Multistate Models for Biomarker Processes

by

Narges Nazeri Rad

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2014

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Multistate models are widely used for describing life history processes. In studies where individuals are observed continuously, the transition times between states are known exactly. However, when individuals are observed intermittently, transition times and even the states visited between successive observations, may be unknown. Irregular intermittent observation is a special case of intermittent observation where the observation times vary across individuals.

In the case of intermittent observation, we may not be able to estimate model parameters precisely. In the first part of the thesis, we review methods of estimation for Markov models in this situation, and provide a numerical study that shows the loss of efficiency in estimation for intermittent observation compared to continuous observation in both progressive and bi-directional multistate models. Then, application to data from the CANOC, Canadian Observational Cohort study of HIV-positive individuals whose virus has been suppressed by combination antiretroviral therapy, illustrates the effect of gap times on estimation efficiency.

Irregular observation is very common in longitudinal data on disease history of individuals in observational studies. However, there are considerable challenges in checking models with these observation schemes, since there is a strong possibility that this irregularity may be induced by the dependency of inter-visit times on previous process history. As a result, followup visits from this kind of data are subject to disease state-dependency, which needs to be taken into account to prevent biased analysis. The second part of this thesis begins with a review on the estimation of marginal process features such as failure time distributions and prevalence probabilities in the context of Markov multistate models with intermittent observations. A method for estimation of these features is developed using Inverse Intensity Weights (IIW). This method corrects the estimation bias due to dependent observation times. Simulation studies illustrate that the proposed method yields estimates that are close to the true values, while the method that ignores the dependency yields estimates that differ substantially from the true values. Then, an application involving viral load dynamics in a group of individuals from the CANOC study is presented.

In practice, we may want to consider models for which transition intensities depend on internal covariates related to previous process history. There are, however, challenges in fitting and checking models involving internal covariates, and in making predictions. In the third part of this thesis, we have developed an algorithm that simulates possible sample paths of individuals' processes, and we use it for prediction and model checking.

Finally, there has been recent discussion of model assessment of multistate models. There remain, however, some difficulties in model assessment with irregular intermittent observations. The last part of this thesis addresses problems that arise with methods based on comparison of empirical and model-based estimates. We propose the use of likelihood ratio tests within the Markov process family, and methods of estimating the power of these tests are given. We also propose a method for comparing models based on different outcome spaces in terms of prediction. Finally, the proposed methods are applied to a group of individuals in the CANOC study.

Acknowledgements

I would like to express my deep and sincere gratitude to my supervisor, Dr. Jerry Lawless. It would not have been possible to write this thesis without his motivation, enthusiasm, patience, encouragement, and continuous support. It was my honor to study with prof. Lawless and have him as my supervisor.

I also extend sincere thanks to my thesis committee, Dr. Richard J. Cook, Dr. Cecilia A. Cotton, Dr. Brian D.M. Tom, and Dr. Suzanne L. Tyas for their helpful suggestions and insightful comments.

I would like to express my gratitude to all faculty members and staff. Especially, I would like to thank Ms. Mary Lou Dufton for her patience with addressing my different questions. I am thankful to Ker-Ai Lee for sharing her comments and helping me with the statistical computing.

My gratitude also goes to Dr. Janet Raboud who kindly provided the CANOC data and shared her experience on this data.

I would like to thank my friends in Waterloo. Especially, Liqun Diao, Meaghan Knight, Nathalie Moon, and Atefeh Zarabadi.

Last but not least, I want to thank my loving and caring family. I am truely grateful to my parents, Soroor Pandidan and Hossein Nazeri Rad for their endless love and support without which I could not get this far. I am also thankful to my wonderful siblings Niloofar, Mohammad, and Maryam for their emotional support and encouragement. Finally, words cannot express my loving gratitude to my husband, Reza Noroozi. This thesis would not have been possible without his support, unwavering love, remarkable patience, and constant encouragement during my PhD journey. I consider myself the luckiest in the

world for having such a supportive family.

To my caring parents, Soroor and Hossein

To the love of my life, Reza

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

A longitudinal study in biomedical or public health research (e.g. clinical trials, observational studies, etc) is one for which participants are followed through time, rather than observed at only one time. By following subjects, investigators can study processes involving features such as (a) times to events (b) time-varying marker processes, and (c) other covariates. A familiar setting is that of persons infected with the human immune deficiency virus (HIV), where covariates, including treatment assignment, demographic information, and physiologic characteristics may be recorded on individuals in a cohort at a baseline time, and measures of immunologic and virologic status such as CD4 count and viral load copy number are biomarkers that are measured at subsequent clinical visits. Time to progression to AIDS or death and recurrent events such as virologic rebounds are also recorded for each participant. Higher CD4 cell counts, lower viral load copy numbers, and a lower rate of virologic rebounds are associated with better survival and fewer

AIDS-related illnesses for HIV- positive patients.

Sometimes events are defined in terms of biomarkers $Y(t)$. An example is in HIV studies where a Viral Blip event (e.g. Grennan et al. (2012)) is defined as a viral load value greater than 50 copies/ mL and less than 1000 copies/ mL, preceded and followed (at consecutive visits) by an undetectable value ($<$ 50 copies/ mL); another event type in these studies is Virologic Rebound which can be defined as either (1) viral load greater than 50 copies/ mL at two consecutive visits at least 30 days apart, or (2) a viral load greater than 1000 copies/ mL.

For convenience, we assume here that individuals are followed from a specified time t=0. Data at time t for each subject $i = 1, ..., n$ is ideally of the form $\{N_i(t), Y_i(t), Z_i(t)\}$ where $N_i(t)$ is a vector of numbers of events of different types occurring over the time interval $[0, t]$, that is $N_i(t) = \{N_{i1}(t), ..., N_{ik}(t)\}$ where k is the number of event types. $\{Y_i(t), t \geq 0\}$ is a longitudinal biomarker trajectory, and $Z_i(t)$ is a vector of external covariates which may be fixed or time dependent. The time origin t=0 is designated according to the processes under study. The observed biomarkers are frequently important health indicators that represent the progression or status of disease. We also might be interested in events like death or drop out from the study, or single or recurrent events such as symptoms of disease, or other disease related events.

Usually observation of individuals is done at intermittent times, in other words, we observe biomarker measures $Y_i(t_{ij})$ at t_{ij} for individual i, where $i = 1, ..., n$ and $j = 1, ..., m_i$. We also usually observe the external covariate $Z(t_{ij})$ at these time points. The event times might be known exactly or interval censored.

Common problems encountered with biomarker data are measurement error, missing data due to the measurement of markers only periodically, process-dependent observation times, and bias due to process-dependent dropout. Measurement errors can be thought of as consisting of both laboratory error and short term biological variability. Similar problems can occur with external covariates. The error in measuring biomarkers or covariates may lead to biased estimation of regression parameters that describe the relationship between event intensities and the true marker or covariate values. Event times may also be measured with error in some settings, or they may be known only to lie in some specified time intervals. In this thesis we deal with missing information, but do not address measurement error; that is, the raw measurements are used as the basis for modelling.

Joint analysis of survival or other events with either repeated measures or recurrent events has been increasingly common in medical areas, for example in HIV or acquired immunodeficiency syndrome (AIDS) studies. Faucett and Thomas (1996), Hogan and Laird (1997 a, b, 1998), Wulfsohn and Tsiatis (1997), Henderson et al. (2000), Xu and Zeger (2001), Tsiatis and Davidian (2001), Yu et al. (2004), Chi and Ibrahim (2006), Borgan et al. (2007), Diggle et al. (2008), and others have discussed joint models for repeated measures and survival times. Tsiatis and Davidian (2004) provide a thorough review on joint models for covariate processes and single survival times or recurrent events. Often in medical studies where the primary interest is in the time to some clinical event, patients are also monitored longitudinally with respect to some biological measurements, or biomarkers. These measurements are important because the trajectory of marker processes may be predictive of times to event occurrence, and then we wish to assess the relationship with event occurrence. Goals in such studies include investigation of the within-subject patterns of change of markers, the effects of treatment and risk factors on the marker processes and on disease-related events or death, the relationship of markers and events, and sometimes whether the biomarkers can be used as surrogate endpoints in a clinical trial. Sometimes markers can also help adjust for dropouts in a study. In AIDS studies, an example of a marker-related dropout process is that people who have faster rates of decline in CD4 counts may be more likely to drop out of the study. Such "dependent" dropout phenomena can lead to biased estimates of process parameters if not handled properly. As well as correcting biases, joint modelling of biomarkers and event occurrence can improve the efficiency of parameter estimates in either part of the model because extra information is being used.

In this thesis, we model a biomarker process and any clinical events of interest jointly by means of multistate models. This approach has been taken by many previous authors (e.g. Gentleman et al. (1994), Copas and Farewell (2001), O'Keeffe et al. (2011), Tom and Farewell (2011)). We discuss several issues in the context of multistate models with intermittent observations. Some of these issues include investigation of the loss of information in estimation when individuals are observed intermittently, challenges in fitting and checking models incorporating internal covariates, and adjusting the estimation of some marginal process features to accomodate dependent observation times. Motivated by studies of viral load dynamics for HIV-positive individuals in the CANOC study, we will focus mainly on models for biomarkers.

