and Kryscio modeled the flow of elderly subjects from intact cognition to dementia with transient cognitive states and implemented a quasi-Monte Carlo method to enable the higher order integration required to account for the uncertainty arising from unobserved transition instants [60]. In implementing their method, however, Wei and Kryscio assumed no unobserved states such that the entire sequence of states was available for each individual. Kang and Lagakos made an important contribution by demonstrating that when the transition intensities from at least one of the states in a model involving back transitions are assumed to be time homogenous, a tractable expression for the likelihood function is possible [57]. The importance of allowing for duration-dependent disease state sojourn distributions in panel data models with back transitions was also emphasized by Lang and Minin who chose to assume an underlying latent continuous-time Markov chain (CTMC) with multiple latent states mapping to each disease state [61]. The latent CTMC framework results in a model that regains analytic tractability when estimated using an iterative expectation-maximization (EM) algorithm. Although the latent CTMC approach has several advantages, implementation requires specification of the structure for the latent CTMC rate matrix and the dimension of the latent space making model misspecification a significant concern. Additionally, for applications in which sojourn times are assumed to follow a specific non-exponential distribution, such as the Weibull, it is often preferable, for reasons of interpretability and comparability, to obtain estimates for the distributional parameters, such as the Weibull shape and scale, as opposed to phase-type functionals arising under the latent CTMC framework.

In this chapter we propose a method for estimating semi-Markov models for panel data in the presence of back transitions. The proposed method accommodates intermittently-observed data and requires minimal assumptions other than the parametric form of the distribution from which the sojourn times arise.

4.2 Notation for a Semi-Markov Model with Back Transitions

A continuous-time multistate stochastic process is one which can take a finite number of states at realizations occurring across time. Let random variable $Y(t)$ denote the state a process occupies at time t giving us a multistate process consisting of $\{Y(t), t \geq 0\}$ where

$t = 0$ can be defined as the time of origin for the given process. At any time t, let $Y(t)$ take on values in the state space $\{1, 2, \ldots, K\}$ for a process with K states. Denote initial and subsequent consecutive states occupied by the process as $s_0, s_1, \ldots$ and denote the time spent in state s_{m-1} prior to any transition event as d_m. We refer to d_m as the sojourn time. Thus, the process $Y(t)$ can be expressed as $\{s_0, d_1, s_1, d_2, \ldots\}$. If the sequence $\{s_0, s_1, \ldots\}$ forms a simple Markov chain and the sojourn times, d_m are independent random variables with distributions depending only on the adjoining states, s_{m-1} and s_m, the process is referred to as semi-Markov [31]. Distinct from Markov processes, semi-Markov processes allow for dependence of the transition intensity functions on duration in the current state. Letting d denote time elapsed since entry into state i following the $(m-1)$th transition, a semi-Markov process can be fully defined by the set of transition intensities

$$\lambda_{ij}(d) = \lim_{\Delta d \to 0} \frac{P\left(d_m < d + \Delta d, s_m = j | d_m \geq d, s_{m-1} = i\right)}{\Delta d}$$

for $i, j = 1, 2, \ldots, K$, and $i \neq j$. The preferred approach to specifying the transition intensities is to allow $\lambda_{ij}(d) = P_{ij} f_{ij}(d)$ where

$$\begin{aligned} P_{ij} &= P(s_m = j | s_{m-1} = i) \\ f_{ij}(d) &= \lim_{\Delta d \to 0} \frac{P\left(d_m < d + \Delta d | d_m \geq d, s_m = j, s_{m-1} = i\right)}{\Delta d}, \end{aligned}$$

in which trajectories are modeled separately from the times of transitions [62]. In this specification, P_{ij} represents the probability that the next transition experienced by a trajectory in state i will be to state j, assuming the trajectory is followed forward an indefinite amount of time without the possibility of censoring. The density function for the sojourn time in state i before transitioning to state j is a function of time elapsed in the state i and is represented by $f_{ij}(d)$. The multistate model is specified such that $0 \leq P_{ij} \leq 1$ for all $i, j = 1, 2, \ldots, K$ and $i \neq j$, and $\sum_{j \neq i} P_{ij} = 1$ for all $i = 1, 2, \ldots, K$. Additionally, assumed model structure may dictate that $P_{ij} = 1$ or $P_{ij} = 0$ for transitions in which j is the only state that can be progressed to directly from state i, or for transitions that cannot occur, such as from death to a healthy state. One example of a model structure typical to disease progression modeling is shown in Figure 4.1 with arrows representing directional transitions for which

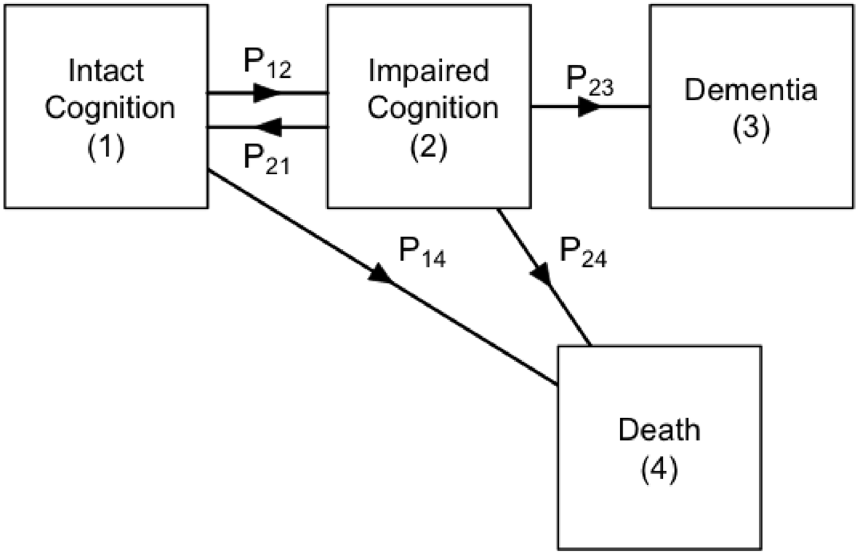


Figure 4.1: Structural diagram for the multistate model assumed in both the simulation study and Nun Study data analysis

$P_{ij} \neq 0$ and boxes representing states an individual can occupy. In this instance the disease for which progression is being modeled is dementia with cognitive impairment as an intermediate disease state from which recovery is possible.

Although we assume an underlying continuous-time semi-Markov process, it is typically not possible to collect complete data for such a process. For each unit of observation sampled, complete data would consist of the chronological sequence of states the process occupies and the associated sojourn times represented by $\{s_0, d_1, s_1, d_2, \ldots\}$. Instead, we typically have panel data consisting of a discrete series of states a process is intermittently observed to occupy, $\{Y(t_n), n = 1, 2, \ldots\}$. When panel data are collected, information is not available about the types of state transitions and the instants that these transitions occur for periods of time between observations.

To demonstrate the issues arising from panel data collection, we will consider the following simple multistate model consisting of two transient states.

$$\boxed{1} \rightleftarrows \boxed{2}$$

Consider an individual observed at four time points, $0 < t_1 < t_2 < t_3 < t_4$, where $Y(t_1) = 1$,

$Y(t_2) = 1$, $Y(t_3) = 2$, and $Y(t_4) = 1$. We depict these observations chronologically using the following diagram where circled numbers indicate observed states.

$$\textcircled{1} \rightarrow \textcircled{1} \rightarrow \textcircled{2} \rightarrow \textcircled{1} \tag{4.1}$$

Throughout all materials presented herein, we assume a random observation process operating independently from the underlying multistate process being modeled. The above sequence of observations indicates that the individual visited state 1 at least twice, state two at least once, and experienced at least one back transition, meaning that the above sequence of observations is consistent with an ordered path equal to **1-2-1**. However, this is the simplest potential path that can be considered. The above set of observations are also consistent with all of the following sequences of states where we let the presence and absence of a circle around the state number indicate observed and non-observed states, respectively.

$$\begin{gathered}
\textcircled{1} \rightarrow 2 \rightarrow \textcircled{1} \rightarrow \textcircled{2} \rightarrow \textcircled{1} \\
\textcircled{1} \rightarrow \textcircled{1} \rightarrow 2 \rightarrow 1 \rightarrow \textcircled{2} \rightarrow \textcircled{1} \\
\textcircled{1} \rightarrow 2 \rightarrow 1 \rightarrow 2 \rightarrow \textcircled{1} \rightarrow \textcircled{2} \rightarrow \textcircled{1}
\end{gathered}$$

In fact, the above represents a small sample from an infinite number of possible paths involving back transitions between states 1 and 2. As implied by the above diagram, the extent to which panel data incorrectly represent the true underlying path depends on the sparseness of the intermittent observation process that generated the data relative to the true bi-directional transition rates of the multistate process. We will explore this dependence using a simulation study.

When an independent sample of individuals are continuously observed such that complete data are available for each individual, likelihood-based inference is straightforward. The likelihood contribution for an individual who occupies a sequence of $M+1$ states during the period of observation is

$$L = \prod_{m=1}^{M} \left[P_{s_{(m-1)} s_m} f_{s_{(m-1)} s_m}(d_m) \right]^{1-\delta_i} \left[\sum_{j \neq s_m} P_{s_m j} S_{s_m j}(d_m) \right]^{\delta_m}, \tag{4.2}$$

where δ_m is an indicator variable equal to one when the duration in state s_m is right censored and zero otherwise. The likelihood function for the entire sample equals the product of all the individual likelihood contributions calculated using the above formula. Parameter estimates with favorable properties can then be obtained by straightforward maximization of the log likelihood function.

When the sequence of states occupied is not observed but is known to belong to one of a limited number of potential sequences, the likelihood contribution for an individual can be calculated by summing over the contributions associated with each potential path. However, in estimating multistate models with back transitions, this approach results in an infinite summation giving rise to an intractable likelihood function.

4.3 A Minimal Path Approach for Semi-Markov Models with Back Transitions

A number of methods to circumvent the intractability of the likelihood function have been proposed. As noted previously, Kang and Lagakos showed that when the sojourn time distribution for at least one of the states of the process is exponential, the resulting joint probability in the likelihood function becomes tractable [57]. Others have regained tractability by assuming the underlying semi-Markov process can be approximated using phase-type sojourn time distributions implied by a latent CTMC structure [63, 61]. By far the most common approach is to assume that, for a given process, the sequence of states observed across the set of discrete observation times are identical to the sequence of states occupied. Following this approach, a sequence that we refer to as the *minimal path* is inferred from the sequence of states observed under the assumption that no additional unobserved back transitions occurred. By assuming the minimal path represents the true underlying path, the likelihood function becomes tractable.

We define the *minimal path* as the shortest possible path given both the chronological sequence of observations and the assumed underlying structure of the multistate model. The model structure is of importance if two consecutive observations identify a process as occupying two states that are not directly connected to one another according to the adopted structure. In this case, the minimal path will include one or more intermediate states as specified by the structure, with preference given to the most direct path. There could exist progressive model structures for which the most direct path is difficult to identify because of equal path lengths among multiple potential state sequences. In this case, the minimal path definition would need to be extended to include criteria for selecting from among the potential sequences. We do not consider such models in the development that follows.

Although assuming the minimal path allows us to regain tractability of the likelihood function, the unobserved transition times translate to uncertainty regarding the sojourn times associated with each state occupied. This form of uncertainty is inherent to all multistate models fit to panel data, regardless of the potential for back transitions. Since the sequence of states occupied is assumed known, this uncertainty can be addressed using methods developed for interval censored time-to-event data since sojourn times are not exactly observed but are known to fall within a bounded interval [64].

In the previous example, we assumed a sequence of states corresponding to $\{s_0, s_1, s_2\} = \{1, 2, 1\}$, which we will refer to as the *minimal path* associated with the observation process depicted in (4.1). Under this assumption, we can address the sojourn time uncertainty by integrating with respect to the transition times. Thus, the likelihood contribution for this individual becomes

$$L = \int_{t_2}^{t_3} \int_{t_3}^{t_4} f_{12}(u_2) f_{21}(u_1 - u_2) S_{12}(t_4 - u_1) du_1 du_2.$$

Likelihood contributions from longer paths and more complicated state spaces can be calculated following a similar approach with the order of integration increasing for each additional transition assumed to have occurred. Although this example demonstrates the necessary approach for dealing with the uncertainty surrounding transition instants, the computational burden associated with calculating higher order integrals can quickly become prohibitive. Noting this issue, Wei and Kryscio implemented a quasi-Monte Carlo method to compute the higher order integrals required for likelihood-based estimation of a semi-Markov model [60]. Even when the computational burden imposed by these high order integrals can be overcome, the potential bias resulting from assuming a minimal sequence of states is a serious concern. Adopting the minimal path is only appropriate when transitions between states can be assumed to occur sufficiently infrequently relative to the rate of the observation process. Unfortunately, with only panel data available, this assumption is often untestable and investigators must instead rely on ancillary knowledge of the mechanism giving rise to the process to verify assumptions. The impact of assuming the minimal path is the true underlying path on semi-Markov model estimates under varying observation process rates remains to be examined.

4.4 An SEM Algorithm for Intermittently Observed Multistate Processes

In this section, we outline the stochastic EM (SEM) algorithm we propose for use with intermittently observed semi-Markov models with back transitions. In approaching the estimation problems detailed above, we choose to rely on the missing information principle which regards the values of the missing data as random variables within the framework of a model for the data [65]. Since its introduction by Dempster et al., the EM algorithm has become a widely used approach for estimating model parameters in the presence of incomplete data [66]. The EM algorithm has proven most useful for models in which maximum likelihood methods provide a straightforward and efficient estimation approach when the complete data are available. As outlined in Section 4.2, the semi-Markov model likelihood function can be easily constructed and maximized when complete data, consisting of the entire state sequence and all sojourn times (censored and uncensored), are readily available. Thus, in addressing the issue of incomplete data arising from the panel observation process, the EM algorithm is a natural consideration.

The EM algorithm uses successive expectation and maximization steps to account for the uncertainty introduced by the incompletely observed data. Let x denote the observed data and z the unobserved data such that (x, z) refers to the complete data. Let $g(x, z|\boldsymbol{\theta})$ be the joint distribution of the complete data conditional on the parameter vector, $\boldsymbol{\theta}$. In accordance with the missing information principle, our objective is to estimate $\boldsymbol{\theta}$ by finding $\hat{\boldsymbol{\theta}}$ that maximizes the marginal likelihood $\int g(x, z|\boldsymbol{\theta})dz$ obtained from integrating out the unobserved data. The EM algorithm accomplishes this through iterative E- and M-steps. At the rth iteration, the E-step computes

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r-1)}) = E\Big[\log\, g(x, z|\boldsymbol{\theta})\Big] = \int \log\, \big[g(x, z|\boldsymbol{\theta})\big] h(z|x, \boldsymbol{\theta}^{(r-1)}) dz,$$

where $h(z|x, \boldsymbol{\theta}^{(r-1)})$ is the conditional distribution of the unobserved data given the observed data and the current parameter estimate vector, $\boldsymbol{\theta^{(r-1)}}$. In the subsequent M-step, the Q-function above is maximized to find the updated estimate, $\boldsymbol{\theta}^r$ that satisifies

$$Q(\boldsymbol{\theta}^r|\boldsymbol{\theta}^{(r-1)}) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r-1)})$$

for all $\boldsymbol{\theta}$ in the parameter space.

In applying the EM algorithm to intermittently observed semi-Markov models with back transitions, we let $g(x, z|\boldsymbol{\theta})$ equal the likelihood function for the entire sample obtained by taking the product of all the individual likelihood contributions calculated using formula (4.2). In this setting, x corresponds to the panel observations $\{Y(t_n), n = 1, 2, \ldots\}$ and z corresponds to the unobserved sequence of states and transition times. As demonstrated in Section 4.3, accounting for uncertainty by integrating $g(x, z|\boldsymbol{\theta})$ with respect to z, results in analytical intractability. As a result, the expectation that must be taken when calculating the Q-function is not available in closed form and the Q-function is intractable. Even under the simplifying assumption that the path is known to belong to a limited set of potential paths, the computational intensity typically necessitates use of an alternative method for completing the E-step described above. To address these complications, Wei and Tanner proposed an algorithm that stochastically approximates the expectation in the Q-function by taking the Monte Carlo average

$$\bar{Q}_{m_r}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r-1)}) = \frac{1}{m_r}\sum_{k=l}^{m_r} \log\, g(x, z_k|\boldsymbol{\theta}),$$

where $z_1, \ldots, z_{m_r}$ are sampled from $h(z|x, \boldsymbol{\theta}^{(r-1)})$ and m_r represents the iteration-specific Monte Carlo sample size [67]. Repeatedly sampling z from the conditional distribution effectively eliminates the need to integrate $g(x, z|\boldsymbol{\theta})$ with respect to z when calculating the expectation in the Q-function. A Monte Carlo approximation is used in place of the true Q-function. This stochastic implementation of the EM algorithm has been used to overcome the burden associated with intractable or high-dimensional integrals in a variety of applications [68, 69].

Under mild regularity conditions, this stochastic EM algorithm (SEM) has been shown to converge to the maximum likelihood estimate when the Monte Carlo sample size, m_r, is increased across successive iterations [70, 71]. The initial sample size and rate of increase differ across implementations of the algorithm but should be selected to balance out the between-iteration variance attributable to the algorithm stepping toward the true value, and the variance attributable to Monte Carlo error. In general, a low sample size is recommended at early iterations to save simulation resources when the step sizes are large relative to the Monte Carlo error. A higher sample size is required at later iterations so that the Monte Carlo error does not overwhelm the directional movement of the algorithm. Care must also be taken when determining how many iterations of the stochastic EM algorithm are necessary to obtain convergence. The deterministic EM algorithm is usually stopped once the relative change in parameter estimates between two successive iterations is smaller than some pre-specified threshold, but applying this stopping rule to the SEM algorithm may be problematic due to the stochastic variation that persists across all iterations. To reduce the chance of prematurely stopping the SEM algorithm, Booth and Hobert recommend applying the deterministic EM stopping rule that relies on relative change in parameter estimates and mandating the rule be satisfied for several successive iterations [72]. After the stopping criteria has been met, final parameter estimates are also typically taken as the average over estimates obtained at the last several iterations to further reduce the impact of variance attributable to Monte Carlo error. There is a rich literature providing guidance on how to select the appropriate Monte Carlo sample size and stopping criterion for various versions of the SEM algorithm.

4.5 Rejection Sampling of Multistate Processes

When implementing the SEM algorithm, there are also a number of different options for sampling from the target distribution, $g(z|x, \boldsymbol{\theta}^{(r-1)})$, including rejection sampling, importance sampling, and Markov Chain Monte Carlo (MCMC) [73]. Performance of importance sampling has been shown to depend heavily on choice of the importance distribution and MCMC sampling yields dependent samples thereby requiring an arbitrary number of initial simulations to be discarded. For the purpose of estimating a multistate model with intermittent observations, we propose a rejection sampling approach that generates independent, identically distributed samples by thinning out samples taken from a candidate distribution through rejection of those that are not appropriate. In this proposed application, we define the term *not appropriate* to mean inconsistent with the entirety of the observed intermittent data, as described in more detail below. Rejection sampling can encounter difficulties when a good candidate distribution is hard to obtain and when the acceptance rate is low. Fortunately, the settings in which estimation of semi-Markov model parameters using the minimal path method is most likely to result in bias are the same sparse observation settings in which a rejection sampling approach is likely to work well. When the panel observation process is relatively sparse, implying a high proportion of incomplete data, it is computationally feasible to use rejection criteria stipulating that the sampled path is entirely congruent with the observed data. Such a criterion ensures that each sampled trajectory retains all the information available in the observed data, x, while sampling at the rth iteration from a broad candidate distribution that exactly matches the multistate model specified by $\theta^{(r-1)}$.

Figure 4.2 depicts implementation of the proposed rejection sampling procedure for a single trajectory. The individual in Figure 4.2 is observed at 5 time points with the first two observations occurring while the individual is in state 1, the next two observations occurring while the individual is in state 2, and the final observation occurring when the individual is in absorbing state 3. Let **A** indicate the first trajectory sampled from the candidate distribution. In comparing **A** to the observation process, we reject **A** because of two discrepancies. Sampled trajectory **A** placed the individual in state 2 at the time of both the second observation (observed to be in state 1) and the fifth observation (observed to be in state 3). After rejecting **A**, we proceed to sample trajectory **B** which is rejected due to one discrepancy at the fifth observation (observed to be in state 3). The third sampled trajectory, **C**, is accepted because it is in no way discrepant with the series of

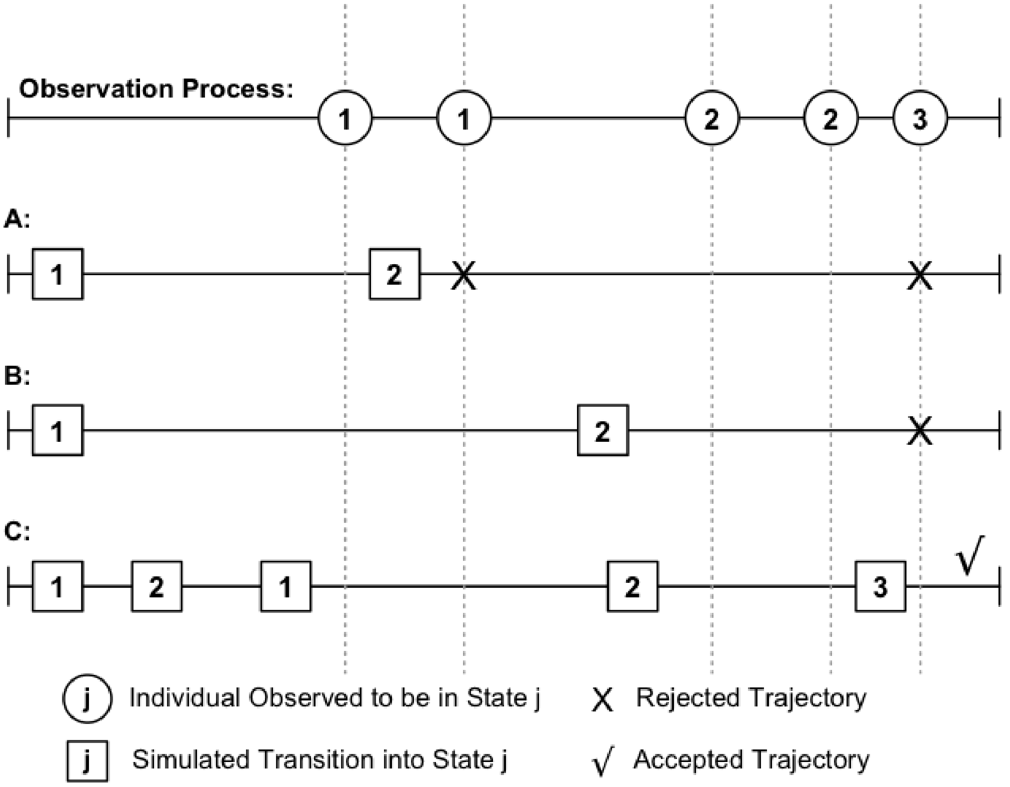


Figure 4.2: The SEM algorithm's rejection sampling scheme applied to a single individual. Trajectories A and B would be rejected after comparison to the observation process at the top and trajectory C would be accepted.

real observations. Generation and comparison of sampled trajectories can be easily and quickly accomplished in most circumstances in which the state space is relatively small and the observation process is somewhat sparse such that the sampling acceptance rate is not prohibitively low.

4.6 A Simulation Study: Design

Simulated data sets were generated to compare three different estimation approaches: estimation with continuously observed paths (Section 4.2), minimal path estimation with intermittently observed paths (Section 4.3), and SEM estimation with intermittently observed paths (Section 4.4). One hundred data sets consisting of 200 individual trajectories spanning a time interval of 300 days were simulated. The model structure shown in Figure 4.1 consisting of two transient states (1 and 2) and two absorbing states (3 and 4), with the option for back transitions between states 1 and 2 was assumed. Sojourn times for all transitions were simulated from a Weibull distribution such that

$$f_{ij}(t) = \left(\frac{\nu_{ij}}{\sigma_{ij}}\right)\left(\frac{t}{\sigma_{ij}}\right)^{(\nu_{ij}-1)} \exp\left(-\left(\frac{t}{\sigma_{ij}}\right)^{\nu_{ij}}\right) \tag{4.3}$$

for $i = 1, 2$, $j = 1, 2, 3, 4$, $i \neq j$, $\sigma_{ij} > 0$, and $\nu_{ij} > 0$. The assumed Weibull distribution also gives us the corresponding survival distribution

$$S_{ij}(t) = \exp\left(-\left(\frac{t}{\sigma_{ij}}\right)^{\nu_{ij}}\right)$$

for $i = 1, 2$, $j = 1, 2, 3, 4$, $i \neq j$, $\sigma_{ij} > 0$, and $\nu_{ij} > 0$. Adopting this formulation, ν_{ij} and σ_{ij} are referred to as the shape and scale parameters, respectively, with $\nu_{ij} = 1$ reducing to a Markov process with exponentially distributed sojourn times. Weibull shape and scale parameter and transition probability values used in these simulations are displayed in Table 4.1 and were selected to represent probabilities that might be observed among an elderly population transitioning between states of health, mild illness and severe illness with death as a competing risk. Each simulated trajectory began the 300-day interval in state 1 and ended after having either entered one of the absorbing states, or by being censored in one of the transient states at the end of the interval.

For each data set, complete data maximum likelihood estimation was implemented by calculating the product of all the individual likelihood contributions in (4.2) and maximizing the log likelihood using the R *constrOptim* function for linearly constrained optimization. Linear constraints were used to ensure the following:

$$\nu_{ij} > 0, \sigma_{ij} > 0 \text{ for } (i, j) \in \{(1, 2), (1, 4), (2, 1), (2, 3), (2, 4)\}$$

$$0 \leq P_{ij} \leq 1 \text{ for } (i, j) \in \{(1, 2), (2, 1), (2, 3)\}$$

$$0 \leq P_{21} + P_{23} \leq 1.$$

Estimates of P_{14} and P_{24} could be obtained given the other estimated transition probabilities.

To implement the other two estimation approaches, an observation process was generated for each trajectory in each of the simulated data sets. In generating the observation processes, inter-observation times were assumed to be exponentially distributed with a 50-day expected inter-observation time. Once generated, the series of observations for each individual trajectory were applied to the corresponding simulated complete data trajectory to obtain new panel data sets. To demonstrate the impact that sparseness of the panel observation process has on the performance of parameter estimates obtained after assuming the minimal path, we also simulated exponential observation processes with a 100-day expected inter-observation time. These more sparse observation processes were applied to the same simulated complete data trajectories to obtain a second set of 100 panel data sets.

Panel data maximum likelihood estimation assuming the minimal path proceeded by the same constrained optimization approach as described above, however, calculation of each individual's likelihood contribution was first completed following the approach detailed in Section 4.3. Specifically, the uncertainty arising due to the unobserved transition instants was addressed using integration. This approach often resulted in high order integrals and substantial computational burden, as addressed by Wei and Kryscio [60]. Both the more sparse (50-day expected inter-observation time) and the less sparse (100-day expected inter-observation time) panel data sets were fit using this minimal path estimation approach and the results compared.

The same 50-day expected inter-observation time panel data sets used when implementing the minimal path estimation approach were then used when implementing the SEM estimation approach. The rejection sampling procedure was chosen to be stringent in the sense that it mandated complete coherence with the observed data resulting in a relatively low acceptance rate. Thus, we found that the Monte Carlo sample size, m_r, did not need to be large to demonstrate reasonable convergence. We chose to set $m_1 = 1$ and to increment m_r by 1 at each successive iteration. The stopping rule we adopted required that the following hold for three consecutive iterations

$$\max_i \left(\frac{|\theta_i^{(r)} - \theta_i^{(r-1)}|}{|\theta_i^{(r-1)}| + \delta_1} \right) < \delta_2, \tag{4.4}$$

where $\delta_1 = 0.001$ and $\delta_2 = 0.05$ for the purpose of this simulation study [72]. After satisfaction of the stopping rule, the final estimates were taken as the average of the estimates obtained during the last 5 iterations of the algorithm. This was done to minimize the impact of Monte Carlo error on the final parameter estimates. A stricter stopping rule or a greater number of iterations to average over could have been selected but for the purpose of demonstrating the accuracy achieved with a relatively small number of iterations and reasonable computational time, the above rules were used for all simulated data sets.

To examine sensitivity of the proposed SEM algorithm to initial starting values, we fit the model to the same 100 data sets (with a 50-day expected inter-observation time) using a different, more extreme set of starting values. Since estimation under the minimal path assumption was extremely computationally difficult (with run times frequently exceeding 250 hours), time did not permit the estimation of more than 100 data sets for comparison across all three estimation approaches. The SEM algorithm was less computationally burdensome to implement and we therefore chose to simulate and fit an additional 100 data sets using just the SEM approach for the purpose of demonstrating the algorithm's performance across a larger sample. The additional panel data sets were simulated using the same true parameter values as displayed in Table 4.1 and a 50-day expected inter-observation time.

To efficiently conduct the rejection sampling and iteration required for the SEM algorithm, we used the R package *Rcpp* which integrates R and C++ to improve performance. To perform the maximization step of the SEM algorithm we once again used the R *constrOptim* function to execute linearly constrained optimization with the same constraints as noted above. All components of this simulation study used computational and storage services associated with the Hoffman2 Shared Cluster provided by the UCLA Institute for Digital Research and Education's Research Technology Group.

4.7 A Simulation Study: Results

The impact of sparseness of the observation process used to generate the panel data on parameter estimates obtained when using the minimal path estimation approach are displayed in Table 4.1. An expected inter-observation time of 0 days corresponds to continuously observed paths and, as anticipated, resulting parameter estimates appear to be unbiased on average with little difference between the parameter estimate means and medians taken across simulated data sets. When increasing the expected inter-observation time to 50 days and thereby increasing sparseness, parameter estimates appear to exhibit moderate bias. Specifically, the median estimated probability of transitioning backwards from state 2 to state 1, given occupancy of state 1, is 0.13 as compared to the true value of 0.30. This result suggests that a portion of the back transitions occurring in the complete simulated data are not identified when the panel observation process is exponential with a 50-day expected inter-observation time. Other parameters impacted by the minimal path assumption include the probability of transition from state 2 to state 3 (median = 0.68, true value = 0.55), the shape parameter for transition from state 2 to state 1 (median = 2.05, true value = 1.25), the shape parameter for transition from state 2 to state 4 (median = 2.47, true value = 1.75), the scale parameter for transition from state 1 to state 2 (median = 74, true value = 60), and the scale parameter for transition from state 2 to state 1 (median = 108, true value = 70).

Increasing the sparseness by increasing the expected inter-observation time to 100 days further increased the bias in parameter estimates. Under this more extreme sparseness, the median estimated probability of transition backwards from state 2 to state 1 is 0.09, as compared to 0.13 for the less sparse observation process, and 0.30 when using the complete data (true value = 0.30). The other parameters previously noted as being impacted by the less sparse observation process were impacted to an even greater extent under the more sparse observation process with the difference between the true value and the median estimate being in the same direction and even larger in magnitude. There also appears to be greater skew in the distribution of parameter estimates across simulated data sets when the observation process is more sparse with the mean estimate greatly exceeding the median in some instances. For instance, in the more sparse setting, the mean estimates for the shape parameters for transitions from state 2 to 1 and state 2 to 4 are 28.4 and 25.9 while the corresponding median values are only 2.6 and 3.8, respectively. This is particularly true for several of the shape parameters that experienced occasional instability during estimation among data sets for which the observation process resulted in a relatively small or anomalous sample of transitions on which estimation was based.

Table 4.1: Mean and median minimal path parameter estimates across different expected inter-observation times for 100 simulated data sets generated using the parameter values displayed in the *True Value* column.

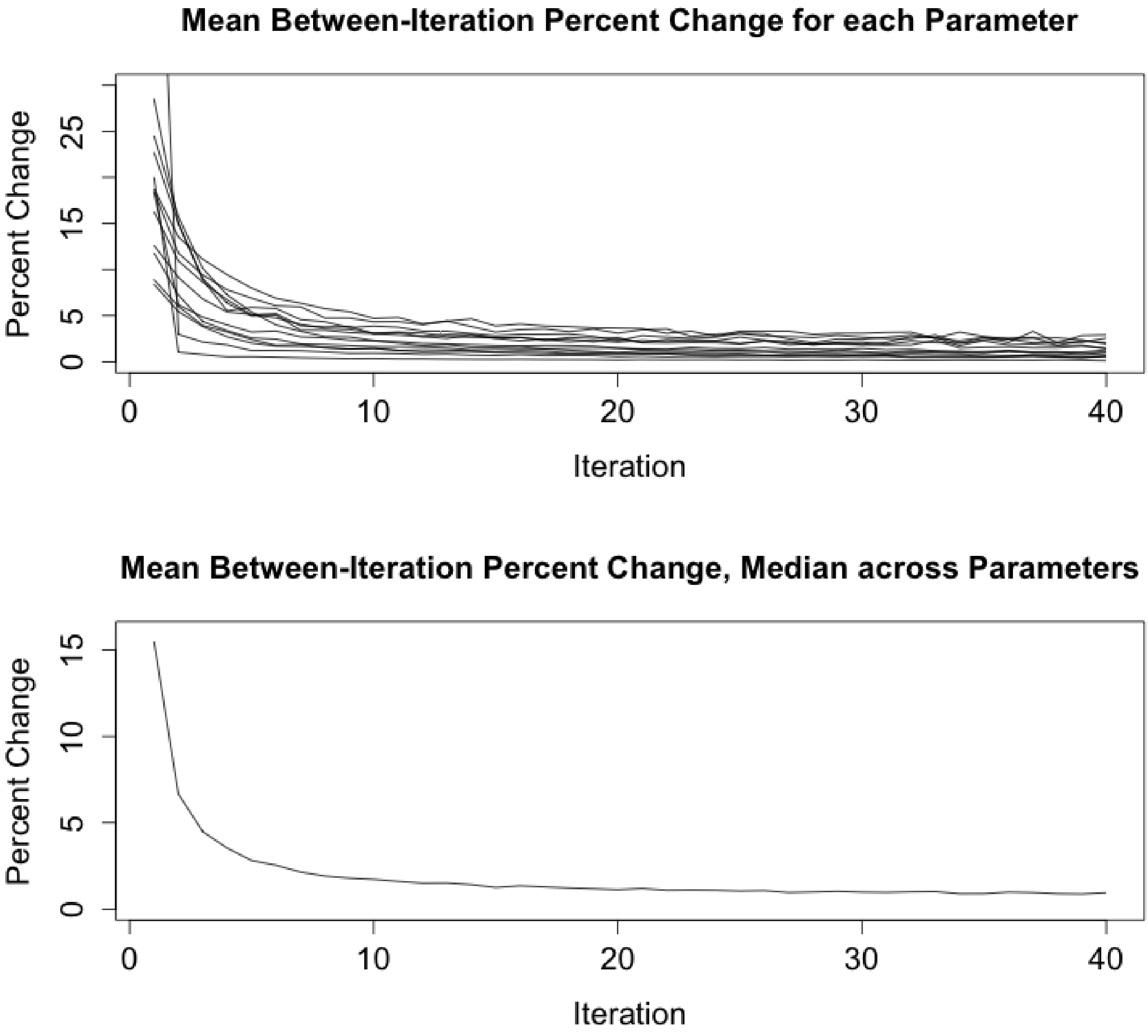
Expected Inter-Observation Time:		**0 Days**		**50 Days**		**100 Days**	
	True	**Estimate**		**Estimate**		**Estimate**	
Parameter	**Value**	Mean	Median	Mean	Median	Mean	Median
P_{12}	0.90	0.90	0.90	0.87	0.87	0.85	0.86
P_{21}	0.30	0.30	0.30	0.14	0.13	0.10	0.09
P_{23}	0.55	0.55	0.56	0.69	0.68	0.76	0.75
ν_{12}	0.75	0.76	0.76	0.80	0.80	0.76	0.75
ν_{14}	1.50	1.70	1.67	1.87	1.77	1.98	1.74
ν_{21}	1.25	1.25	1.23	6.24	2.05	28.39	2.57
ν_{23}	2.00	2.02	2.01	1.86	1.88	1.80	1.70
ν_{24}	1.75	1.93	1.88	2.88	2.47	25.91	3.81
σ_{12}	60	60	60	75	74	77	77
σ_{14}	150	156	155	156	145	161	148
σ_{21}	70	72	72	121	108	140	118
σ_{23}	80	81	80	88	85	97	93
σ_{24}	200	197	196	218	221	199	196

Table 4.2: Comparison of bias and accuracy of parameter estimates across different estimation approaches as applied to 100 simulated data sets generated using the parameter values displayed in the *True Value* column.

		Complete Data				**Minimal Path**				**SEM**			
	True	Mean	SD	%	MSE	Mean	SD	%	MSE	Mean	SD	%	MSE
	Value			Bias	$\times 10^2$			Bias	$\times 10^2$			Bias	$\times 10^2$
P_{12}	0.90	0.90	0.02	0.1	0.0	0.87	0.03	3.2	2.6	0.90	0.03	0.2	0.0
P_{21}	0.30	0.30	0.03	0.9	0.0	0.14	0.04	52.9	58.2	0.31	0.06	1.7	0.0
P_{23}	0.55	0.55	0.04	0.8	0.1	0.69	0.05	24.7	38.1	0.55	0.05	0.2	0.0
ν_{12}	0.75	0.76	0.04	1.6	0.4	0.80	0.07	6.4	3.1	0.78	0.08	3.5	0.9
ν_{14}	1.50	1.70	0.40	13.6	10.5	1.87	0.64	24.5	21.1	2.07	1.12	37.7	28.6
ν_{21}	1.25	1.25	0.16	0.0	0.0	6.24	39.57	398.9	62.8	1.44	0.67	15.5	5.6
ν_{23}	2.00	2.02	0.14	0.9	0.2	1.86	0.32	7.2	6.6	2.05	0.33	2.6	0.8
ν_{24}	1.75	1.93	0.46	10.3	7.1	2.88	1.70	64.6	75.3	2.39	1.54	36.8	26.9
σ_{12}	60	60	6	0.6	2.4	75	9	24.8	2497.4	60	8	0.7	2.2
σ_{14}	150	156	36	4.1	106.0	156	42	4.3	98.0	155	42	3.6	70.7
σ_{21}	70	72	10	3.4	54.5	121	49	72.8	5271.7	76	24	7.9	126.7
σ_{23}	80	81	4	0.6	7.0	88	10	9.6	579.5	80	6	0.3	1.1
σ_{24}	200	197	28	1.4	26.6	218	51	8.8	600.8	194	44	3.1	89.2

A comparison of the three different estimation approaches appears in Table 4.2. The complete data estimates presented in Table 4.2 are equivalent to the 0 day expected inter-observation time estimates displayed in Table 4.1. As indicated in Table 4.1, maximum likelihood estimation using the complete data typically resulted in unbiased parameter estimates. Among these estimates, the largest observed percent bias occurred for the Weibull shape parameters for transitions from state 1 to state 4 and from state 2 to state 4 (13.6% and 10.3%, respectively). This finding likely highlights the sensitivity of shape parameters to the small number of transitions to state 4 that occur in the data. This bias could also be attributable to the limited length of the interval of observation (300 days) since each individual trajectory began in state 1 and may not have had the burn-in time necessary to obtain time-homogeneity, especially for some low probability terminal events such as transition to state 4. In applying the minimal path estimation approach, we identify substantially increased bias across almost all parameters (Table 4.2). As mentioned previously, the minimal path approach drastically underestimates the probability of a back transition from state 2 to state 1 occurring. The minimal path approach also substantially overestimates the shape and scale parameters associated with back transition from state 2 to state 1, and the shape parameter associated with transition from state 2 to state 4 (greater than 60% bias).