The remainder of this Chapter is organized as follows, Section 1.2 reviews intensity-based event history models. Section 1.3 focuses on multistate models, and discusses families of models. Section 1.4 reviews models for biomarker processes and discusses some advantages and disadvantages of multistate models. Section 1.5 contains a motivating example based on the CANOC observational study, and Section 1.6 provides an outline of the rest of the thesis.

1.2 Event history models

Data from a marker process $\{Y(t), t \geq 0\}$ is usually a sequence of repeated measurements which are taken on each individual or subject at several occasions. A single counting process $\{N(t), t \geq 0\}$ records the cumulative number of events of some type up to time t; $N(t) = \sum_{k=1}^{\infty} I(T_k \leq t)$ is the number of events occurring over the time interval $[0, t]$, where $T_1 < T_2$... denote the event times. The individual event processes are considered to start at time t=0, but in some settings, an individual is sampled and observation begins from a time $\tau_0 > 0$, with events observed over the time interval $[\tau_0, \tau]$ where τ is the end of followup time.

1.2.1 Intensity functions

Models for events in continuous time can be specified by the probability of an event in a short interval $[t, t + \Delta t)$, given the history of event occurrence before time t. Let $\Delta N(t) = N(t + \Delta t^-) - N(t^-)$ denote the number of events in the interval $[t, t + \Delta t)$, and $H(t) = \{N(s), 0 \leq s < t\}$ denote the history of the process at time t. Then, assuming that two or more events can not occur at the same time, the event intensity function is given by Andersen et al. (1993); Aalen et al. (2008); Cook and Lawless (2007) as

$$\lambda(t|H(t)) = \lim_{\Delta t \to 0} \frac{P\{\Delta N(t) = 1|H(t)\}}{\Delta t};$$

with J types of events, we can define the intensity function for a type j event as

$$\lambda_j(t|H(t)) = \lim_{\Delta t \to 0} \frac{P\{\Delta N_j(t) = 1|H(t)\}}{\Delta t}, \quad j = 1, ..., J.$$

In this case, $H(t)$ represents the history of all events up to time t, so the previous occurrence of one type of event can influence the occurrence of another event type.

1.2.2 Intensity function with covariates

With external covariates $Z(t)$, we can consider the intensity as being conditioned on the whole covariate history $\overline{Z}(\infty)$ where $\overline{Z}(t) = \{Z(s), 0 \leq s \leq t\}$. Then it is assumed that $\lambda(t|H(t), \overline{Z}(\infty)) = \lambda(t|H(t), \overline{Z}(t))$. If $Z(t)$ is an internal covariate, we can still define intensity function $\lambda(t|H(t), \overline{Z}(t))$; however, it is not generally equivalent to $\lambda(t|H(t), \overline{Z}(\infty))$. A widely used model is the multiplicative intensity model. This encompasses the famous proportional intensity model, or the Cox model. We consider Cox models where, the intensity is of the form

$$\lambda(t|H(t), \overline{Z}(t)) = \lambda_0(t) exp\{Z'(t)\beta\},$$

where β is a vector of regression coefficients, and $\lambda_0(t)$ is the baseline intensity function. The Cox model was introduced by Cox (1972) in the context of survival data. A natural extension of the model to accommodate time-varying covariate effects is

$$\lambda(t|H(t), \overline{Z}(t)) = \lambda_0(t) exp\{Z'(t)\beta(t)\},$$

where $\beta(t)$ is a time-varying regression coefficient.

In some cases, it may be more appropriate to use models where the effect of covariates is modelled on an additive scale. One additive model assumes that the intensity is of the form

$$\lambda(t|H(t), \overline{Z}(t)) = \lambda_0(t) + Z'(t)\beta.$$

The more general model was introduced by Aalen (1980), where

$$\lambda(t|H(t), \overline{Z}(t)) = \lambda_0(t) + Z'(t)\beta(t).$$

1.3 Multistate models

The natural history of a chronic disease can often be expressed in terms of distinct health stages or biomarkers which are health status indicators, and multistate models have been widely used in this situation. For example, O'Keeffe et al. (2011) used it in a study of psoriatic arthritis, Gentleman et al. (1994) in an analysis of HIV-infected patients, and Yu et al. (2008) in a study of cognitive impairment. Here we discuss continuous time multistate models. Suppose individuals independently move among R states, denoted by $r = 1, ..., R$, and let $Y_i(t)$ denote the state occupied by individual $i(i = 1, ..., n)$ at time $t \geq 0$. Then the transition intensity from state r to s for $r \neq s$ is

$$\lambda_{rs}(t|H(t)) = \lim_{\Delta t \to 0} \frac{P\{(Y(t + \Delta t^-) = s|Y(t^-) = r, H(t))\}}{\Delta t}, \quad r \neq s,$$

where $H(t) = \{(Y(u), Z(u)), 0 < u < t\}$. The transition probabilities for such models are defined by

$$P(Y(t+u) = s|Y(t) = r, H(t)) = P_{rs}(t, t+u|H(t)), \quad r, s = 1, ..., R.$$

These are of interest for prediction.

1.3.1 Markov Models

In Markov multistate models, transition intensities are of the form $\lambda_{rs}(t|H(t)) = Y_r(t)q_{rs}(t)$, where $Y_r(t) = I(Y(t^-) = r)$ indicates that an individual is at risk of transition from state r at time t. We define the $R \times R$ transition intensity matrix $Q(t)$ with entries $q_{rs}(t)$ for $r \neq s$, and $q_{rr}(t) = -\sum_{r \neq s} q_{rs}(t)$. The $R \times R$ transition probability matrix $P(u, u+t)$ in this case does not depend on $H(u)$ and has entries $P_{rs}(u, u+t) = P(Y(u+t) = s|Y(u) = r)$.

If none of the transition intensities $q_{rs}(t)$ depends on time, the model is called time homogeneous (stationary). In a time homogeneous Markov model with transition intensity matrix Q, the transition probability matrix $P(u, u+t)$ equals $P(0,t)$, which we denote for simplicity as $P(t)$. If the elements $q_{rs} = q_{rs}(\theta)$ depend on b independent parameters $\theta_1, ..., \theta_b$ with $\theta = (\theta_1, ..., \theta_b)$, then we have (e.g. Kalbfleisch and Lawless (1985))

$$P(t;\theta) = exp\{Q(\theta)t\} = \sum_{l=0}^{\infty} \frac{Q^l(\theta)t^l}{l!}.$$

Kalbfleisch and Lawless (1985) discussed efficient means of computing the transition intensities via a canonical decomposition. When Q is diagonazable, that is Q has R linearly independent eigenvectors and distinct eigenvalues $d_1, ..., d_R$, let A be the $R \times R$ matrix whose jth column is a right eigenvector corresponding to d_j. Then $Q = ADA^{-1}$, where $D = diag(d_1, ..., d_R)$, and

$$P(t) = Adiag(e^{d_1 t}, ..., e^{d_R t})A^{-1}.$$

For some simple time-dependent Markov models, the transition probabilities can be obtained analytically. This has been implemented in the msm package in R (Jackson (2011)). More generally, however, we would need to use numerical methods of some type to obtain transition probabilities from the transition intensities. A simple approach is to use the product integral idea (e.g. Aalen et al. (2008), p. 460). The product integral representation of $P(s,t)$ is given by

$$P(u,t) = \prod_{v \in (u,t]} [I + Q(v)dv],$$

where I is a $R \times R$ identity matrix. Now, in order to approximate $P(u,t)$, we must partition $(u,t]$ into a sequence of intervals $(v_{l-1}, v_l]$ with $u = v_0 < v_1 < ... < v_m = t$, such that to a close approximation $q_{ij}(v) = q_{ij}^{(l)} = \frac{q_{ij}(v_{l-1}) + q_{ij}(v_l)}{2}$ for $v_{l-1} < v \leq v_l$. Then we have

$$P(u,t) \approx \prod_{l=1}^{m} [I + Q^{(l)} \Delta v_l] \approx \prod_{l=1}^{m} exp[Q^{(l)} \Delta v_l],$$

where $\Delta v_l = v_l - v_{l-1}$ and $Q^{(l)} = (q_{ij}^{(l)})$. Titman (2011) gives another numerical approach in terms of the Kolmogorov differential equations for the process.

We often wish to let the transition intensities depend on covariates Z_i or $Z_i(t)$. As described in Section 1.2.2, this can be done in various ways. However, multiplicative models where $q_{rs}(t|H(t), \bar{Z}(t)) = q_{0rs}(t)exp\{Z'(t)\beta_{rs}\}$ are used most often. We note that there is in general a separate baseline intensity and separate regression parameter β_{rs} for each type of transition.