In implementing the SEM algorithm, the specified stopping rule resulted in an average of 28.0 iterations across all 200 simulated data sets with a minimum of 11 and maximum of 52 iterations. The absolute percent change in parameter estimates from one iteration to the next was calculated and the mean percent change across all simulated data sets is displayed in Figure 4.3. As the iterations increase, the displayed means are taken over a smaller number of data sets since data sets required a differing number of iterations before satisfying the stopping criterion. Within the first 5-10 iterations the SEM algorithm demonstrated rapid movement toward the true underlying value for all of the parameters. During later iterations, parameter estimates are shown to stabilize and random variations appear to be primarily attributable to Monte Carlo error, suggesting convergence. Figure 4.4 depicts the distribution of stopping times represented by both number of iterations and mean hours of run time. For the majority of the simulated data sets, the SEM algorithm stopped at between


Mean Between-Iteration Percent Change for each Parameter

Mean Between-Iteration Percent Change, Median across Parameters

Figure 4.3: Convergence of the SEM algorithm based on 200 simulated data sets. Plots display the mean percent change in parameter estimates from one iteration to the next among the 200 data sets. The upper plot displays a line for each of the 13 parameters and the lower plot displays a single line representing the median across all 13 parameters.

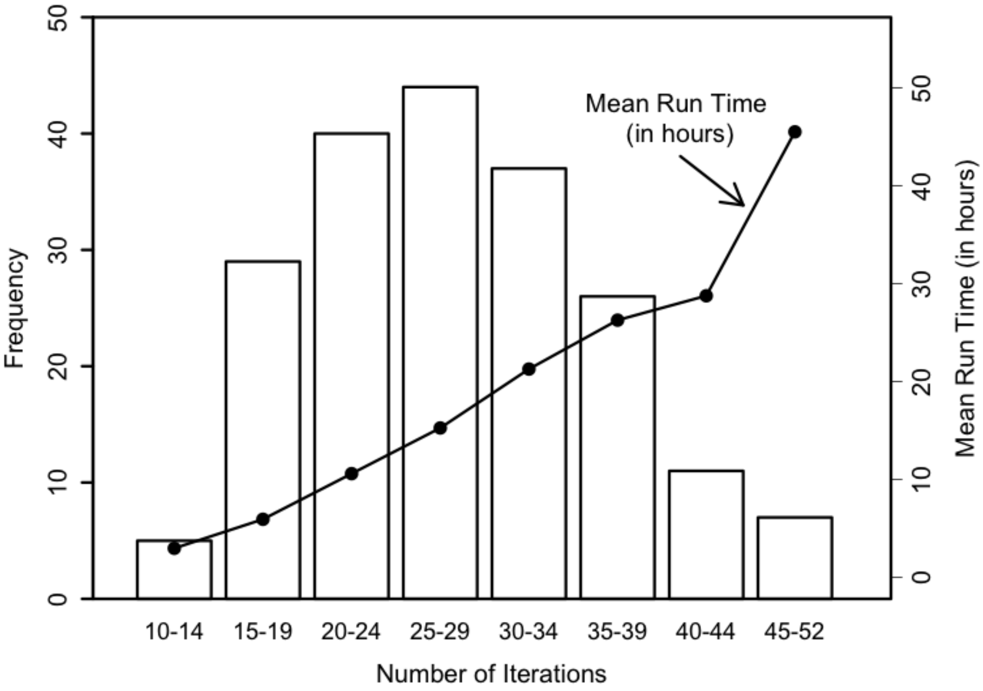


Figure 4.4: Computational requirements of the SEM algorithm among 200 simulated data sets. Vertical bars depict the distribution of the number of iterations required before the stopping criterion was satisfied and the dots connected by a line depict the mean number of hours required for each categorical range of iterations.

20-34 iterations. Among these 121 data sets, the mean run time was 15.6 hours, representing a dramatic improvement over the run time required for minimal path estimation.

As shown in Table 4.2, the SEM parameter estimates demonstrated greatly reduced bias relative to the minimal path estimates and, in many instances, closely approximated the performance of the estimates obtained using the complete data. Transition probabilities, including the probability of back transition, were accurately estimated using the SEM algorithm with standard deviation across data sets only slightly larger when using the SEM estimation approach relative to the complete data estimation approach. Low mean percent bias was also observed for each of the scale parameters (less than 8%) along with moderate increases in standard deviation relative to the complete data estimates. The only parameters for which the mean percent bias remained noticeably high were the shape parameters for

Table 4.3: Statistics describing the distribution of parameter estimates across 200 simulated data sets generated using the parameter values displayed in the *True Value* column and fit using the SEM approach with rejection sampling

	True				**90th**		
Parameter	**Value**	**Mean**	**SD**	**Median**	**Percentile**	**Min**	**Max**
P_{12}	0.90	0.90	0.03	0.90	0.94	0.78	0.97
P_{21}	0.30	0.31	0.06	0.30	0.38	0.13	0.47
P_{23}	0.55	0.55	0.05	0.55	0.61	0.43	0.72
ν_{12}	0.75	0.77	0.09	0.76	0.88	0.59	1.02
ν_{14}	1.50	2.08	1.41	1.80	3.14	0.74	15.69
ν_{21}	1.25	1.42	0.63	1.33	2.23	0.46	5.12
ν_{23}	2.00	2.05	0.38	2.03	2.53	1.19	3.25
ν_{24}	1.75	2.52	2.24	1.94	3.77	0.53	19.35
σ_{12}	60	60	8	60	69	42	94
σ_{14}	150	157	48	147	207	71	346
σ_{21}	70	74	25	69	108	27	177
σ_{23}	80	80	8	79	89	63	125
σ_{24}	200	190	46	188	250	87	368

transitions from state 1 to state 4 and from state 2 to state 4. As mentioned when discussing the complete data results, the bias observed in these estimates may signify the sensitivity of these shape parameters to small transition sample sizes and the limited length of the interval of observation relative to the necessary burn-in time.

To further investigate the source of the mean bias presented in Table 4.2, for the larger sample of 200 data sets fit using the SEM estimation approach we calculated descriptive statistics for the parameter estimates (Table 4.3). For the majority of the parameters, the mean and median taken across the simulated data sets did not differ substantially and the other distributional statistics did not seem to suggest major skew. However, for the two parameters we identified as exhibiting noticeable bias in Table 4.2 (ν_{14} and ν_{24}), the mean and median across the 200 data sets suggested a skewed distribution. In fact, for the shape parameter for transition from state 1 to 4, the mean estimate equals 2.08 and the associated median is only 1.80, much closer to the true value of 1.50. A similar result holds for the shape parameter for transition from state 2 to 4. When examining the 90th percentile relative to the maximum parameter estimate across the 200 data sets, it is clear that the bias observed in Table 4.2 is in part attributable to a small number of extreme shape parameter estimates. After further examination, we determined that these extreme estimates were obtained for a handful of data sets that appear to have had relatively few transitions to state 4 based on the observation series and assuming the minimal path. In general, when applying the minimal path assumption to our panel data sets, if the count of transitions to state 4 was less than 20, the probability of obtaining biased estimates of the associated shape parameters when using the SEM approach was noticeably increased. Since the shape parameter estimates are bounded below by zero, the impact of these extremely high parameter estimates on the mean taken across data sets is not easily mitigated by increasing the number of data sets or samples.

The impact of implementing the SEM algorithm using different starting values was minimal. Final estimates obtained when conducting this sensitivity analysis did not appear to differ from the estimates presented in Table 4.2. The impact of the more extreme starting values was typically completely mitigated by the 3rd or 4th iteration.

4.8 A Simulation Study: Discussion

In this chapter we have presented a simulation-based iterative algorithm that can be used to estimate intermittently-observed semi-Markov multistate models with back transitions. The algorithm provides an attractive alternative to minimal path estimation, which is a commonly-used naive approach that assumes no unobserved back transitions. The minimal path approach produces biased parameter estimates and can easily become prohibitively computationally intense. In contrast, our SEM procedure that replaces the intractable likelihood function with a simulated approximation and updates parameter values at each successive iteration, produces unbiased results after relatively few iterations. The rejection sampling scheme used to sample from the target conditional distribution allowed for incorporation of a high proportion of the observed data which enabled the algorithm to efficiently step from one iteration to the next in the direction of the true parameter values. The method presented here differs from recently proposed alternatives in that it does not require abandonment of the assumed underlying semi-Markov distribution in favor of a Markov or phase-type sojourn distribution.

Importantly, our proposed SEM procedure demonstrated unbiased estimation of the probability of back transition as opposed to the naive approach that drastically underestimated this transition probability parameter. Being able to accurately model back transitions is of importance for researchers interested in identifying protective factors for return from illness to health or risk factors for disease relapse.

Across all three estimation approaches compared in this paper, estimates of the Weibulll scale parameters associated with transitions to death exhibited some amount of bias. Several possible explanations have been given previously. Wei and Kryscio used simulation studies to determine that semi-Markov model parameter estimates were sensitive to the sample size due to the likelihood of observing few transitions [60]. The persistent bias of the Weibull parameter estimates associated with transitions to death in the simulation study results presented herein could be attributed to the low number of transitions observed.

As demonstrated, the proposed algorithm enables accurate estimation of the rate of back transitions through *recovery* of unobserved transitions. As is the case when implementing all missing data methods, the ability to recover unobserved transitions is still limited by the extent of the incompleteness of the data. Consider a disease process with high rates of both forward and backward transition between healthy and diseased states. If this process were combined with a very infrequent observation process for all individuals sampled, the proposed algorithm would inevitably underestimate the rates of transition in both directions. This is a limitation inherent to the minimal amount of information available and would thus be shared by any alternative algorithm attempting to capture unobserved transitions.

The SEM algorithm allows us to borrow information across the entire sample to effectively impute the complete trajectory for each individual based on a limited set of observations. The imputed trajectories are expected to increasingly reflect the true underlying process with each successive iteration of the algorithm until convergence is reached. An advantage of the approach presented herein is that it utilizes the available data in its entirety, thereby incorporating the maximum amount of information when estimating model parameters.

In this chapter, we presented an iterative estimation algorithm that makes use of a simulation-based approximation to overcome the intractability of the likelihood function. We have also described implementation of this algorithm using a rejection sampling method that has the benefit of incorporating the full information at each iteration. In a simulation study, we demonstrated the feasibility and performance of the proposed procedure relative to a naive approach. We determined that in estimating intermittently-observed semi-Markov models, the proposed approach allows for accurate estimation of model parameters and recovery of unobserved back transitions. The proposed method allows researchers interested in modeling recovery from illness and identifying factors impacting the rate and probability of recovery to construct and estimate a semi-Markov model. This result is important because, in many applications, semi-Markov models are more appropriate when considering the underlying disease process, relative to Markov or other alternative models.

We conclude that the proposed estimation approach which relies on the SEM algorithm with rejection sampling is a useful statistical method for obtaining unbiased parameter estimates for semi-Markov disease progression models with back transitions. In the next chapter, we will demonstrate implementation of the proposed estimation approach by fitting a model for dementia onset using cognitive data collected as part of a prospective research study on aging and Alzheimer's disease.

CHAPTER 5

Application to the Semi-Markov Modeling of Dementia Onset

5.1 The Nun Study: Background and Model Specification

Beginning in 1991, the Nun Study began enrolling members of the School Sisters of Notre Dame who were born prior to 1917 and resided in retirement communities in the midwestern, eastern, and southern United States. Data were available for 672 participants who were recruited in phases and received annual cognitive assessments [74, 75, 76]. The study was motivated by scientific interest in examining the onset of dementia in relation to measurable risk factors. For the purpose of this application, cognitive assessment results for each participant were categorized into one of three states: intact cognition, impaired cognition, and dementia. At each visit, dementia was assessed using formal diagnostic criteria. Among participants not meeting the formal criteria for dementia, those who failed one or more of a battery of cognitive and Activities of Daily Living tests (including the Mini-Mental State Exam (MMSE), Boston Naming, Verbal Fluency, and Constructional Praxis tests) were classified as having impaired cognition. Participants who passed all cognitive and Activities of Daily Living tests were classified as having intact cognition. Information regarding date of death was also available for participants who experienced death during the the observation period.

Our analysis is restricted to the 544 participants who did not meet eligibility for dementia at the time of enrollment (165 with intact cognition and 379 with impaired cognition at the time of enrollment). Age in years at the time of enrollment ranged from 75.4 to 90.3 (mean = 79.9) for those participants with intact cognition and ranged from 75.4 to 99.4 (mean =

83.3) for those with impaired cognition. The structure of the multistate model applied to this data is displayed in Figure 4.1. Intact cognition and impaired cognition are considered transient states with back transitions from impaired cognition to intact cognition possible. Dementia and death are considered absorbing states with death functioning as a competing risk to dementia. The model structure does not allow for transitions directly from intact cognition to dementia without passage through the intermediate state of impaired cognition. When considering all 544 participant paths, at last follow-up we find that 6% ended in a state of intact cognition, 9% ended in a state of impaired cognition, 30% ended in dementia, and 55% ended in death. When using the observed panel data to construct the minimal path for each participant, at least one back transition from impaired to intact cognition was observed for 27% of participants. Among these minimal paths, 69% consisted of one back transition, 26% consisted of two back transitions, and 5% consisted of 3 back transitions. Exact times of transitions were not available for transitions between states of intact cognition, impaired cognition, and dementia but the exact time of death was observed. Although annual visits were proposed in the study protocol, actual visits were not evenly spaced and varied across participants. An average of 5.8 observations were available for each participant with a range of 2 to 12 and the mean and median time between observations was 1.4 years. All transitions between the states of intact cognition, impaired cognition, dementia, and death were modeled assuming Weibull-distributed sojourn times based on the parametrization indicated in (4.3).

One obstacle that has been the focus of much attention in the semi-Markov modeling literature is left censoring of the sojourn times associated with the first state occupied by each individual in the sample. At the time of entry into the Nun Study, each participant's age and current state were recorded but no information regarding the amount of time having already elapsed in the current state was collected. When sojourn times are exponentially distributed, implying a Markov model, the time elapsed is irrelevant. However, when constructing a semi-Markov model, one of a variety of approaches for assigning a value to the unknown elapsed length of time is necessary. These sojourn times are left censored because they are known to lie below a value equal to the individual's age at entry into the study. In constructing a semi-Markov model for stroke onset, Kapetanakis et al. selected 40 as the age in years at which all subjects were healthy and then used an EM-inspired algorithm to impute the instant of transition from health to illness among those individuals who had experienced a stoke prior to baseline [46]. In modeling functional status transitions over time, a related approach was taken by Cai et al. who used an analogue to the SEM algorithm to simulate a cohort of 55-year-old subjects from which imputed values of the elapsed sojourn times at entry into the study were drawn [58]. We choose to use a similar approach which is easily implemented within the broader SEM estimation framework proposed. Since no member of the Nun Study sample enrolled at younger than 75 years of age, we assume all participants were in a state of intact cognition at age 70. We also assume that these participants entered into the state that we are referring to as *intact cognition* at exactly age 70 which implies that the intact cognition and impaired cognition states are defined for the purpose of this study only among those individuals 70 years of age or older. This is the same approach as used by Kryscio et al. and is not without limitations when applied to a model with back transitions [77]. The age of 70 was chosen to balance the risk associated with assuming an individual with impaired cognition at age 70 was cognitively intact, with the negative impact that extrapolating outside the range of observed data could have on the resulting estimates.

In applying the SEM estimation approach to the Nun Study, the proposed rejection sampling method was adapted to accommodate the known date of death among study participants. It would have been prohibitively time consuming to continue sampling continuous trajectories until obtaining one which includes a death occurring on the exact date a death was recorded. Thus, we selected a two month acceptance window centered around the date of death within which the simulated death event was allowed to have occurred without rejection of the sampled trajectory. In applying the SEM algorithm to the Nun Study data we decided to implement the stopping criterion expressed in formula (4.4) with $\delta_1 = 0.001$ and $\delta_2 = 0.05$. The SEM algorithm was halted after (4.4) was satisfied for 5 consecutive iterations and the final estimates were taken as the average over the estimates generated during the last 5 iterations.

Estimation of standard errors used to calculate 95% confidence intervals was completed using a nonparametric bootstrap approach for multistate processes [50, 51]. The approach required sampling with replacement a total of 544 individuals from the Nun Study sample and using all of each sampled individual's data to generate parameter estimates using the SEM method. This resampling process was completed until 200 bootstrap samples had been drawn and estimates computed. The variance for each parameter was then calculated as the empirical variance of the 200 replicates.

5.2 The Nun Study: Results

The stopping criterion was met after the 11th iteration. Parameter estimates obtained using the SEM approach are displayed in Table 5.1. For an elderly individual in a state of intact cognition, there is an estimated 0.94 probability that the next state visited will be impaired cognition and only a 0.06 probability that the individual will transition directly to death. For an elderly individual in a state of impaired cognition, there is an estimated 0.58 probability that the next transition that occurs will be back to intact cognition, 0.16 probability that the next transition will be forward to dementia, and a 0.26 probability that the individual will transition directly to death without experiencing a back transition or entering a state of dementia.

Estimated Weibull sojourn time probability distributions for each of the 5 possible transitions are plotted in Figure 5.1. For an elderly individual in a state of intact cognition, the median time to transition from intact cognition to impaired cognition is 3.62 years. The shape parameter estimate of 1.35 for this transition indicates that, given an individual's next transition will be a back transition to impaired cognition, the hazard of transition is increasing as a function of time spent in a state of intact cognition. The shape parameter estimate of 0.99 for the transition from intact cognition directly to death indicates that the hazard of transition directly to death remains fairly constant as a function of time spent in a state of intact cognition. The median time to transition from intact cognition to death is 6.15 years.

Table 5.1: Parameter estimates and 95% bootstrapped confidence intervals for the Nun Study model fit using the proposed SEM approach with rejection sampling

Parameter	**Estimate**	**95% CI**
P_{12}	0.94	(0.93, 0.96)
P_{21}	0.58	(0.53, 0.64)
P_{23}	0.16	(0.13, 0.19)
ν_{12}	1.35	(1.23, 1.47)
ν_{14}	0.99	(0.75, 1.24)
ν_{21}	1.44	(1.24, 1.64)
ν_{23}	1.68	(1.36, 1.99)
ν_{24}	1.34	(1.11, 1.56)
σ_{12}	4.74	(3.92, 5.57)
σ_{14}	8.89	(5.31, 12.47)
σ_{21}	2.47	(2.11, 2.82)
σ_{23}	9.98	(8.72, 11.23)
σ_{24}	9.79	(8.45, 11.13)

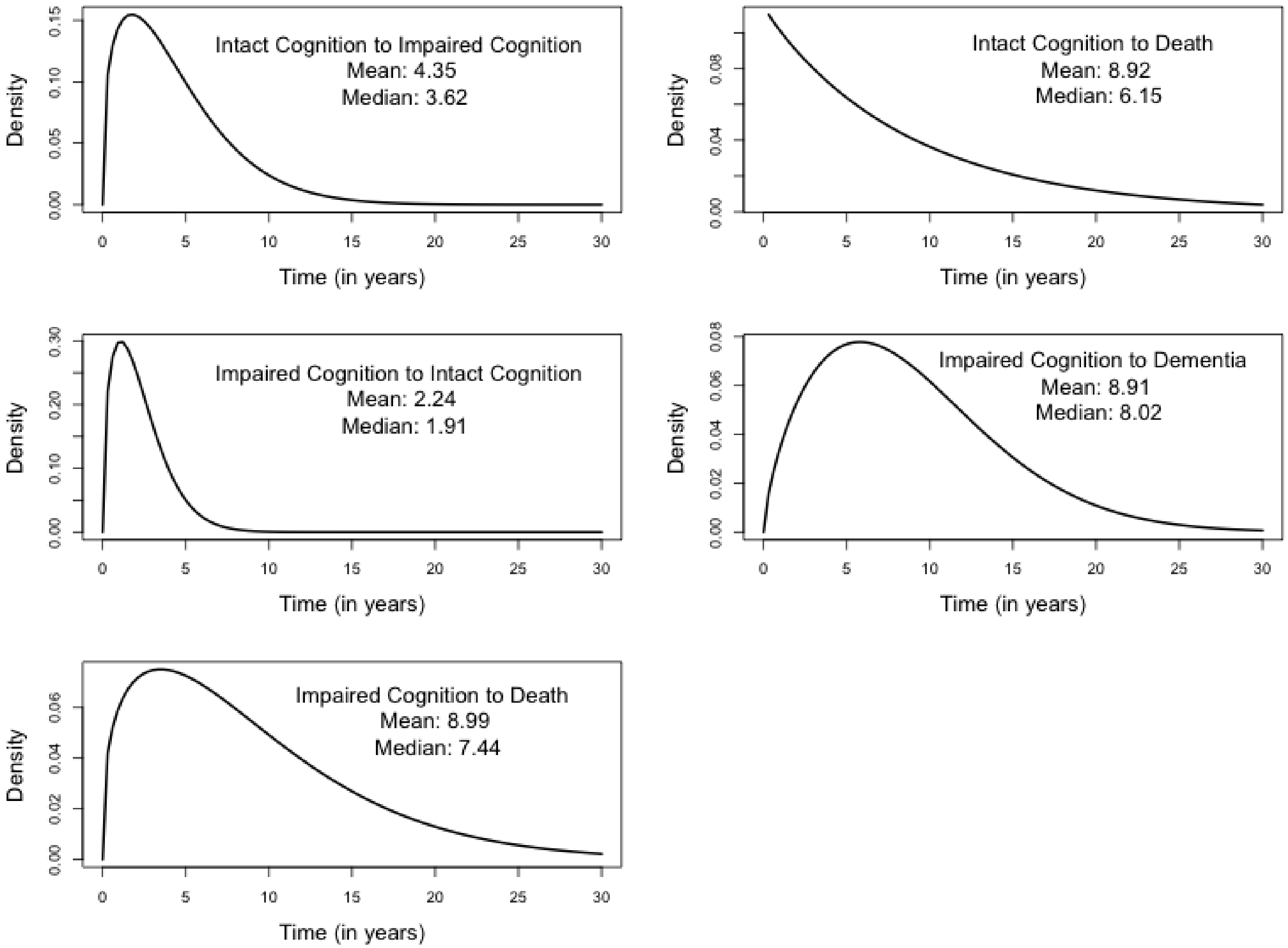


Figure 5.1: Sojourn time probability density functions based on the Weibull parameter estimates obtained from the Nun Study model fit using the proposed SEM approach with rejection sampling

When back transitions occur (from impaired to intact cognition) among this elderly population, they occur relatively quickly (median time: 1.91 years). The shape parameter associated with these back transitions indicates moderately increasing hazard as a function of time spent in the state of impaired cognition. The sojourn time distributions for transitions from impaired cognition to dementia and from impaired cognition to death were similar with median times to transition of 8.02 and 7.44 years, respectively. Shape parameters for these transitions from impaired cognition to dementia and impaired cognition to death both indicated an increasing hazard with increased time elapsed in a state of impaired cognition.

5.3 The Nun Study: Discussion

Our analysis of the Nun Study showed a high probability of back transition from impaired to intact cognition with a 1.9-year median time to recovery of intact cognition. Rates of transition from impaired cognition to dementia and death were shown to increase as a function of time spent in a state of impaired cognition. Previous studies have claimed that the likelihood of transition from impaired back to intact cognition is highest during the time immediately following transition to an impaired state. Our analysis of the Nun Study determined that although the majority of back transitions that occur do so within the first couple years following the transition to an impaired state, the rate of back transition is actually increasing with duration in a state of impaired cognition. Although rates of transition from impaired cognition to dementia and death were also increasing a function of duration, the median length of time until transition to these absorbing states was within the range of 7-9 years.

It is important to emphasize that the Nun Study represents a very specific population comprised of members of the School Sisters of Notre Dame, and thus may not be representative of the general population of elderly persons. The limitations and issues of generalizability of the Nun Study have been previously discussed in a number of publications [60, 75].

There are a number of extensions that could be considered. In our Nun Study example, it would be useful and important to incorporate covariates such as age and genetic markers to assess the impact of such characteristics on transition rates and probabilities in modeling of dementia onset [60, 76]. The present model also does not consider misclassification of states. As in the Nun Study, when subjects are observed to frequently transition back and forth between states such as intact and impaired cognition, misclassification should be considered. If we consider an extreme example, in which the neurodegeneration giving rise to impaired cognition is irreversible meaning that true back transitions are not biologically feasible, the model presented herein would alternatively provide a framework for identifying diagnostic misclassification. Kang and Lagakos demonstrate an approach that can be implemented to account for misclassification when the error probabilities are assumed to satisfy a conditional independence assumption [57]. Building off of this approach, we implement an extension to the present SEM methodology that addresses misclassification in the Chapter 6.

Although we have assumed independence of the random observation process and the underlying multistate process in the materials presented in this paper, there are many instances in which this assumption may be violated. For example, in models of disease progression an individual experiencing symptoms associated with a more severe disease state might seek care at a hospital or clinic resulting in an observation, whereas a healthy individual may be observed only at routine scheduled doctor's appointments. The importance of modeling these disease-driven observations has been recently addressed by Lange et al. and an analog of their proposed joint modeling approach could potentially be incorporated into the estimation procedure described in this paper [45].

As noted by Kryscio and Abner, use of cognitive panel data to model the flow of elderly patients from intact cognition through dementia presents several challenges [78]. As with other cognitive panel studies, the Nun Study enrolled a sample of elderly volunteers which may give rise to missing data or selection bias issues. Since seriously impaired individuals were not likely to participate in the Nun Study, a healthy cohort effect must be considered. Another limitation of the Nun Study results presented herein are that the fitted model does not take age into account. The present model incorporates aging only through the modeling of time-varying hazard rates within a given state. Although the Nun Study data consists largely of observations from individuals of extremely advanced age, it may still be inappropriate, for instance, to assume the rate of transition for an 80-year-old with intact cognition equals that of an 85-year-old with intact cognition. An alternative model could use age at entry into each state as a time-independent covariate in formulating a model for the transition probabilities or sojourn time distributions.

As noted in Section 6.1, there are several options available for imputing the left censored sojourn times for each participant at entry into the study. The option we have chosen is simple to implement but relies on several assumptions that may reduce our ability to generalize the results of our final model. Alternative methods for addressing left censoring have been proposed and should be considered in future applications of the proposed model [58, 46, 79].

CHAPTER 6

A Semi-Markov Model for Disease Progression with Misclassification

6.1 Challenges in Estimating Semi-Markov Models with Misclassification

As discussed in the previous chapter, construction of multistate models that allow for back transitions and rates of transition that vary depending on time elapsed in the current state are of great utility in the field of disease progression research. Methods for the estimation of such models using intermittent observations of the disease process were also presented previously. In practice, continuous observation of the disease process is exceedingly rare and intermittent observations corresponding to patient visits or survey administration waves are far more common. The stochastic method we presented enables estimation of these highly flexible and informative multistate models across a wide range of disease progression applications in which panel data are readily available.

The strengths of this method are apparent, but the potential limitations of such an approach that uses relatively minimal information to estimate the dynamics of a complex process allowing for back transitions need to be carefully considered. Of particular concern, the proposed method relies heavily on the accuracy of the intermittent observations occurring at discrete points in time. In the material presented previously, we made the assumption of no measurement error in the classification of individuals in states of health or illness. The rate of misclassification of individuals in certain states corresponding to severe illness or death may be negligible and can justifiably be ignored when fitting a multistate model.

However, the rate of misclassification is known to be non-negligible in many instances of disease progression modeling. In fact, estimating and evaluating the impact of misclassification rates, often referred to as false positive and false negative rates, continues to be a topic of great interest to epidemiologists and other health scientists. Typically, a false positive refers to the instance in which a healthy individual is assessed and determined to be in a state of illness. Alternatively, a false negative refers to the instance in which an individual in a state of illness is misclassified as being in a healthy state. Accurate estimation of these rates is important in that they are often used to inform health policy decision-making and develop effective screening and treatment guidelines for health practitioners. False positive and negative rates for disease classification instruments are preferably estimated by comparison to a gold standard instrument that can definitively discriminate between the presence and absence of disease. Unfortunately, for many diseases a gold standard either doesn't exist or is not practical for use due to barriers such as expense or patient burden [30].

When the assumed model structure is progressive, such that no state can be transitioned to more than once, misclassification has been dealt with by assuming that each observation implying a back transition is either misclassified or that a misclassification event occurred at a previous observation in such a way as to eliminate the apparent back transition [80, 81]. However, when the assumed model allows for back transitions, non-negligible misclassification rates associated with the transient states can result in a series of observations that imply an artificially higher or lower number of back transitions relative to the true underlying process. For instance, if an individual remains in a state of health across a ten year period and is observed approximately once a year, a single misclassification event occurring at the fifth observation would imply an artificial back transition and would drastically reduce our estimate of the sojourn time spent in a state of health for this individual. When fitting a semi-Markov model to a sample of individuals observed intermittently, even rare misclassification events can impact our perception of the underlying process dramatically. As a first step in examining the limitations of the previously proposed SEM algorithm, we aim to describe the bias in parameter estimates that is introduced when such a method is implemented using data with a non-negligible misclassification rate. Following this description, we aim to propose an extension to the SEM algorithm adopted in the previous chapter that mitigates the bias introduced by potential misclassification.

To address possible misclassification of observed states in multistate modeling, hidden Markov models have been successfully used in instances when the Markov assumption of exponentially distributed sojourn times is appropriate [82, 83, 84]. A hidden Markov model is used to model an underlying, unobserved Markov process given the availability of a series of observations that relate probabilistically to the true states occupied by the hidden process. Far less frequently, hidden semi-Markov models have been attempted in which the Markov assumption is relaxed to allow for alternative sojourn time distributions. Difficulties with estimation, including identifiability concerns, have hindered the wide-spread use of hidden semi-Markov models. Hidden semi-Markov models have not been extensively implemented in the health research setting and have experienced only limited application in fields such as speech recognition and predictive maintenance for engineering systems [85, 86]. Use of hidden semi-Markov models in areas such as manufacturing and operations research may be enabled by the availability of complete sojourn data arising from continuous observation of sampled items such as engineering system components. Since intermittent observations are far more common when studying disease progression among human subjects, methods for fitting hidden semi-Markov models that can accommodate panel data need to be developed before such models can be considered useful in the health research setting and applied to problems of misclassification.

An optimal method for disease progression modeling in the presence of back transitions and misclassification would allow for simultaneous unbiased estimation of the semi-Markov model parameters associated with transition rates and probabilities, and estimation of the true underlying misclassification rate without requiring information from a gold standard. In the remaining sections of this chapter, we will present an extension to the SEM algorithm described in a previous chapter that can be used to address misclassification. We will describe implementation of the proposed method and conduct a simulation to study to evaluate performance of the proposed method relative to a method that ignores misclassification. Next, we will apply the extended version of the SEM algorithm to the Nun Study described previously to estimate the rate at which elderly participants with intact cognition were misclassified into a state of impaired cognition, also referred to as the impaired cognition false positive rate. We conclude by describing the strengths and limitations of the proposed extension and outline continuing areas of research.

6.2 Notation for a Semi-Markov Model with Misclassification

Let $Y = \{Y(t), t \geq 0\}$ denote a continuous-time semi-Markov process with S states denoted $1, \ldots, S$. Let $s_0, s_1, s_2, \ldots$ denote initial and subsequent consecutive states occupied by process Y Let d_n denote the sojourn time in the $(n-1)$th state prior to transitioning to the nth state. Thus, $Y = \{s_0, d_1, s_1, d_2, \ldots, d_n, s_n, \ldots\}$. For a process in which $s_n = i$,

$$\lambda_{ij}(d) = \lim_{\Delta d \to 0} \frac{P(d_{n+1} < d + \Delta d, s_{n+1} = j | d_{n+1} \geq d, s_n = i)}{\Delta d},$$

where d denotes time elapsed since entrance into state i for $i, j = 1, \ldots, S$ and $i \neq j$. The preferred approach for specifying these transition intensities is to set $\lambda_{ij}(d) = P_{ij} f_{ij}(d)$ where

$$\begin{aligned} P_{ij} &= P(s_n = j | s_{n-1} = j) \\ f_{ij} &= \lim_{\Delta d \to 0} \frac{P(d_n < d + \Delta d | d_n \geq d, s_n = j, s_{n-1} = i)}{\Delta d}, \end{aligned}$$

such that the trajectories are modeled separately from the times of transition. P_{ij} represents the probability that the next transition experienced by a process occupying state i will be to state j, assuming the process can be followed forward an indefinite amount of time without the possibility of censoring. The density function for the sojourn time in state i prior to transitioning to state j is represented by $f_{ij}(d)$ which allows for dependence on time elapsed in state i in accordance with the semi-Markov property. Model specification requires that $0 \leq P_{ij} \leq 1$ for all $i, j = 1, 2, \ldots, S$ and $i \neq j$ and $\sum_{i \neq j} P_{ij} = 1$ for $i, j, = 1, 2, \ldots, S$. Additionally, the assumed model structure may dictate that $P_{ij} = 1$ when j is the only state that can be transitioned to from state i, or that $P_{ij} = 0$ if transition from state i to j is not possible. We refer to i as a transient state if $P_{ij} \neq 0$ for at least one $j \in \{1, 2, \ldots, S\}$ and as an absorbing state otherwise.

Consider an intermittently-observed process with M observations at times $0 = t_0 <$

$t_1 < t_2 < \cdots < t_M$. Let $\boldsymbol{Y} = (Y_0, Y_1, \ldots, Y_M)$ where $Y_m = Y(t_m)$ such that $Y_m \in \{1, 2, \ldots, S\}$. When we allow for misclassification, rather than observing $\boldsymbol{Y}$, we observe $\boldsymbol{X} = \{X_0, X_1, X_2, \ldots, X_M\}$ where $X_m \in \{1, 2, \ldots, S\}$. Assume conditional on the entire continuous-time process Y, the distribution of X_m depends only on Y_m for all $m \in \{1, 2, \ldots, M\}$ such that

$$\begin{aligned} P(\boldsymbol{X}|Y) &= P(\boldsymbol{X}|\boldsymbol{Y}) \\ &= \prod_{m=0}^{M} P(X_m|Y_m). \end{aligned}$$

We denote the misclassification probabilities

$$\alpha_{ij} = P(X_m = j|Y_m = i).$$

The likelihood for an individual observed at M visits can then be expressed

$$\begin{aligned} L &= \sum_{\boldsymbol{Y}} P(\boldsymbol{X}|\boldsymbol{Y})P(\boldsymbol{Y}) \\ &= \sum_{\boldsymbol{Y}} P(X_0|Y_0)P(X_1|Y_1)\ldots P(X_M|Y_M)P(\boldsymbol{Y}) \\ &= \sum_{\boldsymbol{Y}} \left(\prod_{m=0}^{M} \alpha_{Y_m, X_m} \right) P(\boldsymbol{Y}), \qquad (6.1) \end{aligned}$$

where $X_m, Y_m \in \{1, 2, \ldots, S\}$ and the summation across $\boldsymbol{Y}$ is the summation over all possible paths. In instances in which the set of possible paths can be enumerated, computation of the likelihood above is straightforward. When observations are intermittent and back transitions are possible, there is no limit to the number of potential paths that must be considered resulting in an infinite sum in (1) and an intractable likelihood function.

6.3 An SEM Algorithm for Multistate Processes with Misclassification

Use of the stochastic expectation-maximization algorithm (SEM) to estimate parameters of an intermittently-observed semi-Markov process with back transitions in application to dementia onset modeling has been discussed in previous Chapters. We will briefly introduce the important notation here. The intermittent observation process combined with the assumed model structure that allows for back transitions results in an intractable likelihood function. Intractability arises because summation across all possible paths cannot be completed when the observation process is such that any number of back transitions could hypothetically have occurred between each set of consecutive observations. The SEM algorithm allows us to circumvent this intractability by using a stochastic approximation while iteratively updating parameter estimates. Thus, the SEM algorithm can be viewed as a modified version of the widely-used deterministic expectation-maximization algorithm (EM) which uses successive expectation and maximization steps to account for the uncertainty introduced by incompletely observed data.

Let $\boldsymbol{X} = \{X_0, X_1, X_2, \ldots, X_M\}$ denote the observed data and $Y = \{Y(t), 0 \leq t \leq t_M\}$ denote the unobserved data such that $(\boldsymbol{X}, Y)$ refers to the complete data. Thus, $g(\boldsymbol{X}, Y|\boldsymbol{\theta})$ represents the joint distribution of the complete data conditional on parameter vector, $\boldsymbol{\theta}$. The rth iteration of the EM algorithm consists of an E-step that involves computing

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r-1)}) = E\Big[\log\, g(\boldsymbol{X}, Y|\boldsymbol{\theta})\Big] = \int \log\, \big[g(\boldsymbol{X}, Y|\boldsymbol{\theta})\big] h(Y|\boldsymbol{X}, \boldsymbol{\theta}^{(r-1)})dY,$$

where $h(Y|\boldsymbol{X}, \boldsymbol{\theta}^{(r-1)})$ is the conditional distribution of the unobserved data given the observed data and the vector of current parameter estimates, $\boldsymbol{\theta}^{(r-1)}$. Subsequently, the M-step is completed by maximizing the Q-function to obtain the updated estimate $\boldsymbol{\theta}^{r}$ that satisfies

$$Q(\boldsymbol{\theta}^{r}|\boldsymbol{\theta}^{(r-1)}) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r-1)})$$

for all $\boldsymbol{\theta}$ in the parameter space.