1.3.2 Semi-Markov models

In Semi-Markov models, we assume that the transition intensity functions are related to the time $B(t)$ since entry to the current state $Y(t)$. That is,

$$\lambda_{rs}(t|H(t)) = Y_r(t)h_{rs}(B(t)), r \neq s,$$

where $h_{rs}(u)$ is a non-negative function. In many situations, it is hard to compute transition probabilities for a Semi-Markov model. As a result, they are not easily handled when observation of individuals is intermittent and we do not consider them further in the thesis. However, see Titman and Sharples (2010b) for more details.

1.3.3 Some pros and cons of multistate models

Multistate models help to provide a comprehensive view of a disease process and they allow the estimation of features such as proportions of individuals in different states at specific times in the future, transition rates, and the distribution of time to first entry to a given state.

Multistate models can also have an advantage over time to event models in survival analysis, for events associated with disease processes. For instance, we can study effects of covariates on the various transition intensities, and gain biological insights which may be ignored when using only a time to event model. In addition, multistate models can limit the effects of process-dependent observation times. As an example, we refer to Gentleman et al. (1994), where they considered a multistate model for the progression of AIDS disease in a group of HIV-positive individuals. They expressed AIDS as an absorbing state, and defined three transient states based on CD4 cell count values. The event of main interest was the entry to the AIDS state, and we could look at this problem from the survival point of view. However, the multistate model with the four states gives us the opportunity to gain a better understanding of the disease process.

The msm package, (Jackson, 2011), has made fitting the multistate models more convenient, but still there are issues that should be mentioned. First, sometimes it is crucial how we discretize a process to a finite number of states. It is important that we define states that are clinically useful, and which are not too susceptible to measurement (misclassification) errors, but there are usually different choices of states that would be sensible. A second issue is how to deal with misclassification errors (i.e. errors in the determination of what state an individual is in at a given time). Jackson et al. (2003) discuss methods to deal with this. The third issue concerns settings where individuals are seen only at discrete times, and their states are only known at these times. In this situation, we do not have full information at times between observations, such as exact transition times, states visited between observation times, and covariate values between visit times. Fitting and assessing multistate models are often challenging, and this is the basis for the research in this thesis.

1.4 Models for biomarker processes

In medicine, a biomarker is an internal covariate that exists only while the subject is alive. A biomarker often indicates the progress or status of a disease in some sense, and their relationship to clinical events is often of interest. In addition, often events are defined in terms of biomarkers.

Two examples of biomarkers are as follows. The first example involves a prostate cancer study (Proust-Lima and Taylor, 2009) where the prostate-specific antigen (PSA) is used as a biomarker for persons who have been treated for prostate cancer, and clinical events of interest include distant metastases, local recurrence, and death. A second example is taken from Liu and Huang (2009), concerning patients with acquired immune deficiency syndrome (AIDS) in which CD4 cell counts are the repeated biomarker measures, and the occurrence of opportunistic diseases are clinical events of interest.

In longitudinal studies, two types of biomarker models are common. The first type is where $Y(t)$ is continuous and stochastic processes such as Gamma or Gaussian process are used for $\{Y(t), t > 0\}$. Tsiatis and Davidian (2001), Xu and Zeger (2001), Liu and Huang (2009), Proust-Lima and Taylor (2009), and many others have considered these types of models. Another type of modelling is where the range of $Y(t)$ is partitioned into a set of R states. For example, Gentleman et al. (1994), Copas and Farewell (2001), Yu et al. (2008), O'Keeffe et al. (2011), Tom and Farewell (2011) have studied this.

There are a number of situations where it is important to consider models for biomarkers and events together, as opposed to cases where only a model for the biomarker process, $Y(t)$, is needed, or to cases where only a model for the transition intensity of the event process, $\lambda(t|\bar{Z}(t), H(t))$, is needed. First, suppose that we focus on modelling the event process. Considering the biomarker process as a time dependent covariate in the event intensity may require the biomarker value at any time t, but the biomarker is usually observed on each individual only intermittently. Particularly, when we have irregular, long gap times in the biomarker measurements, it is required to model biomarker processes to deal with the biomarker's missing values. Second, suppose we are interested in the biomarker processes themselves. For inference on biomarker processes, ignoring the event process can sometimes lead to biased estimation. This bias might be due to informative observation times or censoring which is induced by the clinical events. For example, individuals with more serious HIV disease may be more likely to withdraw from a study or have long gaps between observations than healthier individuals, and this leads to fewer and more widely spaced viral load measurements.

As noted in Section 1.1, a multistate model is one approach for considering event times along with a biomarker process. Other approaches have also been taken (e.g. see Tsiatis and Davidian (2004)), but in this thesis, we focus on multistate models.

1.5 A motivating study: CANOC

CANOC is a Canadian cohort study of antiretroviral naive HIV-positive patients initiating combination antiretroviral therapy (cART) since January 1st 2000 (Raboud et al. (2010), Palmer et al. (2011)). Eligibility criteria for inclusion in this cohort is documented HIV infection, residence in Canada, minimum age of 18 years, initiation of a first antiretroviral therapy (ART) comprised of a minimum of three individual agents, and at least one measurement of HIV-1 RNA and CD4 cell count within six months of cART initiation. Biomarkers that are measured at visit times (approximately every 3 months) for each individual include CD4 cell counts, viral loads, and other measures such as blood lipid (Cholesterol) levels. Clinical events include AIDS defining illnesses (ADIs) such as Candidiasis of esophagus, bronchi, trachea, or lungs (CANDIDA), and other events such as the diagnosis of heart disease or cancer, and death.

In the CANOC data we will also consider issues such as (a) identifying covariates which are significantly related to the viral load biomarker, and assessing how the viral load process changes with time, (b) determining whether the risks of biomarker events are related to previous values and covariates, for instance, whether the occurrence of blips is related to virologic rebounds, (c) examining whether observation times for individuals are related to their biomarker histories, and (d) dealing with process dependent observation times.

1.6 Outline of the thesis

The remaining chapters of this thesis are organized as follows.

In Chapter 2, we review estimation procedures for panel data under Markov models. Estimability issues in getting the maximum likelihood estimates are discussed. The existing methods for assessing the model fit of multistate models are reviewed.

Chapter 3 examines estimation for progressive and bi-directional multistate processes with continuous and intermittent observations. The estimates of asymptotic covariance matrices for parameters of both situations are discussed. Numerical studies are conducted to investigate the effect of observation schedule (in particular, the number and closeness of visit times) on efficiency of estimation. We present applications involving an analysis of time to viral rebound in observational CANOC data.

In Chapter 4, we consider the estimation of prevalence probabilities and failure time distributions associated with multistate models. We introduce inverse intensity weighted (IIW) estimating equations to handle the dependency of inter-visit times on previous process history, and we propose adjusted prevalence estimates that accomodate dependent observation times. Simulations are conducted to evaluate the performance of the proposed estimates, and they show that the IIW prevalence estimates are close to the true prevalence values, whereas estimates that ignore the association between the gap times and previous process history show bias. We present an application to CANOC data, where we show that the gap times between visits depend on the previous viral load states that individuals were in.

Chapter 5 considers the assessment of multistate models with intermittent observations and possibly internal covariates. We present an algorithm for simulating sample paths of individuals' processes which facilitates estimation of the prevalence probabilities in the presence of internal covariates. We also investigate the robustness of prevalence estimates when the Markov assumption is violated, and discuss the implications for model assessment. Simulation studies are given, which demonstrate that the violation of the Markov assumption may induce only small biases in the estimation of prevalence probabilities. We discuss likelihood ratio tests for assessment of multistate models within the family of Markov processes, and give a way to estimate the effects of the gap times between visits on the power of such tests. Finally, we propose a method that allows comparison of models with different state spaces. Then, the proposed methods are applied to the observational CANOC study.

Finally, Chapter 6 briefly summarizes overall findings, and discusses potential areas for further study.

Chapter 2

Fitting and Assessing Multistate Models When Observation is Intermittent

2.1 Introduction

In observational studies, individuals are seen intermittently at followup visits, and clinical information about disease states and covariate values are recorded. However, information for times between visits is missing or incomplete. This is often referred to as panel data (e.g. Kalbfleisch and Lawless (1985), Hwang and Brookmeyer (2003)). In multistate models, the difficulties with intermittent observations are that the exact transition times, and sometimes the states visited between successive observation times are unknown. The intermittent observations can be classified as either balanced (regular) or unbalanced (irregular). In the regular case, individuals have a common set of visit times. Suppose that individual i is seen at times t_{ij}, where $i = 1, ..., n$ and $j = 0, 1, ..., m_i$. The regular observation scheme has $t_{ij} = t_j$ for $i = 1, ..., n$ and $j = 0, 1, ..., m$. Many researchers have studied Markov multistate models with intermittent observations, for example, Gentleman et al. (1994), Kvist et al. (2010), Mandel and Betenesky (2008), Titman and Sharples (2010a), and Cook and Lawless (2014). Titman and Sharples (2010b) and Yang and Nair (2011) have studied semi-Markov models in this context.

In longitudinal studies, we often face incomplete data that may be due to missed study assessments or premature loss to followup (LTF). Little and Rubin (1987) considered analysing missing data mechanisms, and classified them into three categories. The mechanism is called “Missing Completely at Random” (MCAR) if the missing data process does not depend on any variables of interest, and “Missing at Random” (MAR) if the missing data process is independent of the unobserved data, given the observed data. Finally, data are “Missing Not at Random” (MNAR) if the missing data process depends on unobserved data. Cook and Lawless (2014) discussed visit times and LTF in the case of intermittent observations. First, suppose the observation times for individual i are fixed values $t_0 < t_1 < ... < t_m$, and $Y_i(t)$ is the state occupied at time t. Let $Z(t)$ denote covariates. The history of the observed data at the observation time t_j is $H(t_j) = \{(Y(t_l), Z(t_l), t_l), l = 0, 1, 2, ..., j-1\}$. Let $R(t)$ be the missing visit indicator at time t, which equals 1 if the individual is observed at time t and 0 if $Y(t)$ is missing. Cook and Lawless (2014) classified LTF or missed visits as follows:

1. Sequentially Missing at Random (SMAR): Data are said to be SMAR if $R(t_j)$ is conditionally independent of $Y(t_l)$ and $Z(t_l)$ for $l \geq j$, given $H(t_j)$. If the SMAR condition does not hold, we may be able to introduce a vector of observed covariates, $Z^c(t)$, so that $R(t_j)$ becomes conditionally independent of $Y(t_l)$ and $Z(t_l)$ for $l \geq$

j, given $H(t_j)$ and $Z^c(t_j)$. In this situation, $Z^c(t)$ makes the SMAR assumption reasonable.

2. Missing Sequentially Not at Random (MSNAR): The missing data process is said to be MSNAR if it is not SMAR.

Cook and Lawless (2014) also investigated the case that observation times for individual i are random values $t_{i1}, ..., t_{im_i}$; this can occur, for example, when t_j is scheduled based on observed history up to time t_{j-1}. They defined the SMAR assumption as T_j is independent of $\{Y(t), Z(t), t > t_{j-1}\}$, given history to t_{j-1}.

This chapter reviews maximum likelihood estimation of Markov multistate model parameters based on intermittent observation of individuals. It also reviews methods for checking time homogeneity and other assumptions.

2.2 Fitting multistate models

We focus here on Markov models, first considering the case with no covariates. Suppose that individual i has observation times $t_{i0} < t_{i1} < ... < t_{im_i}$ that satisfy the SMAR requirement. Then, the likelihood function based on n independent individuals and under the Markov assumption is

$$L = \prod_{i=1}^{n} \prod_{j=1}^{m_i} P\{Y_i(t_{ij}) | Y_i(t_{ij-1}); \theta\}, \tag{2.1}$$

where θ is the vector of unknown transition intensity parameters. This likelihood is conditional on the state occupied by the individual at time t_{i0}. In many studies, all individuals have the same initial state at the time of entry into the study. For a time homogeneous model, we can compute the transition probabilities $P_{rs}(t_{i,j-1}, t_{ij}; \theta)$ by the matrix exponential function $exp(Q(t_{ij} - t_{i,j-1}))$ as discussed in Section 1.3.1.

In the case where there are fixed covariates, the likelihood function can be written as

$$L = \prod_{i=1}^{n} \prod_{j=1}^{m_i} P\{Y_i(t_{ij}) | Y_i(t_{i,j-1}), Z_i; \theta\}, \tag{2.2}$$

where Z_i is a vector of fixed covariates for individual i. We should remark that the computation of transition probabilities in this case requires separate calculations for each distinct covariate value. Therefore, if the number of distinct covariate values is large, maximum likelihood estimation may require a lot of computation. Considering time-varying covariates is more challenging, since the covariate values are only known at visit times. Cook and Lawless (2014) mention two methods for addressing this problem. First, we can use simplifying assumptions that would enable us to interpolate the covariate values at all times between visits. For instance, we can assume that the covariate values for each individual are constant between observation times. Then, the likelihood function is

$$L = \prod_{i=1}^{n} \prod_{j=1}^{m_i} P\{Y_i(t_{ij}) | Y_i(t_{i,j-1}), Z_i(t); \theta\}, \tag{2.3}$$

where $Z_i(t) = Z_{ij}$ for $t_{ij-1} < t \leq t_{ij}$. Note that the vector Z_{ij} is allowed to be a function of observed history $\{Y_i(t_{il}), Z_{il}, l = 0, 1, ..., j-1\}$. We will discuss other interpolation methods later in Section 2.3.2. Second, Cook and Lawless (2014) mention procedures for the joint modelling of the covariate process with the multistate process, and provide discussion of this. In this thesis we focus primarily on the first approach.

Jackson (2011) discussed algorithms for the computation and maximization of all the above likelihood functions for models with time homogeneous baseline intensities. This is implemented in the msm package in R. There are, however, still some challenges in fitting multistate models. First, when the times between observations are fairly large, or when there are many parameters in the model, then the likelihood can be “flat” in certain places, and msm often encounters convergence problems and fails to maximize the likelihood. The number of parameters in the models is increased when there are many covariates or types of transitions. Second, if for some intensities, there are no observed direct transitions in the original data, we may encounter estimability problems. A third difficulty is that the optimization algorithms in the msm package are very sensitive to the initial values that we give for covariate effects and baseline transition intensities, and apparent convergence to parameter vectors that do not maximize the likelihood is a common occurrence.

2.2.1 Time dependency

When the process is time homogeneous, the transition probabilities are obtained via a canonical decomposition as discussed in Section 1.3.1. However, the assumption of time homogeneity can be unsatisfactory. We mention two methods for checking this assumption. First, we can use the piecewise constant intensities models. In these models, we assume that for a set of cut points $0 = a_0 < a_1 < ... < a_K = \infty$ we have transition intensities

$$q_{rs}(t) = q_{rs}^{(l)} \quad a_{l-1} \leq t < a_l. \tag{2.4}$$

Piecewise constant transition intensities models are flexible models and can be implemented by the msm package. The time homogeneity assumption can be assessed by using a likelihood ratio test for a time homogeneous null against the piecewise constant model. We note, however, that the choice of the location and the number of cut points can be crucial. This method has been considered by Gentleman et al. (1994), among others.

A second method for checking departures from time homogeneity is to use local score tests. De Stavola (1988), Kalbfleisch and Lawless (1989), Gentleman et al. (1994), and Titman and Sharples (2010a) have discussed this approach. In this method, we consider the alternative hypotheses that the transition intensities $q_{rs}, r \neq s$ depend on time linearly or through a power relationship that is common across all transitions. That is, we consider $H_A : q_{rs}(t) = q_{rs} + t\gamma$ or $H_A : q_{rs}(t) = q_{rs}t^{\beta-1}$ and then we can test $H_0 : q_{rs}(t) = q_{rs}$ against these alternatives. The authors develop score tests which have the advantage that only the time homogeneous model needs to be fitted. In addition, we can apply this test separately to each transition intensity, to assess where the time homogeneity assumption might be violated.

2.3 Multistate models vs empirical estimates

One way that we can assess the adequacy of a multistate model is by comparing certain estimates based on it with empirical estimates that are valid more generally. Three types of model features are usually considered: (i) distributions of time to first entry in a state, (ii) prevalence probabilities $P_r(t) = P(Y(t) = r)$, and (iii) transition probabilities $P_{rs}(u, u + t) = P(Y(t + u) = s|Y(u) = r)$. We discuss (i) and (ii) briefly here, and (iii) in the next section.

2.3.1 Survival curves

The Kaplan-Meier estimate can be used to estimate the distribution of time to first entry to a given state. Using this method, we often compare the estimated probability of entry to an absorbing state based on the multistate model with the Kaplan-Meier estimate.

Gentleman et al. (1994) used this method when the time of entry into the absorbing state was exactly known. We can similarly use this method when times into the absorbing state are interval censored, using the Turnbull (1976) estimate in place of the Kaplan-Meier. If the Markov model is correct, we expect a close agreement between the model-based and empirical estimates. We can consider, say 95% pointwise confidence intervals based on the multistate model, and check if the empirical survival curve is inside this bounds. When we do not have an internal covariate and the exact state entry times are known, this can be implemented by the function plot.survfit.msm in the msm package. However, we can no longer use this function in the presence of internal covariates. In Chapter 4, we will discuss this approach in more detail.

2.3.2 Prevalence counts

Similar to empirical survival curves for time to first entry in a state, we can have empirical curves for the probability of occupancy of all the states. These probabilities are called prevalence functions (probabilities). We can define the prevalence probability for state r at time t as $P_{y_0 r}(0,t) = P(Y(t) = r|Y(0) = y_0)$, where y_0 is a specified initial state at time 0, or as $P_r(t) = \sum_{y_0} P(y_0)P_{y_0 r}(0,t)$, where $P(y_0) = P(Y_0 = y_0)$. However, in many studies, individuals all start from some state 1 at time 0, so the prevalence probability becomes $P_{1r}(t) = P(Y(t) = r|Y(0) = 1)$.