The E-step involves integrating with respect to Y to account for the uncertainty inherent in the incomplete data. Since this uncertainty arises in part from the infinite number of potential paths, the E-step encounters analytical intractability. To address this issue, Wei and Tanner (1990) proposed an algorithm that stochastically approximates the integral in the E-step by taking the Monte Carlo average

$$\bar{Q}_{K_r}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r-1)}) = \frac{1}{K_r} \sum_{k=1}^{K_r} \log\, g(\boldsymbol{X}, \dot{Y}^{k}|\boldsymbol{\theta}), \tag{6.2}$$

where $\dot{Y}^1, \ldots, \dot{Y}^{K_r}$ are continuous paths sampled from $h(Y|\boldsymbol{X}, \boldsymbol{\theta}^{(r-1)})$ and K_r represents the Monte Carlo sample size that is typically increased across successive iterations [67]. Note that here we introduce the dot notation to distinguish sampled elements, such as $\dot{Y}$, from their true counterparts, Y. Adopting the approach proposed by Wei and Tanner, a Monte Carlo approximation is used in place of the true Q-function. Under mild regularity conditions, the SEM algorithm has been shown to converge with increasing K_r. The SEM algorithm is usually stopped after the criterion that the relative change in parameter estimates between two successive iterations is smaller than some pre-specified threshold is satisfied for several consecutive iterations.

6.4 Rejection Sampling of Multistate Processes with Misclassification

In implementing the SEM algorithm described previously, an approach must be specified for sampling from the conditional distribution, $h(Y|\boldsymbol{X}, \boldsymbol{\theta}^r)$. Rejection sampling from a target distribution has been effectively implemented in a previous study. Specifically, rejection sampling can be accomplished by drawing samples from a broad proposal distribution and thinning out these samples by rejecting those that are inconsistent with the observed data. Using this approach, independent, identically distributed samples from the target distribution can be generated. To sample from $h(Y|\boldsymbol{X}, \boldsymbol{\theta}^r)$, we would like to draw samples, $\dot{Y}^k$ with probability proportional to $P(Y|\boldsymbol{X}, \boldsymbol{\theta}^r)$ which can be expressed as follows

$$\begin{aligned} P(Y|\boldsymbol{X}, \boldsymbol{\theta}^r) &= \frac{P(Y, \boldsymbol{X}|\boldsymbol{\theta}^r)}{P(\boldsymbol{X}|\boldsymbol{\theta}^r)} \\ &= \frac{P(\boldsymbol{X}|Y, \boldsymbol{\theta}^r)P(Y|\boldsymbol{\theta}^r)}{P(\boldsymbol{X}|\boldsymbol{\theta}^r)} \\ &\propto P(\boldsymbol{X}|, \boldsymbol{\theta}^r)P(Y|\boldsymbol{\theta}^r). \end{aligned}$$

Expressing $P(Y|\boldsymbol{X}, \boldsymbol{\theta}^r)$ as a function proportional to $P(\boldsymbol{X}|Y, \boldsymbol{\theta}^r)P(Y|\boldsymbol{\theta}^r)$ provides justification for the following proposed rejection sampling algorithm.

1. Sample continuous path $\dot{Y} = \{\dot{Y}(t), 0 \le t \le t_M\}$ from distribution $Y|\boldsymbol{\theta}^r$.

2. For each $t_1, t_2, \ldots, t_M$, sample $\dot{X}_m = \dot{X}(t_m)$ from the distribution $X(t_m)|\dot{Y}(t_m), \boldsymbol{\theta}^r$.

3. If $\dot{X}_m = X_m \; \forall \; m \in \{1, 2, \ldots, M\}$ accept $\dot{Y}$, otherwise reject $\dot{Y}$ and repeat steps (1)-(3).

In executing the above rejection sampling algorithm we are using $P(\boldsymbol{X}|Y, \boldsymbol{\theta}^r)P(Y|\boldsymbol{\theta}^r)$ to define our proposal distribution and thinning out samples by applying knowledge available from the set of observations $\boldsymbol{X}$. An accepted $\dot{Y}$ can be used to augment the observed data resulting in pseudocomplete data and can be used in calculation of (2). The kth accepted $\dot{Y}$ in the rth iteration yields pseudocomplete data $(\boldsymbol{X}, \dot{Y}^k)$. After K_r trajectories have been accepted, the Monte Carlo average is calculated and the SEM algorithm proceeds to the subsequent M-step.

In practice, step 1 can be completed by alternating between sampling transitions from the categorical distribution with probabilities equal to P_{ij} for $j \in \{1, 2, \ldots, S\}$ and sampling state durations from f_{ij} where j represents the state previously sampled from the categorical distribution. These two sampling approaches are repeated until a continuous-time trajectory extending from time $t = 0$ to time $t > t_M$ has been obtained. Censoring is then assumed to have occurred at $t = t_M$. Step 2 can be completed for each $m = 1, 2, \ldots, M$ by sampling from the categorical distribution with probabilities equal to $\alpha^r_{\dot{Y}_m, j}$ where $j \in \{1, 2, \ldots, S\}$ and $\dot{Y}_m = \dot{Y}(t_m)$.

With the availability of the sampled continuous time process, $\dot{Y}^k$, calculation of $g(\boldsymbol{X}, \dot{Y}^k|\boldsymbol{\theta})$ is straightforward. Let $\dot{\boldsymbol{Y}}^k = \{\dot{Y}^k_0, \dot{Y}^k_1, \ldots, \dot{Y}^k_M\}$ be the vector of states occupied by sampled process $\dot{Y}^k$ at times $t_0, t_1, \ldots, t_M$ such that $\dot{Y}^k_m = \dot{Y}^k(t_m)$. For a sampled continuous time process, $\dot{Y}^k$ that occupies a sequence of $N+1$ states during the period of observation, let $\dot{s}^k_0, \dot{s}^k_1, \ldots, \dot{s}^k_N$ denote the initial and subsequent states occupied and $\dot{d}^k_n$ denote the time spent in state $\dot{s}^k_{n-1}$ prior to transition or censoring. Calculation can be completed using the following expression

$$\begin{aligned} g(\boldsymbol{X}, \dot{Y}^k|\boldsymbol{\theta}) &= P(\boldsymbol{X}|\dot{Y}^k, \boldsymbol{\theta})P(\dot{Y}^k|\boldsymbol{\theta}) \\ &= P(\boldsymbol{X}|\dot{\mathbf{Y}}^k, \boldsymbol{\theta})P(\dot{Y}^k|\boldsymbol{\theta}) \\ &= \left[\prod_{m=0}^{M} P(X_m|\dot{Y}_m^k, \boldsymbol{\theta})\right] P(\dot{Y}^k|\boldsymbol{\theta}) \\ &= \left[\prod_{m=0}^{M} \alpha_{\dot{Y}_m^k, X_m}\right] \prod_{n=1}^{N} \left[P_{\dot{s}_{(n-1)}^k, \dot{s}_n^k} f_{\dot{s}_{(n-1)}^k, \dot{s}_n^k}(\dot{d}_n^k)\right]^{1-\delta_n^k} \left[\sum_{j \neq \dot{s}_n^k} P_{\dot{s}_n^k, j} S_{\dot{s}_n^k, j}(\dot{d}_n^k)\right]^{\delta_n^k}, \end{aligned} \quad (6.3)$$

where $\delta_n^k = 1$ when the duration in state $\dot{s}_n^k$ is right censored and $\delta_n^k = 0$ otherwise.

6.5 Implementation of the SEM Algorithm for Dementia Onset Modeling

In dementia onset modeling, we assume the model structure displayed in Figure 6.2 with two transient states and two absorbing states. State 1 corresponds to the state of intact cognition, from which an individual can transition to impaired cognition (State 2) or death (State 4). From a state of impaired cognition, an individual can experience a back transition to State 1, indicating recovery of intact cognition, or the individual can transition forward to a state of either dementia (State 3) or death. Death is treated as a competing risk to dementia and we do not estimate parameters associated with transition from dementia to death. We assume a Weibull sojourn time distribution for each of the transitions such that

$$f_{ij}(d) = \left(\frac{\nu_{ij}}{\sigma_{ij}}\right)\left(\frac{d}{\sigma_{ij}}\right)^{(\nu_{ij}-1)} \exp\left(-\left(\frac{d}{\sigma_{ij}}\right)^{\nu_{ij}}\right),$$

where $\sigma_{ij} > 0$ and $\nu_{ij} > 0$ for $i = 1, 2,\ j = 1, 2, 3, 4$, and $i \neq j$. The assumed Weibull distribution also gives us the corresponding survival distribution

$$S_{ij}(d) = \exp\left(-\left(\frac{d}{\sigma_{ij}}\right)^{\nu_{ij}}\right),$$

where $\sigma_{ij} > 0$ and $\nu_{ij} > 0$ for $i = 1, 2,\ j = 1, 2, 3, 4$, and $i \neq j$. Under this framework, ν_{ij} and σ_{ij} are referred to as the shape and scale parameters, respectively. Note that $\nu_{ij} = 1$ would reduce the multistate process to a Markov process with exponentially distributed sojourn times corresponding to constant hazards of transition given occupancy of a known state. In cognitive impairment multistate modeling, we consider misclassification that occurs when an individual is in a state of intact cognition but is erroneously classified in a state of impaired cognition at a given point in time, often referred to as a *false positive*. Thus, while the notation presented previously is general enough to describe methodology relevant to multiple types of misclassification, the only misclassification rate we choose to estimate is $\alpha_{12} = P(X_m = 2|Y_m = 1)$. We assume $\alpha_{ij} = 1 \ \forall \ i = j$ and $\alpha_{ij} = 0 \ \forall \ i \neq j$ and $(i,j) \notin \{(1,2)\}$. If we model P_{ij} using $P_{ij} = \rho_{ij}$ for $i = 1, 2,\ j = 1, 2, 3, 4$, and $i \neq j$ and $P_{ij} = 0$ otherwise, we can specify the vector of parameters we are interested in estimating as

$$\boldsymbol{\theta} = \{\rho_{12}, \rho_{21}, \rho_{23}, \alpha_{12}, \nu_{12}, \nu_{14}, \nu_{21}, \nu_{23}, \nu_{24}, \sigma_{12}, \sigma_{14}, \sigma_{21}, \sigma_{23}, \sigma_{24}\}.$$

To estimate $\boldsymbol{\theta}$, the SEM algorithm described previously can be implemented. When completing the E-step, $\dot{Y}^k$ are obtained by following the rejection sampling steps outlined above. Since we are only considering one type of misclassification, sampling from $\dot{X}(t_m)|\dot{Y}(t_m), \boldsymbol{\theta}^r$ in step 2 results in $\dot{X}(t_m) = \dot{Y}(t_m)$ for any $\dot{Y}(t_m) \neq 1$. When $\dot{Y}(t_m) = 1$, we complete a Bernoulli trial and set $\dot{X}(t_m) = 2$ with probability equal to α_{12}^r. Using equations (6.2) and (6.3), the E-step requires computation of

$$\begin{aligned} \bar{Q}_{K_r}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r-1)}) &= \frac{1}{K_r}\sum_{k=1}^{K_r} \log\, g(\boldsymbol{X}, \dot{Y}^k|\boldsymbol{\theta}) \\ &= \frac{1}{K_r}\sum_{k=1}^{K_r} \log\left[\prod_{m=0}^{M} \alpha_{\dot{Y}_m^k, X_m}\right] \prod_{n=1}^{N}\left[P_{\dot{s}_{(n-1)}^k, \dot{s}_n^k} f_{\dot{s}_{(n-1)}^k, \dot{s}_n^k}(\dot{d}_n^k)\right]^{1-\delta_n^k}\left[\sum_{j\neq \dot{s}_n^k} P_{\dot{s}_n^k, j} S_{\dot{s}_n^k, j}(\dot{d}_n^k)\right]^{\delta_n^k}. \end{aligned}$$

We note that the above equation can alternatively be expressed as follows

$$\begin{aligned} \bar{Q}_{K_r}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r-1)}) &= \frac{1}{K_r}\sum_{k=1}^{K_r} \log\left[g_1(\boldsymbol{X}, \dot{\boldsymbol{Y}}^k|\alpha_{12}) g_2(\dot{Y}^k|\boldsymbol{\theta}_{\bar{\boldsymbol{\alpha}}_{\mathbf{12}}})\right] \\ &= \left[\frac{1}{K_r}\sum_{k=1}^{K_r} \log\, g_1(\boldsymbol{X}, \dot{\boldsymbol{Y}}^k|\alpha_{12})\right]\left[\frac{1}{K_r}\sum_{k=1}^{K_r} \log\, g_2(\dot{Y}^k|\boldsymbol{\theta}_{\bar{\alpha}_{12}})\right], \end{aligned} \qquad (6.4)$$

where $\boldsymbol{\theta}_{\bar{\alpha}_{12}}$ denotes a vector obtained by removing element α_{12} from $\boldsymbol{\theta}$. In completing the M-step, we can take advantage of the separability of the function $\bar{Q}_{K_r}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r-1)})$ with respect to $\boldsymbol{\theta}$ and can complete the maximization by independently maximizing the two components of the product in (4). Define $\eta^k_{1.m}$ such that $\eta^k_{1.m} = 1$ if $(\dot{Y}^k_m, X_m) = (1,1)$ and $\eta^k_{1.m} = 0$ otherwise. Define $\eta^k_{2.m}$ such that $\eta^k_{2.m} = 1$ if $(\dot{Y}^k_m, X_m) = (1,2)$ and $\eta^k_{2.m} = 0$ otherwise. Let C represent the first component of (4) above, such that

$$\begin{aligned} C &= \frac{1}{K_r}\sum_{k=1}^{K_r} \log\, g_1(\boldsymbol{X}, \dot{\boldsymbol{Y}}^k | \alpha_{12}) \\ &= \frac{1}{K_r}\sum_{k=1}^{K_r} \log\left[\prod_{m=0}^{M} \alpha_{\dot{Y}^k_m, X_m}\right] \\ &= \frac{1}{K_r}\sum_{k=1}^{K_r} \log\left[\alpha_{12}^{\left(\sum_{m=0}^{M}\eta^k_{2.m}\right)}(1-\alpha_{12})^{\left(\sum_{m=0}^{M}\eta^k_{1.m}\right)}\right]. \end{aligned}$$

Solving for the maximum with respect to α_{12},

$$\begin{aligned} \frac{dC}{d\alpha_{12}} &= \frac{1}{K_r}\sum_{k=1}^{K_r}\left[\frac{\sum_{m=0}^{M}\eta^k_{2.m}}{\alpha_{12}} - \frac{\sum_{m=0}^{M}\eta^k_{1.m}}{1-\alpha_{12}}\right] \\ &= \frac{1}{K_r}\left[\frac{\sum_{k=1}^{K_r}\sum_{m=0}^{M}\eta^k_{2.m}}{\alpha_{12}} - \frac{\sum_{k=1}^{K_r}\sum_{m=0}^{M}\eta^k_{1.m}}{1-\alpha_{12}}\right]. \end{aligned}$$

Setting $\frac{dC}{d\alpha_{12}} = 0$ and solving for α_{12} gives us

$$\hat{\alpha}_{12} = \frac{\sum_{k=1}^{K_r}\sum_{m=0}^{M}\eta^k_{2.m}}{\sum_{k=1}^{K_r}\sum_{m=0}^{M}\eta^k_{2.m} + \sum_{k=1}^{K_r}\sum_{m=0}^{M}\eta^k_{1.m}}.$$

Therefore, we can obtain the maximum likelihood estimate of α_{12} in the rth iteration, α^r_{12}, by calculating the proportion of $\dot{Y}^k_m = 1$ for which $X_m = 2$ across all $m = 0, \ldots, M$ and $k = 1, \ldots, K_{r-1}$.

The second component of the product in (6.4) represents the Monte Carlo average of a set of complete data log likelihood functions for a process without misclassification, represented by $\dot{Y}^k$. As demonstrated in previous studies, the maximum likelihood estimate of $\boldsymbol{\theta}_{\bar{\alpha}_{12}}$ can be obtained via numerical optimization. Following each iteration of the SEM algorithm, α^r_{12} and $\boldsymbol{\theta}^r_{\bar{\alpha}_{12}}$ are combined to give the updated parameter estimate, $\boldsymbol{\theta}^r$ which is carried forward to the next iteration.

6.6 A Simulation Study: Design

A simulation study was completed to evaluate performance of the proposed algorithm for use with semi-Markov processes with back transitions, intermittently-observed data, and potential misclassification of the healthy state as diseased. Simulated data sets were generated t?o compare the performance of the proposed SEM algorithm that addresses misclassification to the performance of the SEM algorithm described in Chapter 4 under the assumption of no misclassification. To do so, we first simulated 100 data sets consisting of 200 individual trajectories spanning a time interval of 600 days. Once again, the model structure shown in Figure 4.1, consisting of two transient states (1 and 2) and two absorbing states (3 and 4) with the option for back transitions from state 2 to state 1 was assumed. Sojourn times for all transitions were simulated from a Weibull distribution with probability density function

$$f_{ij}(t) = \left(\frac{\nu_{ij}}{\sigma_{ij}}\right)\left(\frac{t}{\sigma_{ij}}\right)^{(\nu_{ij}-1)} \exp\left(-\left(\frac{t}{\sigma_{ij}}\right)^{\nu_{ij}}\right)$$

for $i = 1, 2,\ j = 1, 2, 3, 4,\ i \neq j,\ \sigma_{ij} > 0$, and $\nu_{ij} > 0$. The assumed Weibull distribution also gives us the corresponding survival distribution

$$S_{ij}(t) = \exp\left(-\left(\frac{t}{\sigma_{ij}}\right)^{\nu_{ij}}\right)$$

for $i = 1, 2,\ j = 1, 2, 3, 4,\ i \neq j,\ \sigma_{ij} > 0$, and $\nu_{ij} > 0$. We refer to ν_{ij} and σ_{ij} as the shape and scale parameters, respectively. In the instance that $\nu_{ij} = 1$, the semi-Markov process reduces to a Markov process with exponentially distributed sojourn times. Weibull shape and scale parameters and transition probability parameters used to generate the simulated data sets are displayed in Table 6.1 and are identical to those parameters used in the simulation study described in Chapter 4. These parameter values were selected to reflect the values that might be observed when modeling disease progression from health through mild and severe illness with death as a competing risk among an elderly population. Each simulated trajectory began the 600-day interval in state 1 and ended after having either entered one of the two absorbing states or by being censored in one of the two transient states at the end of the interval. To convert the complete continuous multistate data generated for each of the 100 simulated data sets, we next generated independent observation processes for each data set for each of the 200 trajectories. We assumed independent, exponentially distributed inter-observation times with a 50-day expected inter-observation time. Once generated, the series of observation times for each individual trajectory were applied to the corresponding simulated complete data trajectory to obtain panel data sets.

Using the simulated panel data, for each observation of a trajectory in state 1, a Bernoulli trial was completed with probability of success equal to the assumed misclassification probability, α_{12}, shown in Table 6.1. In the instance of a successful trial, the panel observation of the trajectory was converted such that the process was observed to occupy state 2 instead of state 1. No such conversion was enacted for instances in which the Bernoulli trial resulted in a failure. Thus, the Bernoulli trial served to replicate the conduct of a diagnostic test with false positive rate equal to α_{12}. A misclassification rate of 0.20 was selected because it corresponds to a specificity of 0.80 which is commonly observed in practice. While a higher specificity, corresponding to a lower false positive rate would be preferable in most clinical settings, we selected a specificity of 0.80 to examine the performance of the proposed method in instances where the misclassification rate is likely to introduce substantial bias in the estimation of transition rates and probabilities. In such instances, methods such as the proposed method that address misclassification may find their greatest utility.

Figure 6.1 depicts implementation of the rejection sampling procedure in its entirety for a single trajectory. The individual in Figure 6.1 is observed at 5 time points with the first two observations occurring while the individual is in state 1, the next two observations occurring while the individual is in state 2, and the final observation occurring when the individual has reached absorbing state 3. Let **A** indicate the first continuous trajectory sampled from the candidate distribution, $\dot{Y}$. Let the gray boxes beneath trajectory **A** indicate the results of the second stage of sampling, $\dot{X}_m$ for $m = 1, \ldots, 5$. We reject trajectory **A** because $\dot{X}_2 \neq X_2$. We then proceed to sample continuous trajectory **B** and subsequently complete the second stage of sampling using the Bernouli trial approach described above. Trajectory **B** is accepted because $\dot{X}_m = X_m$ for all $m = 1, \ldots, 5$ such that it is entirely congruent with the observation process above.

For each simulated, misclassified, data set, the SEM algorithm was implemented, first, under the assumption of no misclassification and, second, while addressing misclassification using the approach described in the present Chapter. For both implementations, the M-step of the SEM algorithm was carried out using the R *constrOptim* function for linearly

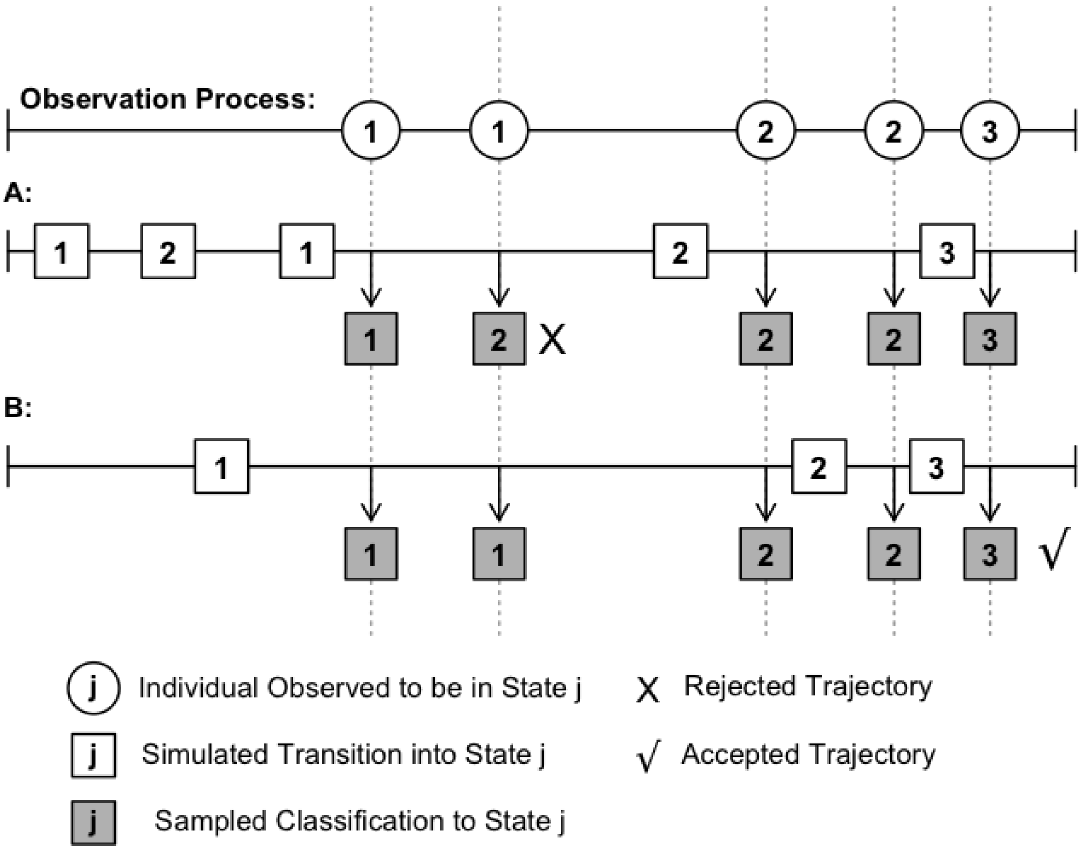

Observation Process:
1
1
2
2
3
A:
1
2
1
2
3
1
2
X
2
2
3
B:
1
2
3
1
1
2
2
3
√
j Individual Observed to be in State j
j Simulated Transition into State j
j Sampled Classification to State j
X Rejected Trajectory
√ Accepted Trajectory

Figure 6.1: The SEM algorithm with misclassification's rejection sampling scheme applied to a single individual. Following both stages of sampling, trajectory A would be rejected after comparison to the observation process at the top and trajectory B would be accepted.

constrained optimization. Linear constraints were used to ensure the following:

$$\nu_{ij} > 0, \sigma_{ij} > 0 \text{ for } (i,j) \in \{(1,2),(1,4),(2,1),(2,3),(2,4)\}$$
$$0 \leq P_{ij} \leq 1 \text{ for } (i,j) \in \{(1,2),(2,1),(2,3)\}$$
$$0 \leq P_{21} + P_{23} \leq 1$$
$$0 \leq \alpha_{12} \leq 1.$$

Estimates of P_{14} and P_{24} could be obtained given the other estimated transition probabilities. For both implementations, we determined that a modest Monte Carlo sample size, m_r, would suffice as a result of the stringent rejection sampling procedure employed which resulted in a relatively low acceptance rate. We set $m_1 = 1$ and incremented m_r by 1 at each successive iteration. We adopted the same stopping rule as outlined in Section 4.6 and required that the following hold for three consecutive iterations

$$\max_i \left(\frac{|\theta_i^{(r)} - \theta_i^{(r-1)}|}{|\theta_i^{(r-1)}| + \delta_1} \right) < \delta_2, \tag{6.5}$$

where $\delta_1 = 0.001$ and $\delta_2 = 0.05$. To minimize the impact of Monte Carlo error on the final parameter estimates, once the stopping rule was satisfied, the final estimates were taken as the average of the estimates obtained during the last 5 iterations of the algorithm.

The two implementations of the SEM algorithm differ with respect to the execution of the rejection sampling approach. As described in Chapter 4, the implementation of the SEM algorithm ignoring misclassification relies on repeated sampling of continuous paths, $\dot{Y}$, until $\dot{Y}(t_m) = Y(t_m)$ for all $m = 1, \ldots, M$, in which case a sampled path is accepted and retained for use in completing the next M-step. Alternatively, the implementation of the SEM algorithm addressing misclassification relies on the series of steps outlined in Section 6.4. Continuous sample paths, $\dot{Y}$, are sampled followed by a second stage of sampling in which $\dot{X}_m$ are obtained by probabilistic application of the current estimate of the misclassification rate.

To conduct the computationally intense rejection sampling and iterative maximization, we used the R package *Rcpp* which integrates R and C++ to improve performance. All components of this simulation study used computational and storage services associated with the Hoffman2 Shared Cluster provided by the UCLA Institute for Digital Research and Education's Research Technology Group.

Table 6.1: Comparison of bias and accuracy of parameter estimates across different estimation approaches as applied to 100 simulated data sets generated using the parameter values displayed in the *True Value* column.

		Complete Data				**SEM Ignoring Misclassification**				**SEM Addressing Misclassification**			
	True Value	Mean	SD	% Bias	Median	Mean	SD	% Bias	Median	Mean	SD	% Bias	Median
P_{12}	0.90	0.90	0.02	0.3	0.90	0.96	0.01	6.5	0.96	0.90	0.02	0.3	0.90
P_{21}	0.30	0.30	0.03	1.6	0.29	0.65	0.05	117.5	0.66	0.27	0.06	11.0	0.26
P_{23}	0.55	0.55	0.03	0.8	0.55	0.27	0.04	51.0	0.27	0.58	0.05	5.2	0.58
ν_{12}	0.75	0.75	0.03	0.3	0.75	0.81	0.10	8.1	0.83	0.78	0.08	3.5	0.77
ν_{14}	1.50	1.58	0.29	5.3	1.54	1.81	0.55	20.9	1.78	1.76	0.67	17.6	1.61
ν_{21}	1.25	1.28	0.13	2.3	1.26	1.00	0.10	20.1	1.01	1.20	0.30	4.3	1.17
ν_{23}	2.00	2.02	0.15	0.9	1.99	2.05	0.29	2.5	2.05	1.97	0.27	1.5	1.96
ν_{24}	1.75	1.84	0.27	5.3	1.83	1.72	0.56	1.8	1.55	2.17	0.71	24.1	2.01
σ_{12}	60	61	5	1.4	60	26	4	56.4	26	65	9	8.1	65
σ_{14}	150	151	24	0.6	150	89	36	41.0	81	155	35	3.2	154
σ_{21}	70	71	7	1.2	71	25	5	63.9	25	74	20	6.0	72
σ_{23}	80	80	3	0.3	80	79	5	1.1	79	82	6	2.1	81
σ_{24}	200	199	21	0.5	198	194	42	2.8	192	202	33	0.9	202
α_{12}	0.20	-	-	-	-	-	-	-	-	0.21	0.03	2.7	0.20

6.7 A Simulation Study: Results

Results from estimation of the complete simulated data including observation of all trajectories continuously over the entire 600-day interval, are displayed in Table 6.1 for comparison purposes. These complete data results were obtained by applying the straightforward maximum likelihood estimation approach to each of the simulated data sets prior to application of the simulated panel data observation process. As anticipated, the complete data parameter estimates display minimal bias. Among these estimates, the largest observed bias occurred for the Weibull shape parameters associated with transition from state 1 to state 4 and from state 2 to state 4 (both 5.3% bias). As noted in previous chapters, these parameters appear to be highly sensitive to the limited length of the window of observation as related to the assumption of time-homogeneity. All other complete data parameter estimates were effectively unbiased ($\leq$ 2.3% bias).

The impact of ignoring misclassification when misclassification exists is displayed in Table 6.1. Parameter estimates obtained from implementation of the SEM algorithm ignoring misclassification exhibited considerable bias. Of specific interest, transition probability estimates for the probability of transition from state 1 to state 2 and from state 2 back to state 1 were overestimated while the probability of transition from state 2 to state 3 was drastically underestimated. For a process in state 2, the SEM algorithm ignoring misclassification estimated that the next transition would be a back transition to state 1 with probability 0.65, a transition to state 3 with probability 0.27, and a transition to state 4 with probability 0.08. Relative to the true underlying process, the estimated probability of back transition is substantially higher (0.65 versus 0.30) when misclassification is ignored because erroneously classifying an observation as being in state 2 between two adjacent observations in state 1 implies an additional back transition which would not be present for a process observed without measurement error. Shape parameters that were estimated with substantial bias when ignoring misclassification include the parameter for the transition from state 1 to state 4, which was overestimated, and the shape parameter for the transition from state 2 back to state 1, which was underestimated. When ignoring misclassification, the scale parameters associated with transitions from state 1 to state 2, state 1 to state 4, and state 2 to state 1 were substantially underestimated. While the true value of the scale parameter for transition from state 1 to state 2 was equal to 60, the SEM algorithm ignoring misclassification resulted in a median estimate of only 26. This observed underestimation in the presence of ignored misclassification can be explained by the introduction of artificial implied back transitions in the data which likely shortened the apparent average sojourn time spent in state 1.

Performance of the proposed SEM algorithm which addresses misclassification is displayed in Table 6.1. In implementing the SEM algorithm addressing misclassification, the specified stopping rule resulted in an average of 22.2 iterations across all 100 data sets with a minimum of 14 and a maximum of 38 iterations. The parameter estimates obtained from implementing the SEM algorithm while addressing misclassification demonstrated greatly reduced bias relative to the estimates that resulted when misclassification was ignored. In many instances, estimates closely approximated the performance of the estimates obtained using the complete data. Transition probabilities, including the probability of back transition, were estimated with minimal bias when misclassification was addressed and the standard deviations across data sets were only slightly larger than when implementing complete data estimation. Remarkably, the SEM algorithm addressing misclassification was able to accurately estimate the probability of misclassification, with the median across all 100 simulated data sets equal to the true value of 0.20. In addressing misclassification, there was a tendency towards slight underestimation of the probability of back transition (11.0% bias) and corresponding overestimation of the probability of transition from state 2 to state 3 (5.2% bias). Parameter estimates that demonstrated noticeable bias even after implementation of the SEM algorithm addressing misclassification include the shape parameter for transition from state 1 to state 4 (17.6% bias) and the shape parameter for transition from state 2 to state 4 (24.1% bias). As mentioned when discussing the complete data results, the bias observed in these estimates may signify the heightened sensitivity of these shape parameters to the limited length of the interval of observation required to establish time-homogeneity.

6.8 The Nun Study: Design and Results

When applying the SEM algorithm with rejection sampling and misclassification to the Nun Study described in section 5.1 we consider the possibility that there was a non-negligible probability of misclassification. We consider misclassification of healthy individuals with intact cognition into a state of cognitive impairment. The impact of false positive diagnoses in the evaluation of cognitive impairment is an important issue in the field of dementia and Alzheimer's research [87, 88]. Experts have noted that there is substantial risk of misclassifying mild cognitive impairment (MCI) in healthy older adults and the risk is even greater when multiple cognitive measures are administered, as was the case with the Nun Study. Brooks et al. attribute false positives to numerous factors such as long-standing and static relative cognitive weaknesses, reversible causes of poor performance on memory measures, and situational influences on performance [89]. Nevertheless, continued efforts to accurately identify true positives are necessary in research and practice because there exists a large body of consistent evidence that individuals with MCI are at increased risk of subsequent progression to dementia and Alzheimer's disease [90, 91, 92]. We therefore aim to build upon the applied results presented in the previous chapter by adapting our model to include the possibility of misclassification.

In this secondary analysis of the Nun Study, we use the exact same data and model specifications as outlined in Section 5.1. This includes the assumption that all 544 participants entered into the state we refer to as *intact cognition* at exactly age 70 implying that the intact and impaired cognition states are defined only among those individuals 70 years of age or older. We chose to implement the same two month acceptance window around the exact date of death and used the stopping criteria expressed in formula (4.4) with $\delta_1 = 0.001$ and $\delta_2 = 0.08$.

In implementing the SEM algorithm addressing misclassification to model the Nun Study data, the stopping criterion was met after the 31st iteration. Parameter estimates obtained using this modeling approach are displayed in Table 6.2. For comparison purposes, the parameter estimates obtained in Chapter 5 using the SEM algorithm but ignoring misclassification are also re-displayed in Table 6.2. The impact of addressing misclassification on the transition probability and other parameter estimates was dramatic. The transition probability parameter estimates obtained when fitting the SEM algorithm addressing misclassification to the Nun Study data are displayed in Figure 6.2. For an elderly individual in a state of intact cognition, after accounting for misclassification, there is an estimated 0.75 probability that the next state visited will be impaired cognition and a 0.25 probability that the individual will transition directly to death. In comparison, the model fit while ignoring misclassification estimated a 0.94 probability that the next state visited will be impaired cognition and only a 0.06 probability of transition directly to death. The model addressing misclassification resulted in an estimated probability of falsely classifying an individual with intact cognition as having impaired cognition of 0.31. Correspondingly, the estimated probability that the next transition experienced by an elderly individual in a state of impaired cognition would be a back transition to intact cognition was only 0.05 in the model addressing misclassification. This compares to an estimated probability of 0.58 in the model fit under the assumption of no misclassification.

Many of the Weibull shape and scale parameter estimates associated with the sojourn time distributions were also greatly impacted by implementing the modeling approach that addresses misclassification. The shape parameter estimates associated with transitions from intact to impaired cognition, intact cognition to death, and impaired cognition to death were all higher when estimated while accounting for misclassification. These higher estimates imply that the hazard rates associated with these transitions are estimated to be increasing to an even greater extent as a function of time elapsed in the current state when addressing versus not addressing misclassification. Alternatively, the shape parameter estimate associated with back transition from impaired to intact cognition was lower when addressing misclassification, implying a more constant back transition hazard rate as a function of time elapsed in a state of impaired cognition ($\nu_{21} = 1.10$).

Weibull scale parameter estimates associated with transitions from intact to impaired cognition and intact cognition to death were higher when estimated while accounting for misclassification. These higher estimates imply longer average durations of time spent in

Table 6.2: Parameter estimates for the Nun Study model fit using the proposed SEM approach with rejection sampling for the two cases in which misclassification is ignored and addressed. (1 = Intact Cognition, 2 = Impaired Cognition, 3 = Dementia, 4 = Death)

	SEM Ignoring Misclassification	**SEM Addressing Misclassification**
Parameter	Estimate	Estimate
P_{12}	0.94	0.75
P_{21}	0.58	0.06
P_{23}	0.16	0.44
ν_{12}	1.35	2.33
ν_{14}	0.99	4.90
ν_{21}	1.44	1.10
ν_{23}	1.68	1.53
ν_{24}	1.34	2.26
σ_{12}	4.74	13.08
σ_{14}	8.89	20.03
σ_{21}	2.47	1.10
σ_{23}	9.98	8.54
σ_{24}	9.79	9.59
α_{12}	-	0.31

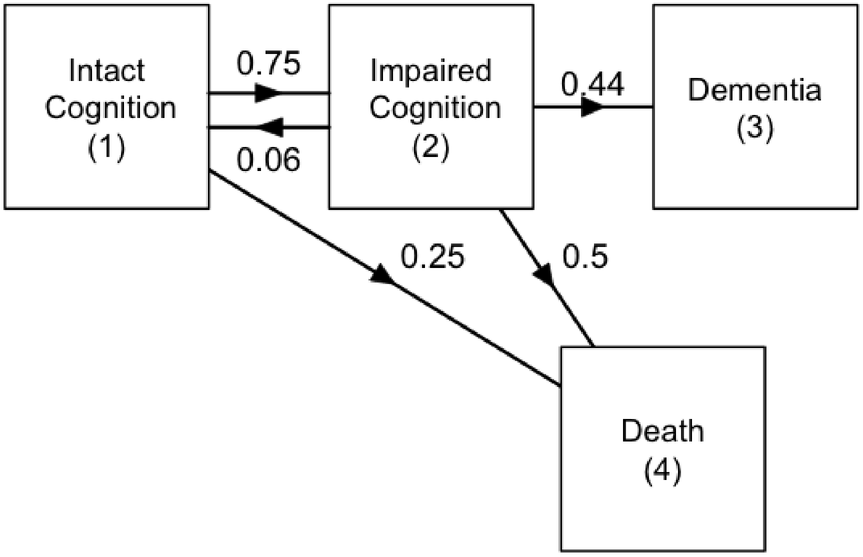


Figure 6.2: Structural diagram for the multistate model assumed in both the simulation study and Nun Study data analysis with Nun Study transition probability estimates obtained using the SEM approach addressing misclassification. The estimated probability of misclassification of an individual with intact cognition in a state of impaired cognition is 0.31.

intact cognition before progressing to either impaired cognition or death. As shown in Figure 6.3, the median sojourn time for an elderly individual in a state of intact cognition was estimated to be 11.18 years if their next transition was to a state of impaired cognition or 18.59 years if their next transition was to death. The estimated sojourn time distributions for these transitions from intact cognition are both fairly symmetrical with similar mean and median values. In implementing the model addressing misclassification, the scale parameter associated with back transition from impaired to intact cognition was estimated to be lower, relative to the estimate obtained when ignoring misclassification. As depicted in Figure 6.3, in the model addressing misclassification, the vast majority of back transitions were estimated to occur during the 2-3 years following entrance into a state of impaired cognition. The median time to transition back from impaired to intact cognition was estimated to be 0.79 years with a very small proportion of sojourn times extending out beyond 5 years. Scale parameter estimates associated with transitions from impaired cognition to dementia and death did not differ substantially when comparing estimates obtained while ignoring and accounting for misclassification (Table 6.2). The median sojourn time for transition from impaired cognition to dementia was estimated to be 6.72 years and the median sojourn time for transition from impaired cognition to death was estimated to be 8.16 years. The sojourn time distribution for transition from impaired cognition to death demonstrated is relatively symmetrical, while some right skew is noticeable for the sojourn time distribution for the transition from impaired cognition to dementia (Figure 6.3).