A check on a multistate model can be made by comparing prevalence probabilities based on it with empirical prevalence probabilities. An equivalent approach is to compare prevalence counts, or frequencies. In the prevalence counts method, we compare observed or "empirical" state occupancy frequencies with the expected values based on the fitted multistate models, at a series of times, $0, t_1 ..., t_m$. In the case of a balanced observation scheme, when everyone starts from state 1 at time $t_0 = 0$ and is seen at common times $t_0, t_1, ..., t_m$, we have observed prevalence counts

$$o_{jr}(t_j) = \sum_{i=1}^{n} I(t_j \leq c_i) I(Y_i(t_j) = r), \tag{2.5}$$

and

$$e_{jr}(t_j) = \sum_{i=1}^{n} I(t_j \leq c_i) \hat{P}_{1r}(0, t_j | z_i), \tag{2.6}$$

where c_i is the censoring time of individual i, $I(t_j \leq c_i)$ indicates that the individual was under observation, z_i is the vector of fixed covariates for individual i, and $\hat{P}$ denotes estimates from the multistate model. A large deviation of observed counts, o_{jr}, and expected counts, e_{jr}, indicates a poor fit. Formal chi-squared tests can also be given.

Approaches for computing empirical prevalence counts for observational studies with irregular sampling times have been rather ad hoc. Since the precise state each individual occupies at the assigned times $t_1, ..., t_m$ are not generally known, it is necessary to interpolate in some way. Gentleman et al. (1994) assumed that individuals would remain in the state they were observed in at the previous observation time. An alternative is to assume that transitions occur half way between successive observation times. Both approaches have inaccuracy, but it is suggested that this might not be significant if individuals are observed fairly frequently. The function prevalence.msm in the msm package R computes prevalence counts and allows a choice of interpolation methods. Mathieu et al. (2005) proposed another approach which does not use any interpolation method. They only considered subgroups of individuals who had a visit time close to the time of interest.

2.4 Formal tests of fit

2.4.1 Model expansion

A likelihood ratio test is often used to compare the fit of two models, where one is nested within the other. This is usually used when we want to test whether a simplifying assumption for a model is valid. In the context of multistate models, we might want to test that some covariates effects are zero, for example, or that a process is time homogeneous. Some likelihood ratio tests can be implemented by the function lrtest.msm in the msm package. Likelihood ratio tests will be discussed in Section 5.4.

2.4.2 Pearson goodness of fit test

Formal goodness of fit tests in Markov models have been based on a comparison of observed transition counts with expected transition counts based on the model. Such tests are often referred to as Pearson tests, given their connection to multinomial tests of fit.

In the case of the balanced observation scheme, when $t_{ij} = t_j$ for all i, and categorical covariates, Kalbfleisch and Lawless (1985) grouped the transition counts according to the observation number ($j = 1, 2, ..., m$) and the covariate value. Thus, the observed and expected number of r to s transitions between t_j and t_{j+1} for individuals with covariate value c are as follows:

$$o_{rs}^{(jc)} = \sum_{i:z_i=c} I\{Y_i(t_{j+1}) = s, Y_i(t_j) = r)\}, \tag{2.7}$$

$$e_{rs}^{(jc)} = \hat{P}_{rs}(t_j, t_{j+1}|z_i = c)n_{jcr}, \tag{2.8}$$

where n_{jcr} is the number of individuals with covariate value c observed in state r at time t_j, and having an observation at time t_{j+1}.

Within each group (jc) of observation number and covariate value, the transition probabilities for each individual are the same in this case. Hence, within each group, the transition counts are multinomial, given the values n_{jcr}, and we can obtain a Pearson goodness of fit statistic. These groups are conditionally independent if the Markov model is of order one, so an overall goodness of fit statistic is obtained by adding the statistics calculated for each group. The overall Pearson goodness of fit statistic is then

$$X^2 = \sum_j \sum_c \sum_r \sum_s \frac{(o_{rs}^{(jc)} - e_{rs}^{(jc)})^2}{e_{rs}^{(jc)}}, \tag{2.9}$$

which is asymptotically equivalent to the likelihood ratio test statistic

$$\Lambda = \sum_j \sum_c \sum_r \sum_s o_{rs}^{(jc)} log(\frac{o_{rs}^{(jc)}}{e_{rs}^{(jc)}}). \tag{2.10}$$

Both statistics have an asymptotic Chi-square distribution with degrees of freedom given by $d - |\theta|$, where d is the number of independent cells and $|\theta|$ is the number of unknown parameters for the fitted model. Suppose that n_r is the number of possible values for $Y(t_j)$ given $Y(t_{j-1}) = r$. Then, the number of independent cells is $G\sum_r(n_r - 1)$, where G is the product of the number of observation times multiplied by the total number of covariate values. We note that in the case where χ^2 or Λ are large, a comparison of the individual $o_{rs}^{(jc)}$ and $e_{rs}^{(jc)}$ indicate the nature of the lack of fit.

We can not use this goodness of fit test if the observation scheme is unbalanced (irregular) or when we are not able to group observations due to the large number of possible covariate values. Aguirre-Hernandez and Farewell (2002) proposed a Pearson goodness of fit test that allows irregular sampling times and continuous covariates, and this has been extended by Titman and Sharples (2008). An excellent review of this and other methods of fit is given by Titman and Sharples (2010a). For each allowable transition, the data $(Y_i(t_{ij-1}), Y_i(t_{ij}))$ are partitioned by observation categories, h, which are groups based on quantiles of the time since the start of the process. It is also partitioned within each observation category h, by a time interval category, l_h, based on quantiles of the time interval between observations. Additionally, observations are grouped by covariates, c, according to quantiles of the estimated transition intensity $q_{rs}(z)$. For instance, suppose that we have a three-state progressive model, and we consider $Q_{33.3}$, and $Q_{66.7}$ quantiles of transition intensities $q_{12}(z)$ and $q_{23}(z)$ for all individuals. Then, the covariate groupings can be based on the values of $q_{12}(z_i)$ and $q_{23}(z_i)$ compared to their quantiles for the whole sample. So, in this case, we have 9 covariate groups.

Thus, for each transition r to s for individual i with observations at times $t_{ij}, j = 1, ..., m_i$, we calculate for each group (hl_hc) the frequencies

$$o_{rs}^{(hl_hc)} = \sum I\{Y_i(t_{ij+1}) = s, Y_i(t_{ij}) = r\}, \tag{2.11}$$

$$e_{rs}^{(hl_hc)} = \sum \hat{P}(Y_i(t_{ij+1}) = s, Y_i(t_{ij}) = r) I(Y_i(t_j) = r), \tag{2.12}$$

where the sum for each h is over the set of observations:

$$\forall i, j : t_{j+1} - t_j \in l_h, \quad q(z) \in c.$$

Then the statistic, $T = \sum_h \sum_{l_h} \sum_r \sum_s \sum_c \frac{(o_{rs}^{(hl_hc)} - e_{rs}^{(hl_hc)})^2}{e_{rs}^{(hl_hc)}}$, was proposed. Since the time intervals and individuals within the groups are non-identical, the observed transitions are approximate realizations from a set of independent but non-identical multinomials, and the limiting distribution is a linear combination of variables with $\chi^2(1)$ distribution, where the coefficients depend on unknown parameter values. Titman (2009) described this in detail. In some cases the distribution of the test statistic is approximately Chi-squared with degrees of freedom $d - |\theta|$, similar to the degrees of freedom in the standard Pearson goodness of fit test. However, this is in general a poor approximation.

Aguirre-Hernandez and Farewell showed that the Chi-squared distribution was adequate for models without covariates, but for models with fitted covariates, T had a distribution with a mean higher than the degrees of freedom. So, they proposed a parametric bootstrap to approximate the distribution of the T statistic. Each bootstrap sample requires, first, generating realizations of the process based on the maximum likelihood estimates of the original data at the observation times from the original data. Then, fitting the multistate model to the new data and computing the T statistic must be done. This method can require an excessive amount of computation and in many cases is not practically applicable. This becomes more infeasible as the number of states and time-dependent covariates in the multistate models increase.

The Aguirre-Hernandez and Farewell method is not applicable in the common situation where the process includes an absorbing state, such as death, for which the time of entry is known exactly. Their method is based on partitioning the data set on the basis of intervals between observations which are independent of the process. Titman and Sharples (2008) proposed a modified method to extend the tests to deal with Hidden Markov models (HMM) as well as to handle terminal events. In their modified method, they impute estimated times of the next observation assuming that the individuals survived. The pearson.msm function in the msm package uses the Aguirre-Hernandez and Farewell approach when there are no absorbing states and the Titman and Sharples goodness of fit test when there is an absorbing state.

Titman (2009) proposed another asymptotic approximation for the Aguirre-Hernandez and Farewell (2002) and Titman and Sharples (2008) statistics. This approximation is obtained by showing that the asymptotic distribution of these statistics can be considered as the joint distribution of the observed counts in the contingency table and the score function of the log likelihood. The new approximation provides more accurate p values in cases where the bootstrapping is time consuming and $\chi^2(d-|\theta|)$ is not a good approximation for the distribution of the statistic. However, it is cumbersome to apply and Titman (2009) illustrates it only for a very simple model. In view of these various difficulties, we focus on likelihood ratio tests for goodness of fit, and discuss this approach in Section 5.4.