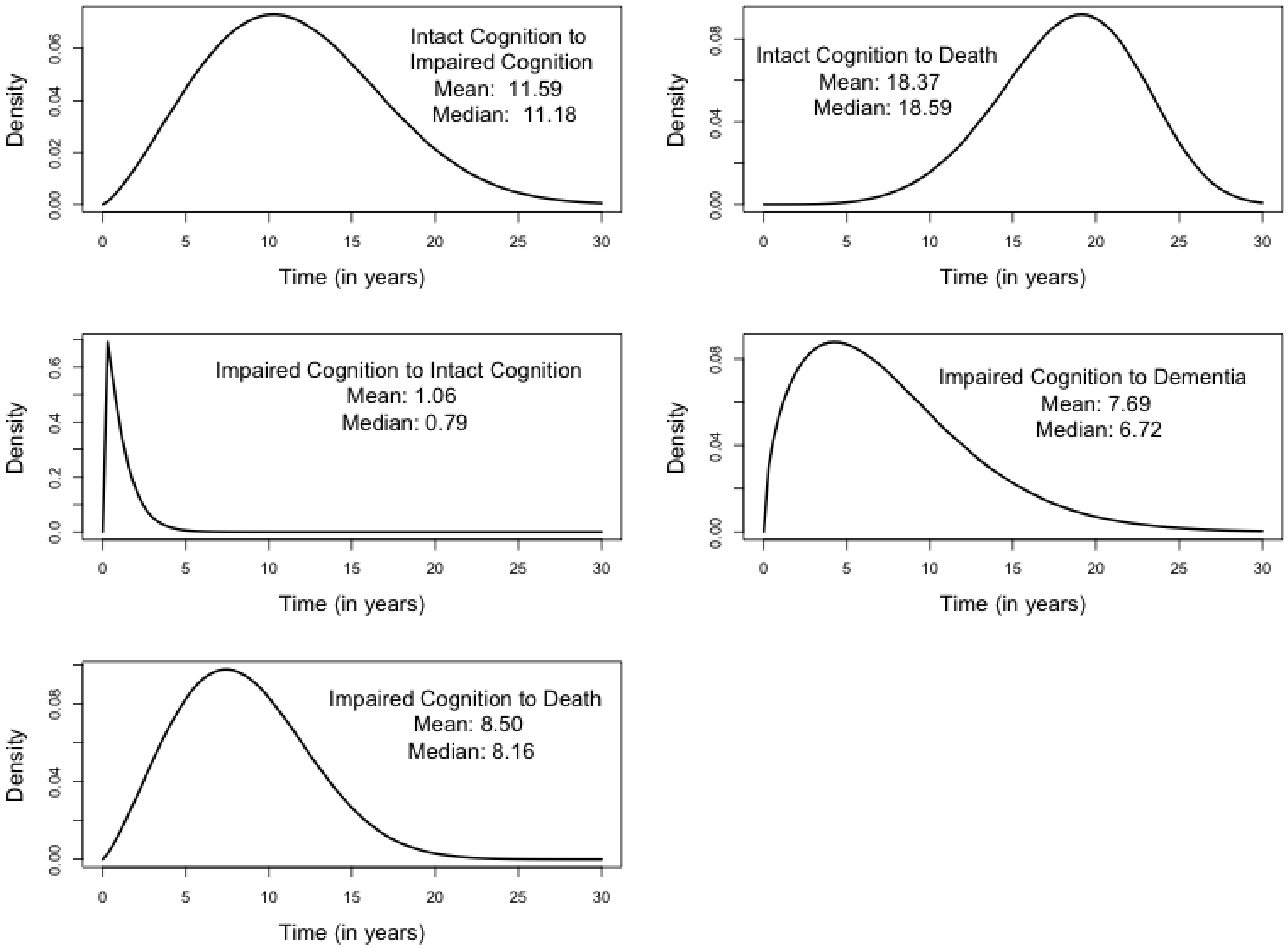


Figure 6.3: Sojourn time probability density functions based on the Weibull parameter estimates obtained from the Nun Study model fit using the proposed SEM approach with rejection sampling after addressing misclassification

6.9 Discussion

In this chapter we proposed an approach to addressing misclassification when estimating semi-Markov models with back transitions and intermittently-observed data. The proposed method simultaneously accomplishes two tasks: mitigating the parameter bias introduced when fitting semi-Markov models while ignoring misclassification, and producing unbiased estimates of the misclassification rate. The method we proposed relies on the same SEM algorithm described and implemented in Chapters 4 and 5 with an extension which enables iterative estimation of one or more additional parameters representing misclassification rates. Another distinguishing feature of the approach described in this chapter is the adaptation of the rejection sampling strategy to include two stages of sampling corresponding to the sampling of continuous multistate trajectories and the subsequent sampling of the misclassification mechanism at the discrete observation time points. Rejection of sampled trajectories that are not coherent with the observed data now occurs after both stages of sampling have been completed.

Using a simulation study, we demonstrated that estimating transition probabilities and sojourn time distribution parameters while ignoring misclassification that truly exists can introduce non-trivial bias. Preferably, misclassification arising from imperfect disease classification instruments would be estimated by comparison to a gold standard instrument that can definitively discriminate between states of disease and health. As noted previously, for many diseases a gold standard either doesn't exist or is not practical for use. As an alternative, we have shown that the misclassification rate can be estimated with no gold standard information available. Instead of relying on a gold standard, the proposed estimation method relies on the parametric modeling assumptions and the richness of the data provided by the entire sample of trajectories to disentangle true from false positives. Additionally, the misclassification rate can be estimated simultaneously, and under the same general framework as implemented previously for estimation of a semi-Markov model with back transitions and panel data. The estimation approach relying on the SEM algorithm with rejection sampling is capable of accurately estimating the misclassification rate because the stochastic components of the procedure minimize potential identifiability issues while enabling exploration of different misclassification rates and the coherence of those rates with the observed multistate data. Results of the simulation study we completed suggest that within each iteration the estimation approach was successfully able to identify observations that were anomalous when considering the expected behavior of the underlying semi-Markov model being concurrently estimated. Recognition of these anomalous observations and their frequency within the observed data allowed the procedure to estimate a misclassification rate within each iteration and led to convergence over the course of numerous iterations. The method presented here accomplishes the same objective as targeted when fitting a hidden semi-Markov model. Use of hidden semi-Markov models, however, has been historically limited by difficulties with estimation which are exacerbated when attempting to fit a hidden semi-Markov model to panel data while allowing for back transitions. The method proposed and implemented in this chapter may serve to extend the utility of such models in disease progression research when misclassification is a real concern and gold standard information is not readily available.

The impact of addressing misclassification on our interpretation of the progression of Nun Study participants through various states of cognitive impairment was significant. Whereas the model ignoring misclassification seemed to imply that back transitions from impaired to intact cognition were relatively common and that elderly individuals remained in a state of intact cognition for periods of less than 5 years on average, the model accounting for misclassification suggests that back transitions are relatively infrequent and that the average elderly individual with intact cognition will remain in a state of intact cognition for over 10 years. This dramatic shift in our perspective on the rate of back transition is attributable to the estimated misclassification rate of 0.31, which suggests that a cognitively intact elderly individual who shows up for an assessment at any given point in time has a 31% chance of being mistakenly classified as having impaired cognition. This is certainly a non-negligible false positive rate but is not incongruent with previously published estimates. A meta-analysis published in 2009 estimated a pooled proportion for specificity of 65.4%, corresponding to a false positive rate of 34.6% when examining the accuracy of the MMSE in identifying subjects with MCI versus healthy subjects in a specialist setting (memory clinic or Alzheimer's disease center) [93]. Although the Nun Study was conducted outside of a specialist setting, this finding is directly applicable because participants in the Nun Study who failed the MMSE were classified in a state of cognitive impairment.

In addition to the limitations mentioned in the previous chapters, implementation of the SEM algorithm with rejection sampling while addressing misclassification is subject to several additional limitations that are worth mentioning. In addressing misclassification, we chose only to consider misclassification for individuals in a state of intact cognition and did not consider the potential for cognitively impaired individuals to be misclassified in a state of intact cognition. Thus, in a general sense, our model allowed only for uni-directional movement of an individual observed at a given point in time from a state of intact to impaired cognition. While this did not appear to introduce bias in our simulation study results, the impact of assuming only one type of misclassification event on the Nun Study estimates is unknown. Furthermore, each misclassification rate considered is associated with at least one additional parameter (more if the misclassification rate is expressed as a function of explanatory variables) that needs to be estimated using the iterative approach described. There is real concern that the addition of parameters will result in identifiability or convergence issues, especially when the additional parameters are associated with misclassification. Assuming another misclassification event is possible increases the proportion of incompleteness of the observed data and reduces our ability to recover accurate and meaningful estimates of other important model parameters.

Implicit in the modeling framework adopted in the current Chapter was the assumption that the misclassification mechanism operated independent of individual, external time, and time elapsed in the current state. We also assumed that misclassifications occurred as independent events. In practice, the probability of misclassification for an individual at a given instant in time may vary across individuals and time. An important area for future work may be attempting to explain this heterogeneity among individuals and over time using explanatory variables and/or random effects. It is also reasonable to consider the probability of misclassification being associated with the occurrence of previous misclassification events, in which case a more complicated structure for the misclassification rate would need to be considered.

In future work, we plan to conduct further simulation studies to examine the sensitivity of the proposed method to varying rates of misclassification and rates of back transition. Intuitively, our ability to accurately estimate both the misclassification rate and semi-Markov model parameters is increasingly compromised as the true underlying misclassification rate is increased. One can imagine an extreme scenario in which the true false positive rate is 50%, such that a healthy individual has equal probability of being classified as healthy versus diseased. For an individual in a state of health for 10 years and observed once annually, 5 of the 10 years could result in classification in a disease state and the estimation approach described would have difficulty distinguishing between true and false positives. Identifying at what point, and under what conditions the performance of the proposed estimation approach is seriously compromised will improve our recommendations for use of the method across different disease progression applications.

CHAPTER 7

Discussion

7.1 Multistate Modeling Challenges and Applications Addressed

In this dissertation, I outlined several challenges of recent and continued importance in the field of multistate modeling. I proposed, applied, and evaluated appropriate methodology to address these challenges. The three methodological challenges addressed were motivated by distinct applied research questions and the limitations of the data available for use in answering these questions.

We were faced with the question of whether concurrent partnership patterns are associated with increased rates of HIV transmission relative to serially monogamous patterns, when holding all other sexual partnership dynamics, such as number and duration of partnerships, fixed. To address this question, it was necessary to develop a modeling approach that would yield estimates of both the extent and the magnitude of concurrency within a population. The available data consisted of retrospective sexual history reports provided cross-sectionally by a sample of men having sex with men and seeking HIV testing at a Los Angeles clinic. Among this sample, a high percentage (60%) of the partnerships reported were one-offs, meaning that they were reported as having the same first and last date of sexual intercourse. Previous attempts to analyze retrospective sexual history data for use in measuring concurrency used the partnership as the unit of observation and thus ignored heterogeneity across individuals and time. To improve upon the existing methodology, we developed a multistate model that treated the individual as the independent unit of observation and inherently modeled the dependence among partnerships engaged in by the same individual at the same or different times. The model we developed enabled straightforward estimation of the point prevalence of concurrency and the mean duration of concurrency while being flexible enough to allow incorporation of explanatory variables. In formulating the model, we also jointly fit a point process which accounted for one-off sexual encounters and allowed the occurrence of one-offs to potentially impact the subsequent formation and dissolution of ongoing partnerships. By describing this novel approach for the joint modeling of sexual partnership patterns using retrospective sexual history data we have provided a tool that can be used to answer pertinent questions in the field of HIV transmission research. Model parameter estimates can be used to draw inferences about the population from which the sample was drawn, compare partnership patterns across samples, or as input to dynamic mathematical models with the capacity to answer questions about the impact of potential interventions on rates of HIV transmission within a community.

A second motivating question regarded the modeling of disease progression for diseases in which back transitions from illness to health are possible and data are available in the form of intermittent observations of an individual's disease status. In this instance, interest lied in the accurate estimation of transition probabilities and sojourn time distribution parameters. Unbiased estimation of such parameters would enable construction of models that could be used to identify factors impacting rates and probabilities of transition between various states of health and disease. Of specific interest was the transition back from a state of disease to a state of health since this reverse transition is typically associated with patient recovery and is the target of many public health interventions. Multistate models for disease progression can be estimated using a straightforward maximum likelihood approach when knowledge of the sequence of states visited and the duration of time spent in each state is available for all individuals in a random sample drawn from the population of interest. When panel data are collected, however, the sequence of states an individual occupies across an observation time interval is typically unknown. When back transitions are possible, consideration of all potential state sequences between two discrete observations leads to an intractable likelihood function. Thus, the challenge we faced was the development of a computationally feasible method for the unbiased estimation of semi-Markov models using panel data in the presence of back transitions. In response, we developed a simulation-based iterative algorithm involving rejection sampling of complete trajectories to complete the traditional E-Step of the EM algorithm for incomplete data. We evaluated the performance of the proposed algorithm by completing a simulation study in which we compared bias and accuracy of parameter estimates obtained using the proposed SEM algorithm to estimates obtained using a commonly-employed naive approach that assumed no unobserved states. Whereas the naive approach resulted in biased parameter estimates and underestimation of the rate of back transition, the SEM approach produced unbiased estimates through recovery of unobserved back transitions after relatively few iterations. After demonstrating the performance of the proposed SEM method in the simulation study, we applied the method to dementia onset modeling using the Nun Study panel data collected from a sample of elderly individuals participating in a longitudinal study on aging and Alzheimer's disease. Results of this application consisted of estimated transition probabilities and Weibull sojourn time parameters for transitions occurring between the two transient states of intact and impaired cognition, and transitions from either of these two states to either of the two absorbing states of dementia or death.

To improve utility of the proposed SEM approach for disease progression modeling in applications such as the Nun Study, we next considered an extension that allows for misclassification of a healthy individual in the diseased state. We developed this extended estimation approach by modifying the rejection sampling strategy to include two stages and formulating an expression for the iterative estimation of an additional parameter representing the misclassification rate. We demonstrated the performance of the proposed extension using a simulation study in which we compared performance of the proposed method that addresses misclassification to the SEM method presented previously which ignores misclassification. Results demonstrated that estimating transition probabilities and sojourn time distribution parameters while ignoring misclassification that truly exists can introduce non-trivial bias. Importantly, the extended method was shown to simultaneously mitigate the parameter bias introduced when ignoring misclassification and produce unbiased estimates of the true misclassification rate. This method can be useful in a wide range of disease progression applications in which panel data are collected and a non-negligible false positive rate is suspected with no gold standard diagnostic information available. We used this new extended methodology to re-fit a multistate model to the Nun Study data while considering the potential for misclassification of individuals with intact cognition into a state of impaired cognition. Findings from this re-analysis suggest that the rate of misclassification may be relatively high among this sample and true back transitions from impaired to intact cognition may be relatively rare. These results emphasize the importance of accounting for misclassification in dementia onset modeling when considering cognitive impairment as an intermediate disease state predictive of eventual progression to dementia and Alzheimer's disease.

7.2 Future Work

This dissertation lays the foundation from which to explore a number of other important research questions in the area of multistate modeling with back transitions.

As mentioned previously, the most immediate priority for future work will entail additional simulation studies to examine the sensitivity of the proposed SEM method addressing misclassification to varying rates of misclassification. Identifying the lower bound on the true underlying misclassification rate, above which the performance of the proposed estimation approach becomes significantly compromised would improve our recommendations for use of the method across different disease progression applications. In a similar study, we will examine the sensitivity of the proposed method which addresses misclassification to varying rates of back transition. These results will help define settings in which the methods presented in Chapter 6 are likely to perform optimally.

In addressing the three methodological questions, much of our effort was focused on developing flexible models capable of providing unbiased parameter estimates given the type of data typically available to public health researchers. Having demonstrated the viability of the modeling frameworks and estimation procedures developed herein, we believe there remains several important opportunities to utilize these multistate models to answer specific applied research questions regarding the impact of hypothesized explanatory variables on the rates and probabilities of transition. In our Nun Study application, for example, incorporation of covariates such as age and genetic markers would provide an important contribution to the applied literature by allowing us to assess the impact of such characteristics on transition rates and probabilities in modeling of dementia onset [60, 76]. In further utilizing the joint model for sexual partnership patterns, we would like to consider the use of additional covariates and perhaps sexual history data collected from different samples to draw further inference about community or individual characteristics associated differences in concurrency. Identification of such differences could improve our understanding of variations in HIV transmission observed across different populations and sexual networks.

A third area of interest for future work is related to state-dependent biased sampling of multistate trajectories. In traditional survival models designed to model time to a single event, methods to account for sampling bias in which only survival times occurring within a certain interval are observed are well-established. In multistate modeling we often encounter a related but distinct issue in which sampling bias is introduced because individuals are observed only if their trajectories over a specified time interval meet certain criteria. For instance, eligibility criteria for participation in the MetroMates study described in Chapter 3 stipulated that an individual report at least one instance of sexual intercourse with a male partner in the past year. This data collection strategy certainly necessitates the careful interpretation of results as being only generalizable to a population of sexually active individuals. Even so, the results obtained after excluding these individuals who reported no sexual activity during the previous year could potentially provide a biased representation of the sexually active population. If an independent sample of sexually active individuals were recruited, enrolled, and followed prospectively for a year, a certain percentage of these study participants would report no instances of sexual intercourse and this information is important for incorporation into the likelihood function when estimating a multistate model. Other biased sampling scenarios can be imagined, such as the exclusion of individuals who do not enter a disease state at least once during a pre-specified time interval when modeling disease progression. In future work, we aim to extend the SEM modeling approach to accommodate state-dependent biased sampling of trajectories. This work can be completed first by conducting a simulation study and then by application of the proposed method to the modeling of sexual partnership patterns.

Another extension to the SEM modeling approach that we would like to pursue is to develop methodology to effectively account for the left censoring of the sojourn times associated with the first state occupied by each individual in the sample. In a recent paper, Cai et al. used an analogue of the SEM algorithm to simulate a cohort of subjects from which imputed values of the elapsed sojourn times at entry into the study were drawn [58]. The SEM approach we developed herein can be extended to incorporate imputation of elapsed sojourn times under a single unified framework relying on the stochastic approximation to the expectation for the complete data including the left censored sojourn times.

7.3 Closing Thoughts

Both the sexual partnership joint modeling procedure and the SEM-based estimation approach for disease onset models are important in that they enable public health researchers to answer crucial questions that had been previously addressed using inferior or naive methods. Cox proportional hazards and other traditional survival models exploring time to event for a single outcome are ubiquitous. As an alternative to constructing separate survival models for each of a number of interrelated outcomes, multistate models have gained popularity in recent years. Multistate models can be used to examine pathways of association across multiple alternative or intermediate states which are not directly apparent when modeling survival separately for each state or outcome. Applications of multistate models with back transitions in the health sciences are many and varied, nevertheless, use of such models has been hindered by insufficient methods for model formulation and estimation. The limited availability of complete data consisting of all states occupied and times of transition between states has likely further limited the use of such models in health research settings in which intermittent observations are the norm.

We anticipate that the methods presented in this dissertation will enable greater utility of multistate models with back transitions in public health research. With each passing year, as members of the statistical community, we notice a shift in the preference for stochastic and computational methods that can be used to circumvent intractable problems. These methods have been increasingly adopted in recent years to answer previously unanswerable questions and to explain phenomena of historical uncertainty. It is our hope that the statistical challenges and applications described herein will encourage the use of multistate models with back transitions in answering difficult questions of importance in the field of public health research.

References

[1] Aralis HJ, Gorbach PM, Brookmeyer R. Measuring concurrency using a joint multistate and point process model for retrospective sexual history data. *Statistics in Medicine* 2016; **35**(24):4459–4473.

[2] Jepsen P, Vilstrup H, Andersen PK. The clinical course of cirrhosis: the importance of multistate models and competing risks analysis. *Hepatology* 2015; **62**:292–302.

[3] Gangnon RE, Lee KE, Klein BEK, Iyengar SK, Sivakumaran TA, Klein R. The Y402H variant in the complement factor H gene affects incidence and progression of age-related macular degeneration: results from multi-state models applied to the Beaver Dam Eye Study. *Archives of Ophthalmology* 2012; **130**(9):1169–1176.

[4] Duffy A, Horrocks J, Doucette S, Keown-Stoneman C, McCloskey S, Grof P. The developmental trajectory of bipolar disorder. *The British Journal of Psychiatry* 2014; **204**:122–128.

[5] Zhang X, Li Q, Rogatko A, Tighiouart M, Hardison RM, Brooks MM, Kelsey SF, Kaul S, Merz CNB. Analysis of the bypass angioplasty revascularization investigation trial using a multistate model of clinical outcomes. *The American Journal of Cardiology* 2015; **115**:1073–1079.

[6] Clark DE, Ostrander KR, Cushing BM. A multistate model predicting mortality, length of stay, and readmission for surgical patients. *Health Services Research* 2015; .

[7] Eulenburg C, Schroeder J, Obi N, Heinz J, Seibold P, Rudolph A, Chang-Claude J, Flesch-Janys D. A comprehensive multistate model analyzing associations of various risk factors with the course of breast cancer in a population-based cohort of breast cancer cases. *American Journal of Epidemiology* 2016; **183**(4):325–334.

[8] Jazic I, Schrag D, Sargent DJ, Haneuse S. Beyond composite endpoints analysis: semicompeting risks as an underutilized framework for cancer research. *Journal of the National Cancer Institute* 2012; **108**(12):1–6.

[9] Liu Y, Wang M, Morris AD, Doney ASF, Leese GP, Pearson ER, Palmer CNA. Glycemic exposure and blood pressure influencing progression and remission of diabetic retinopathy. *Diabetes Care* 2013; **36**(12):3979–3984.

[10] Pan SL, Lien IN, Yen MF, Lee TK, Chen THH. Dynamic aspect of functional recovery after stroke using a multistate model. *Archives of Physical Medicine and Rehabilitation* 2008; **89**(6):1054–1060.

[11] Marioni RE, Hout A, Valenzuela MJ, Brayne C, Matthews FE. Active cognitive lifestyle associates with cognitive recovery and a reduced risk of cognitive decline. *Journal of Alzheimer's Disease* 2012; **28**:223–230.

[12] Foxman B, Newman M, Percha B, Holmes KK, Aral SO. Measures of sexual partnerships: lengths, gaps, overlaps, and sexually transmitted infection. *Sexually Transmitted Diseases* 2006; **33**(4):209–214, doi:10.1097/01.olq.0000191318.95873.8a.

[13] Baggaley RF, White RG, Marie-Claude B. HIV transmission risk through anal intercourse: systematic review, meta-analysis and implications for HIV prevention. *International Journal of Epidemiology* 2010; **39**(4):1048–1063, doi:10.1093/ije/dyq057.

[14] Royce RA, Sena A, Cates W, Cohen MS. Sexual transmission of HIV. *The New England Journal of Medicine* 1997; **336**(15):1072–1078, doi:10.1056/NEJM199704103361507.

[15] Boily M, Alary M, Baggaley RF. Neglected issues and hypotheses regarding the impact of sexual concurrency on HIV and sexually transmitted infections. *AIDS Behavior* 2012; **16**:304–311, doi:10.1007/s10461-011-9887-0.

[16] Morris M, Kurth AE, Hamilton DT, Moody J, Wakefield S. Concurrent partnerships and HIV prevalence disparities by race: linking science and public health practice. *American Journal of Public Health* 2009; **99**(6):1023–1031, doi:10.2105/AJPH.2008.147835.

[17] Mercer CH, Aicken CRH, Tanton C, Estcourt CS, Brook MG, Keane F, Cassell JA. Serial monogamy and biologic concurrency: measurement of the gaps between sexual partners to inform targeted strategies. *American Journal of Epidemiology* 2013; **178**(2):249–259, doi:10.1093/aje/kws467.

[18] Epstein H. The mathematics of concurrent partnerships and HIV: a commentary on Lurie and Rosenthal, 2009. *AIDS Behavior* 2010; **14**:29–30, doi:10.1007/s10461-009-9627-x.

[19] Morris M. Barking up the wrong evidence tree. Comment on Lurie and Rosenthal, concurrent partnerships as a driver of the HIV epidemic in Sub-Saharan Africa? the evidence is limited. *AIDS Behavior* 2010; **14**:31–33, doi:10.1007/s10461-009-9639-6.

[20] Lurie MN, Rosenthal S. Concurrent partnerships as a driver of the HIV epidemic in Sub-Saharan Africa? the evidence is limited. *AIDS Behavior* 2010; **14**:17–24, doi:10.1007/s10461-009-9583-5.

[21] Lurie MN, Rosenthal S. The concurrency hypothesis in Sub-Saharan Africa: convincing empirical evidence is still lacking. Response to Mah and Halperin, Epstein, and Morris. *AIDS Behavior* 2010; **14**:34–37, doi:10.1007/s10461-009-9640-0.

[22] Mah TL, Halperin DT. The evidence for the role of concurrent partnerships in Africa's HIV epidemics: a response to Lurie and Rosenthal. *AIDS Behavior* 2010; **14**:25–28, doi:10.1007/s10461-009-9617-z.

[23] Kalbfleisch JD, Lawless JF. The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association* 1985; **80**(392):863–871.

[24] Hwang W, Brookmeyer R. Design of panel studies for disease progression with multiple stages. *Lifetime Data Analysis* 2003; **9**:261–274.

[25] Joly P, Commenges D, Helmer C, Letenneur L. A penalized likelihood approach for an illness-death model with interval-censored data: application to age-specific incidence of dementia. *Biostatistics* 2002; **3**(3):433–443.

[26] Lawless JF, Rad NN. Estimation and assessment of Markov multistate models with intermittent observations on individuals. *Lifetime Data Analysis* 2015; **21**:160–179.

[27] Kvist K, Andersen PK, Angst J, Kessing LV. Event dependent sampling of recurrent events. *Lifetime Data Analysis* 2010; **16**:580–598.

[28] Cook RJ, Lawless JF. Statistical issues in modeling chronic disease in cohort studies. *Statistics in Biosciences* 2014; **6**:127–161.

[29] Mandel M, Betensky RA. Estimating time-to-event from longitudinal ordinal data using random-effects Markov models: application to multiple sclerosis progression. *Biostatistics* 2008; **9**(4):750–764.

[30] Hui SL, Zhou XH. Evaluation of diagnostic tests without a gold standard. *Statistical Methods in Medical Research* 1998; **7**:354–370.

[31] Cox DR, Miller HD. *The theory of stochastic processes*, vol. 134. CRC Press, 1977.

[32] Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society Series B Methodological* 1972; **34**:187–220.

[33] Goodreau SM, Cassels S, Kasprzyk D, Montano DE, Greek A, Morris M. Concurrent partnerships, acute infection and HIV epidemic dynamics among young adults in Zimbabwe. *AIDS Behavior* 2012; **16**:312–322, doi:10.1007/s10461-010-9858-x.

[34] Rhee A. An agent-based approach to HIV/AIDS modelling: A case study of Papua New Guinea. Master's Thesis, Massachusetts Institute of Technology 2006.

[35] Heuveline P, Sallach D, Howe T. The structure of an epidemic: modelling AIDS transmission in Southern Africa. *Symposium on Agent-Based Computational Modelling*, Vienna, Austria, 2003.

[36] Brookmeyer R, Boren D, Baral SD, Bekker L, Phaswana-Mafuya N, Beyrer C, Sullivan PS. Combination HIV prevention among MSM in South Africa: results from agent-based modeling. *PLoS ONE* 2014; **9**(11):e112 668, doi:10.1371/journal.pone.0112668.

[37] Mah TL, Haperin DT. Concurrent sexual partnerships and the HIV epidemics in Africa: evidence to move forward. *AIDS Behavior* 2010; **14**:11–16, doi:10.1007/s10461-008-9433-x.

[38] Morris M, M K. A microsimulation study of the effect of concurrent partnerships on the spread of HIV in Uganda. *Mathematical Population Studies* 2000; **8**:109–133, doi: 10.1080/08898480009525478.

[39] Morris M, Epstein H, Wawer M. Timing is everything: international variations in historical sexual partnership concurrency and HIV prevalence. *PLoS ONE* 2010; **5**(11):e14 092, doi:10.1371/journal.pone.0014092.

[40] Glynn JR, Dube A, Kayuni N, Floyd S, Molesworth A, Parrott F, French N, Crampin AC. Measuring concurrency: an empirical study of different methods in a large population-based survey and evaluation of the UNAIDS guidelines. *AIDS* 2012; **26**:997–985, doi:10.1097/QAD.0b013e328350fc1f.

[41] UNAIDS Reference Group on Estimates, Modelling, and Projections: Working Group on Measuring Concurrent Sexual Partnerships. HIV: Consensus indicators are needed for concurrency. *Lancet* 2014; **375**:621–622, doi:10.1016/S0140-6736(09)62040-7.

[42] Eaton JW, McGrath N, Newell M. Unpacking the recommended indicator for concurrent sexual partnerships. *AIDS* 2012; **26**(8):1037–1039, doi:10.1097/QAD.0b013e328351f726.

[43] Commenges D. Multi-state models in epidemiology. *Lifetime Data Analysis* 1999; **5**:315–327, doi:10.1023/A:1009636125294.

[44] Andersen PK, Keiding N. Multi-state models for event history analysis. *Statistical Methods in Medical Research* 2002; **11**:91–115, doi:10.1191/0962280202SM276ra.

[45] Lange JM, Hubbard RA, Inoue LYT, Minin VN. A joint model for multistate disease processes and random informative observation times, with applications to electronic medical records data. *Biometrics* 2015; **71**:90–101.

[46] Kapetanakis V, Matthews FE, Van Den Hout A. A semi-Markov model for stroke with piecewise-constant hazards in the presence of left, right and interval censoring. *Statistics in Medicine* 2013; **32**:697–713.

[47] Fischer W, Meier-Hellstern K. The Markov-modulated Poisson process (MMPP) cookbook. *Performance Evaluation* 1992; **18**:149–171, doi:10.1016/0166-5316(93)90035-S.

[48] Bocharov PP, D'Apice C, Pechinkin AV, Salerno S. *Queueing Theory*. Walter de Gruyter: New York, 2003.

[49] Fearnhead P, Sherlock C. An exact Gibbs sampler for the Markov-modulated Poisson process. *Journal of the Royal Statistical Society: Series B* 2006; **68**:767–784, doi:10.1111/j.1467-9868.2006.00566.x.

[50] Efron B. Bootstrap methods: another look at the jackknife. *Annals of Statistics* 1979; **7**:1–26.

[51] Efron B. Missing data, imputation and the bootstrap. *Journal of the American Statistical Association* 1994; **89**:972–985.

[52] Gorbach PM, Javanbakht M, Bolan R. Behavior change following HIV diagnosis: findings from a cohort of Los Angeles MSM. *Oral Abstract 2347, AIDS Impact 2015, The 12th International Conference*, Amsterdam, Netherlands, 2015.

[53] Klein JP, Moeschberger ML. *Survival Analysis: Techniques for Censored and Truncated Data*. 3 edn., Springer: New York, 2010.

[54] Goodreau SM, Carnegie NB, Vittinghoff E, Lama JR, Sanchez J, Grinsztejn B, Koblin BA, Mayer KH, Buchbinder SP. What drives the US and Peruvian HIV epidemics in men who have sex with men (MSM)? *PloS ONE* 2012; **7**(11):e50 522, doi: 10.1371/journal.pone.0050522.

[55] Gorbach PM, M J, Kornbleth L, Bolan R, Blum ML. Behaviors associated with transmitted drug resistant HIV: Findings from a cohort of Los Angeles MSM with new HIV diagnosis. *JAIDS* In Review; .

[56] Gentleman RC, Lawless JF, Lindsey JC, Yan P. Multi-state Markov models for analysing incomplete disease history data with illustrations for HIV disease. *Statistics in Medicine* 1994; **13**:805–821.

[57] Kang M, Lagakos SW. Statistical methods for panel data from a semi-Markov process, with application to HPV. *Biostatistics* 2007; **8**(2):252–264.

[58] Cai L, Schenker N, Lubitz J. Analysis of functional status transitions by using a semi-Markov process model in the presence of left-censored spells. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2006; **55**(4):477–491.

[59] Crimmins EM, Saito Y. Getting better and getting worse: transitions in functional status among older Americans. *Journal of Aging and Health* 1993; **5**:3–36.

[60] Wei S, Kryscio RJ. Semi-Markov models for interval censored transient cognitive states with back transitions and a competing risk. *Statistical Methods in Medical Research* 2014; .

[61] Lange JM, Minin VN. Fitting and interpreting continuous-time latent Markov models for panel data. *Statistics in Medicine* 2013; **32**:4581–4595.

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

[63] Titman AC, Sharples LD. Semi-Markov models with phase-type sojourn distributions. *Biometrics* 2010; **66**:742–752.

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

[65] Orchard T, Woodbury MA. A missing information principle: theory and applications. *Proceedings of the 6th Berkeley Symposium on mathematical statistics and probability*, vol. 1, University of California Press Berkeley, CA, 1972; 697–715.

[66] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B* 1977; **39**:1–38.

[67] Wei GCG, Tanner MA. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association* 1990; **85**(411):699–704.

[68] McCulloch CE. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association* 1997; **92**(437):162–170.

[69] Booth JG, Hobert JP, Jank W. A survey of Monte Carlo algorithms for maximizing the likelihood of a two-stage hierarchical model. *Statistical Modelling* 2001; **1**:333–349.

[70] Fort G, Moulines E. Convergence of the Monte Carlo expectation maximization for curved exponential families. *The Annals of Statistics* 2003; **31**(4):1220–1259.

[71] Chan KS, Ledolter J. Monte Carlo EM for time series models involving counts. *Journal of the American Statistical Association* 1995; **90**(429):242–252.

[72] Booth JG, Hobert JP. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 1999; **61**(1):265–285.

[73] Jank W. *The EM algorithm, its randomized implementation and global optimization: some challenges and opportunities for operations research*. Springer: New York, NY, USA, 2006.

[74] Snowdon DA, Kemper SJ, Mortimer JA, Greiner LH, Wekstein DR, Markesbery WR. Linguistic ability in early life and cognitive function and alzheimer's disease in late life: findings from the Nun Study. *JAMA* 1996; **275**:528–532.

[75] Snowdon DA. Healthy aging and dementia: findings from the Nun Study. *Annals of Internal Medicine* 2003; **139**(5):450–454.

[76] Wei S. Multi-state models for interval censored data with competing risks. PhD Thesis, University of Kentucky 2015. Paper 10.

[77] Kryscio RJ, Abner EL, Lin Y, Cooper GE, Fardo DW, Jicha GA, Nelson PT, Smith CD, Van Eldik LJ, Wan L, *et al.*. Adjusting for mortality when identifying risk factors for transitions to MCI and dementia. *Journal of Alzheimers Disease* 2013; **35**(4):823–832.

[78] Kryscio RJ, Abner EL. Are Markov and semi-Markov models flexible enough for cognitive panel data. *Biometrics and Biostatistics* 2013; **4**(1):1–2.

[79] Satten GA, Sternberg MR. Fitting semi-Markov models to interval-censored data with unknown initiation times. *Biometrics* 1999; **55**(2):507–513.

[80] Van Den Hout A, Matthews FE. Multi-state analysis of cognitive ability data: a piecewise-constant model and a Weibull model. *Statistics in Medicine* 2008; **27**:5440–5455.

[81] Van Den Hout A, Jagger C. Estimating life expectancy in health and ill health by using a hidden Markov model. *Journal of the Royal Statistical Society Series C Applied Statistics* 2009; **58**:449–465.

[82] Benoit JS, Chan W, Luo S, Yeh HW, Doody R. A hidden Markov model approach to analyze ternary outcomes when some observed states are possibly misclassified. *Statistics in Medicine* 2016; **35**:1549–1557.

[83] Bureau A, Shiboski S, Hughes JP. Applications of continuous time hidden Markov models to the study of misclassified disease outcomes. *Statistics in Medicine* 2003; **22**:441–462.

[84] Jackson CH, Sharples LD. Hidden Markov models for the onset and progression of bronchiolitis obliterans syndrome in lung transplant recipients. *Statistics in Medicine* 2002; **21**:113–128.

[85] Yu SZ. Hidden semi-Markov models. *Artificial Intelligence* 2010; **174**:215–243.

[86] Vaseghi SV. State duration modelling in hidden Markov models. *Signal Processing* 1995; **41**:31–41.

[87] Klekociuk SZ, Summers JJ, Vickers JC, Summers MJ. Reducing false positive diagnoses in mild cognitive impairment: the importance of comprehensive neuropsychological assessment. *European Journal of Neurology* 2014; **21**:1330–1336.

[88] de Rotrou J, Wenisch E, Chausson C, Dray F, Faucounau V, Rigaud AS. Accidental MCI in healthy subjects: a prospective longitudinal study. *European Journal of Neurology* 2005; **12**:879–885.

[89] Brooks BL, Iverson GL, Holdnack JA, Feldman HH. Potential for misclassification of mild cognitive impairment: a study of memory scores on the Wechsler Memory Scale - III in healthy older adults. *Journal of the International Neuropsychological Society* 2008; **14**:463–478.

[90] Bennett DA. Update on mild cognitive impairment. *Current Neurology and Neuroscience Reports* 2003; **3**:379–384.

[91] Bruscoli M, Lovestone S. Is MCI really just early dementia? a systematic review of conversion studies. *International Psychogeriatrics* 2004; **16**(2):129–140.

[92] DeCarli C. Mild cognitive impairment: prevalence, prognosis, aetiology, and treatment. *Lancet Neurology* 2003; **2**:15–21.

[93] Mitchell AJ. A meta-analysis of the accuracy of the mini-mental state examination in the detection of dementia and mild cognitive impairment. *Journal of Psychiatric Research* 2009; **43**:411–431.