Chapter 3

Effects of Observation Schedule on Efficiency of Estimation

3.1 Introduction

Sometimes it is possible, at additional cost, to determine the exact times at which certain transitions or events occur. However, with panel data, we only know the states of an individual at each observation time t_{ij}, and the exact transition times are unknown. In this case, it is of interest to know the loss of information due to the intermittent observation, and how this changes if times between visits are increased or decreased. Hwang and Brookmeyer (2003) considered the effects of gap time between visits on precision of the transition intensity estimates in progressive multistate models. Progressive or uni-directional models refer to models in which the individuals can only move from state r to state s $(s \neq r)$ in one direction, whereas in bi-directional models, individuals can have reversible transitions. Mehtala et al (2011) and Cook (2000) discuss two-state bi-directional models. In this chapter, we study and compare uni-directional and bi-directional models, and consider estimation of baseline intensities, covariate effects, and transition probabilities.

The remainder of this chapter is as follows. Section 3.2 discusses estimation of parameters in Markov models based on intermittent observation, and describes how we can obtain asymptotic covariance matrices. Section 3.3 presents some numerical studies of the effects of gap time between visits on estimation in both progressive and bi-directional models, and Section 3.4 considers applications to the observational CANOC data described in Chapter 1. Finally, Section 3.5 contains some concluding remarks.

3.2 Estimation and asymptotic covariance matrices

3.2.1 Maximum likelihood estimation

In Section 2.2 of Chapter 2, we presented the likelihood function for Markov models based on panel data. In this section, we focus on the special case, where $t_{ij} = t_j$ for $i = 1, ..., n$ and $j = 1, ..., m$. The likelihood function based on these n individuals with observation times $t_0 < t_1 < ... < t_m$ is

$$L(\theta) = \prod_{i=1}^{n} \prod_{j=1}^{m} P_{y_i(t_{j-1}), y_i(t_j)}(w_j; \theta) \tag{3.1}$$

where $P_{rs}(w; \theta) = P(Y_i(t + w) = s | Y_i(t) = r; \theta)$ and $w_j = t_j - t_{j-1} \quad (j = 1, ..., m)$. The transition probability matrix $P(w)$ can be computed using the methods described in Section 1.3.1.

3.2.2 Asymptotic covariance matrice for panel data

Kalbfleisch and Lawless (1985) discussed the expected (Fisher) information matrix $I(\theta)$ for θ based on n independent individuals and gave an algorithm for computing it. The $I(\theta)$ has elements

$$\begin{aligned} I_{uv}(\theta) &= E(\frac{-\partial^2 logL(\theta)}{\partial\theta_u\theta_v}) \\ &= \Sigma_{j=1}^{m}\Sigma_{r,s=1}^{K}\frac{E_r(t_{j-1};\theta)}{P_{rs}(w_j;\theta)}\frac{\partial P_{rs}(w_j;\theta)}{\partial\theta_u}\frac{\partial P_{rs}(w_j;\theta)}{\partial\theta_v}, \end{aligned} \tag{3.2}$$

where $E_r(t_{j-1};\theta) = \Sigma_{i=1}^{n} P\{Y_i(t_{j-1}) = r|Y_i(t_0);\theta\}$ is the expected number of individuals in state r at time t_{j-1}. Here, we assumed that all individuals are observed at common times $t_1, ..., t_m$. Sometimes individuals can be lost to followup before t_m. In this case, we multiply the ith term in $E_r(t_{j-1};\theta)$ by $\pi_i(t_{j-1})$ which is the probability that individual i is still under observation at time t_{j-1}.

We can either use the Kalbfleisch and Lawless (1985) algorithm for computing these derivatives, or use the numerical differentiation formula

$$\frac{\partial P_{rs}(w;\theta)}{\partial\theta_u} \doteq \frac{P_{rs}(w;\theta+\Delta_u) - P_{rs}(w;\theta-\Delta_u)}{2\delta_u}, \tag{3.3}$$

where Δ_u is a zero vector except for the small value δ_u for the element corresponding to θ_u. Then, $\sqrt{n}(\hat{\theta}-\theta)$ has the asymptotic covariance matrix lim $nI(\theta)^{-1}$ when the number of individuals n approaches infinity. Here, we assume that there is an initial distribution for individuals at time t_0 and the Markov model is correct and that θ is the true parameter vector.

3.2.3 Asymptotic covariance matrices for continuous observations

The asymptotic covariance matrices for the case of continuous observations are discussed by Andersen et al. (1993), Chapter 6. We will focus here on time-homogeneous models for which $q_{rs}(t) = q_{rs}$ for the time interval $0 \leq t < \tau$, and individuals are observed over $(0, \tau)$. The likelihood function is

$$L = \prod_{r \neq s} q_{rs}^{n_{rs}} exp\{-q_{rs} \int_0^\tau \delta_r(t)dt\}, \tag{3.4}$$

where $\delta_r(t) = \sum_{i=1}^n \delta_{ir}(t)$, $\delta_{ir}(t) = I(Y_i(t) = r)$, and n_{rs} is the total number of r to s transitions among the n individuals across all observation times.

The derivative of the loglikelihood function with respect to q_{rs} is

$$\frac{dl}{dq_{rs}} = \frac{n_{rs}}{q_{rs}} - \int_0^\tau \delta_r(t)dt. \tag{3.5}$$

The observed information is obtained from

$$\frac{-d^2l}{dq_{rs}^2} = \frac{n_{rs}}{q_{rs}^2}, \tag{3.6}$$

where $\frac{-d^2l}{dq_{rs}dq_{r's'}} = 0$ if $rs \neq r's'$. The expected information matrix has non-zero entries

$$E\{\frac{-d^2l}{dq_{rs}^2}\} = \frac{En_{rs}}{q_{rs}^2}. \tag{3.7}$$

Now, by equalizing the expected value of the score function to zero, we have

$$\begin{aligned} En_{rs} = q_{rs}E\{\int_0^\tau \delta_r(t)dt\} &= q_{rs}\sum_{i=1}^n \int_0^\tau E\{\delta_{ir}(t)\}dt \\ &= q_{rs}\sum_{i=1}^n \int_0^\tau P\{Y_i(t) = r|Y_i(0)\}dt. \end{aligned} \tag{3.8}$$

By substituting En_{rs} in the expected information, we thus get

$$E\{\frac{-d^2l}{dq_{rs}^2}\} = \frac{1}{q_{rs}}\sum_{i=1}^{n}\int_0^{\tau} P\{Y_i(t) = r|Y_i(0)\}dt. \tag{3.9}$$

So, for the time homogeneous Markov model, the asymptotic covariance matrix for the transition intensity estimates is a diagonal matrix with elements

$$asvar(\sqrt{n}(\hat{q}_{rs} - q_{rs})) = \frac{nq_{rs}}{\sum_{i=1}^{n}\int_0^{\tau} P\{Y_i(t) = r|Y_i(0)\}dt}. \tag{3.10}$$

3.3 Numerical studies

Hwang and Brookmeyer (2003) showed in their numerical study that, for a three-state time-homogeneous progressive model with a fixed followup time, the loss of efficiency for estimating $q_{r,r+1}(r = 1, 2)$ is small when the inter-visit time, w, is less than the average sojourn time q_{rr+1}^{-1} in state r. However, in the case of bi-directional models, we are facing a different situation. Suppose for example, we have a simple two-state bi-directional model with transitions q_{12} and q_{21}. The transition probability matrix $P(t, t + w)$ is given by Kalbfleisch and Lawless (1985) as

$$\begin{pmatrix} 1 - \pi(1 - exp(-\alpha w)) & \pi(1 - exp(-\alpha w)) \\ (1 - \pi)(1 - exp(-\alpha w)) & \pi + (1 - \pi)exp(-\alpha w) \end{pmatrix}, \tag{3.11}$$

where $\alpha = q_{12} + q_{21}$ and $\pi = q_{12}\alpha^{-1}$. As w increases, both rows of the above matrix approach the limiting distribution $(1 - \pi, \pi)$. Thus, if the gap time w is large, we might be able to estimate π precisely, but not the transition intensities q_{12} and q_{21}.

In this section, we consider the effects of a constant inter-visit time, w, on the precision of estimation of transition intensities q_{rs}, transition probabilities $P_{rs}(t)$, and covariate effects

β_{rs}. For this purpose, first we compute the asymptotic variance of the maximum likelihood estimators of parameters obtained from panel data which we discussed in Section 3.2.2, and then we calculate the asymptotic variance of the estimates in the special case where the individuals are monitored continuously in time, as discussed in Section 3.2.3. Finally, the relative efficiency of a specific estimator from panel data relative to continuous-time data is defined as

$$reff(\hat{\psi}^{(w)}) = \frac{asvar(\hat{\psi}^{(0)})^{1/2}}{asvar(\hat{\psi}^{(w)})^{1/2}}, \tag{3.12}$$

where $asvar(\hat{\psi}^{(0)})$ is the asymptotic variance of the maximum likelihood estimator for ψ when we have continuous observation data (i.e. $w = 0$), and $asvar(\hat{\psi}^{(w)})$ is the asymptotic variance of the maximum likelihood estimator for ψ when we have panel data. The diagonal elements of the $I(\theta)$ matrix in Section 3.2.2, and equation (3.10) multiplied by $1/n$ represent $asvar(\hat{\psi}^{(0)})$, and $asvar(\hat{\psi}^{(w)})$, respectively. In the following two subsections, we consider several multistate models, and compute the relative efficiencies of different estimators. For convenience, tables on these relative efficiencies are given all together in an Appendix (Section 3.6) at the end of the chapter.

3.3.1 Two-state models

We first investigate the effect of gap times between visits on the precision of estimation in two-state bi-directional and progressive models with no covariates. We compare models M1a with $q_{12}, q_{21} > 0$ and models M1b, where $q_{12} > 0, q_{21} = 0$. We considered two models M1a with different average sojourn times in state 2. In M1a(i), the average sojourn times in both states 1 and 2 equal 1, but in M1a(ii) the average sojourn time in state 2 is 1/4, as might occur when state 2 represents a recurring transient condition. Additionally, we considered two followup times, $\tau = 4$ and 10 years, with fixed gap times $w = \Delta t$ ranging from 0.25 to 2. Tables 3.10 and 3.11 show the asymptotic relative efficiencies of estimators of transition intensities and prevalence probabilities for models M1a and M1b, with $\tau = 4$ and 10, respectively. We decided not to show $P_{12}(t)$ to concur a little space in these tables.

As we can see from both Tables 3.10 and 3.11, the relative efficiency of transition intensity estimators decreases drastically with the increase of the gap times in model M1a. However, the relative efficiency of these estimators does not change much in model M1b. The tables also show that the relative efficiencies of transition intensity estimators for M1a(i) model are higher than the corresponding values in M1a(ii). This is due to the smaller average sojourn time in state 2 for model M1a(ii), which causes more transitions between the two states. The results in Tables 3.10 and 3.11 show that the increase of the followup time from 4 to 10 does not affect the relative precision of transition intensity estimators, but can affect the prevalence probabilities, depending on the speed with which these values approach the limiting distribution as w increases.

3.3.2 Three-state models

Here, we investigate the effect of Δt on the relative efficiency of estimation for the multistate models in Figure 3.1. First, we focus on Figure 3.1(b), which will be used later for modeling viral load dynamics in the CANOC study. For models M2a and M2b, there are no covariates and the transition intensities are taken to be $q_{12} = 1, q_{21} = 2, q_{23} = 1$ and $q_{12} = 1, q_{21} = 0, q_{23} = 1$ respectively. We will also consider these models in the presence of a binary covariate z, and we will name them M3a and M3b respectively. Covariate effects are of multiplicative form, as described in Section 1.3.1. That is, $q_{rs}(z) = q_{0rs} exp(\beta_{rs} z)$ for $r \neq s$.

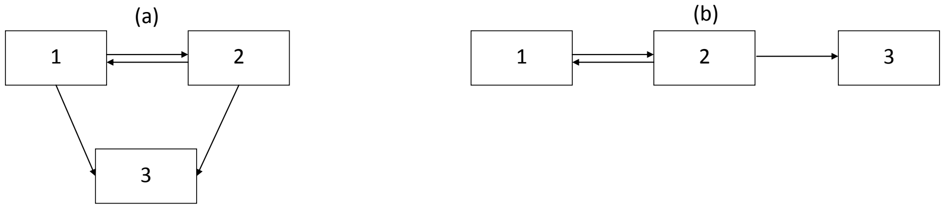

Figure 3.1: Two multi-state models

The covariate effects that we used in our numerical calculations are $\beta_{12} = -0.2, \beta_{21} = -0.5, \beta_{23} = -0.9$ for model M3a and $\beta_{12} = -0.2, \beta_{21} = 0, \beta_{23} = -0.9$ for model M3b. We again considered two followup times 4 and 10 years for each scenario. The choice of parameter values for the numerical studies is based on some preliminary analysis of viral load processes in specific groups of the CANOC data.

Tables 3.12 and 3.13 show similar results to Tables 3.10 and 3.11. We again decided to not show $P_{12}(t)$ to concur a little space in these tables. For the bi-directional model M2a, we have a considerable drop in the relative efficiencies of transition intensity estimators as Δt increases. However, for the transition intensity q_{23} this drop is smaller, because this transition is uni-directional. The loss of efficiency for prevalence probability estimators is very small, except at $t = 1$ in Table 3.12 and $t = 2$ in Table 3.13. For the progressive model M2b, the loss of efficiency for all estimators is negligible, except when $\Delta t = 2$.

Tables 3.14 and 3.16 show the relative efficiencies for baseline transition intensities and covariate effects in model 3.1(b) with the binary covariate. We assumed that the numbers of individuals having covariate values $z = 0$ and $z = 1$ are equal. As we can see from these two tables, the relative efficiencies for transition intensity estimators are similar to Tables 3.12 and 3.13. For the M3a model, the relative efficiency of covariate effects β_{12} and β_{21} drastically decreases as Δt increases, just as for the relative efficiency of the transition intensities. This was expected, since $exp(\beta_{rs})$ is a ratio of transition intensities for $z = 1$ and $z = 0$. However, for β_{23} we do not have loss of efficiency, as this transition is uni-directional. For the uni-directional model M3b, the loss of efficiencies is very small. Tables 3.15 and 3.17 represent the relative efficiencies for prevalence probability estimators of a group of individuals with $z = 1$ for the cases where $\tau = 4$ and 10. Results for individuals with covariate $z = 0$ are as in Tables 3.12 and 3.13. From Tables 3.15 and 3.17, we can not see any substantial loss of efficiency except in model M3a for $P_{11}(1)$ and $P_{11}(2)$ for followup times $\tau = 4$ and 10 years, respectively.

Tables 3.18 to 3.21 show similar results for models M4a and M4b, illustrated in Figure 3.1(a). We assumed transition intensities $q_{12} = 1, q_{21} = 2, q_{13} = 0.25, q_{23} = 0.5$ and $q_{12} = 0.5, q_{21} = 0, q_{13} = 0.25, q_{23} = 0.5$ for models M4a and M4b respectively. In addition, we also presented the relative efficiencies for transition probabilities in Tables 3.19 and 3.21. As we can see from Tables 3.18 and 3.20, the pattern of drops in relative efficiencies is similar to the results in Tables 3.12 and 3.13. In model M4a, the loss of efficiency is very substantial for transition intensities, but much less for the uni-directional transitions (i.e. for q_{13} and q_{23}). It is interesting that estimators of transition probabilities $P_{23}(t)$ in model M4a lose more efficiency than for probabilities $P_{13}(t)$.

3.4 Application to data on viral load dynamics

In this Section, we consider the analysis of CANOC data which we described earlier in Section 1.5. We focus on the use of multistate models for viral load (VL) dynamics, and the challenges in fitting such models based on data with irregular followup times. The numerical results in Section 3.3 included equally spaced gap times between visits, but we expect the impact to be similar when they vary within and between subjects.

In the present analysis, individuals were included in the study if they achieved virologic suppression ($VL < 50$) copies/mL on two consecutive visits. One main interest in this study is the event of virologic rebound, and so we consider the time from virologic suppression to virologic rebound ($VL \geq 1000$). For individuals not having a viral rebound during the followup period, the viral rebound times are censored at the last viral load measurement times.

Grennan et al. (2012) studied viral rebounds in the CANOC data. Since there are differences in patient characteristics across CANOC sites, we focus here only on males in the British Columbia cohort, which has a good record for followup. We considered 1869 participants who had their first antiretroviral therapy between January 4, 2000 and Febuary 25, 2009. We divided participants into two groups based on the year of their first antiretroviral therapy (FARV), in order to allow for developments in cART over time: FARV1 refers to those who started cART in years $2000 - 2004$, and FARV2 refers to those who started in years $2005 - 2009$. There are $n = 834$ individuals in the FARV1 and $n = 1035$ individuals in the FARV2 group.

In our analysis, we are considering the same covariates as Grennan et al. (2012). Baseline covariates include age, indication of injection drug use (IDU), indication of sex with men (MSM), and type of the initial combination antiretroviral therapy (cART). Age is taken to be a binary covariate which categorizes the age at entry to the study in years as age ≤ 45 or age > 45. The reference level is age ≤ 45. Covariate IDU has three categories: the reference category is for non injection drug users (IDU=No); the second category is for those individuals who are injection drug users (IDU=Yes); (IDU=NA) refers to those whose IDU status is unknown, of which they are many. Covariate cART has two categories of treatment (nonnucleoside reverse transcriptase inhibitor (NNRTI), or protease inhibitor (PI)); the reference level is NNRTI. We also considered covariate MSM, but we dropped it for the present discussion, since it did not have any significant effect in the cohort that we considered. Table 3.1 shows some basic characteristics of the study individuals.

Cohort members are nominally seen about every three months, but are sometimes seen more or less often. Table 3.2 shows the quartiles of length of followup, number of visits, and gap times between visits (in years). In FARV1, there were 368 gap times which were smaller than 0.05 years, and 13 gap times greater than 2 years. Similarly, in FARV2, we had 222 gap times smaller than 0.05 years, and 1 gap time greater than 2 years. Figure 3.2 and 3.3 show plots of last visit times versus first antiretroviral visit times for FARV1 and FARV2 groups. The time origins for plots in Figures 3.2 and 3.3 are January 4, 2000 and January 1, 2005, respectively. In these plots, crosses indicate individuals who died; in this case, the time of death is the last "visit" time. For individuals who were still alive as of the administrative censoring time, dots indicate the time of their last visit. If every inter-visit time were less than, say, 6 months then the dots would all be within 6 months of the administrative end date. We see that this is the case for the majority of persons, but that at the end of follow up some individuals have not been seen for a long time. These persons might be deemed lost to followup, though in principle they could reappear in the future.

3.4.1 Model fitting and estimation

In this section, we analyse viral loads up to the time of a first viral rebound by considering the multistate model shown in Figure 3.1(b). We defined 3 states based on viral load values: State 1 is $VL < 50$, state 2 is $50 \leq VL < 1000$, and state 3 is $VL \geq 1000$ copies/mL.

Table 3.1: Characteristics of cohort members at time of cART initiation.

Group	Age		IDU			MSM			cART	
	≤ 45	> 45	No	Yes	NA	No	Yes	NA	NNRTI	PI
FARV1	531	303	272	272	290	285	259	290	402	432
	(64%)	(36%)	(33%)	(33%)	(35%)	(34%)	(31%)	(35%)	(48%)	(52%)
FARV2	576	459	323	287	425	318	292	425	402	633
	(56%)	(44%)	(31%)	(28%)	(41%)	(31%)	(28%)	(41%)	(39%)	(61%)

Table 3.2: Quartiles of length of follow up, number of visits, and gap times (in years) between visits for the FARV1 and FARV2 groups.

Group	Length of followup			Number of visits			Gap times		
	Q_{25}	Q_{50}	Q_{75}	Q_{25}	Q_{50}	Q_{75}	Q_{25}	Q_{50}	Q_{75}
FARV1	4.441	5.704	7.221	18	26	34	0.123	0.208	0.266
FARV2	0.986	1.863	2.973	6	10	15	0.112	0.189	0.255

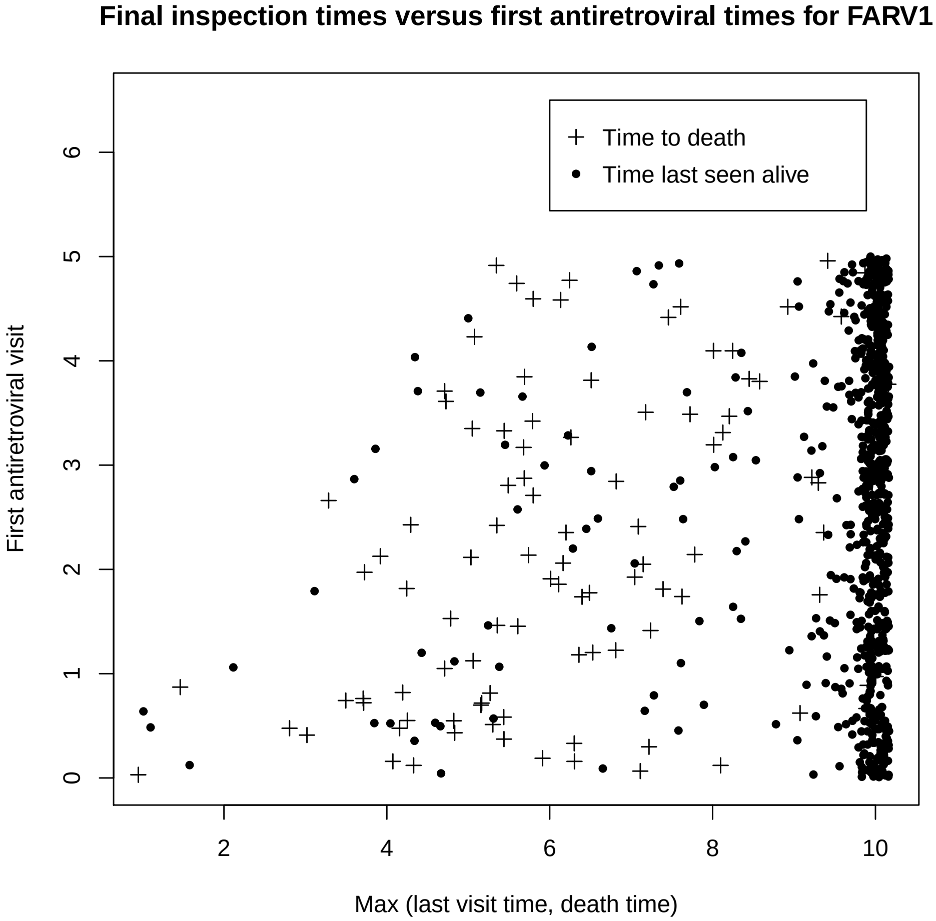

Figure 3.2: The plot of length of followup times vs first antiretroviral visit for the FARV1 group. Time is in years, measured from January 1, 2000.

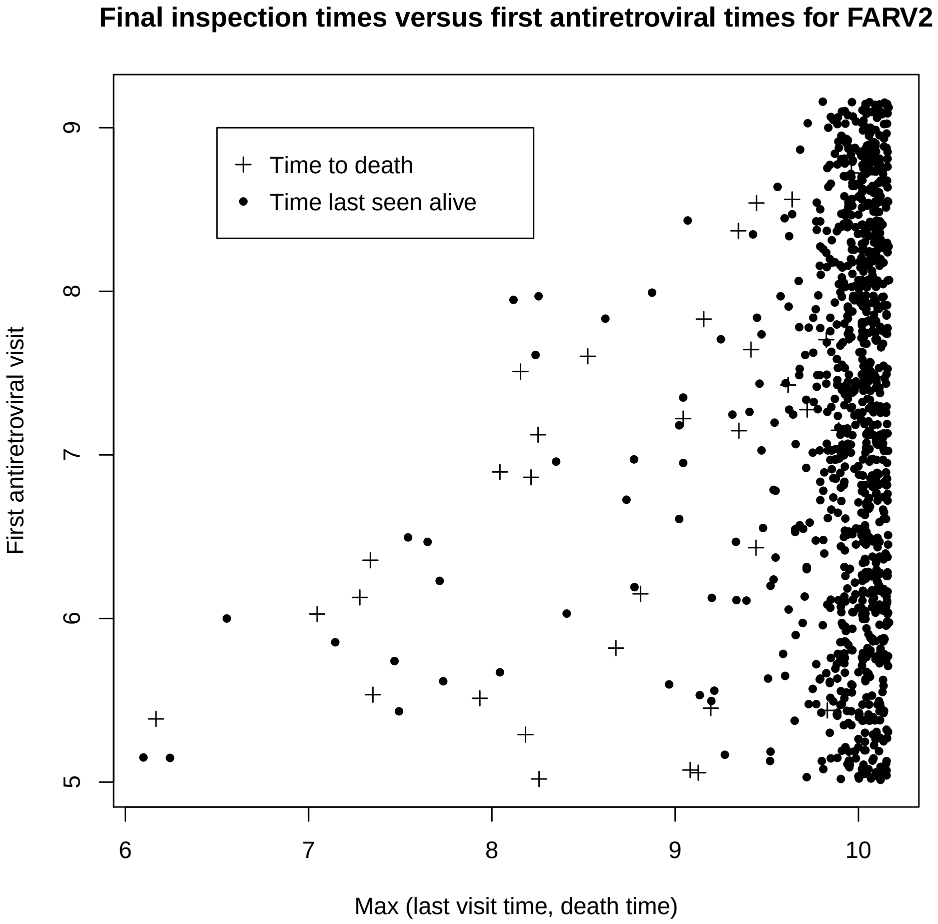

Figure 3.3: The plot of length of followup times vs first antiretroviral visit for the FARV2 group. Time is in years, measured from January 1, 2000.

Table 3.3: Observed transitions for the three-state multi-state model in Figure 3.1-(b) for the FARV1 group.

	Transition to		
Transition from	State 1	State 2	State 3
State 1	13764	426	275
State 2	359	191	52

Table 3.4: Observed transitions for the three-state multi-state model in Figure 3.1-(b) for the FARV2 group.

	Transition to		
Transition from	State 1	State 2	State 3
State 1	8228	464	112
State 2	400	211	16

State 1 effectively means that HIV in the blood is undetected, and state 3 represents viral rebound. We assumed state 3 as an absorbing state, since after an individual enters this state, treatment is adjusted so that they can achieve viral suppression. Then, after viral suppression, we can model the next viral rebound once again. Tables 3.3 and 3.4 present the transition counts for this multistate model in the FARV1 and FARV2 groups. The counts in Tables 3.3 and 3.4 are totals across all pairs of successive visit times for all individuals, and because the inter-visit times vary a good deal, should just be considered as summaries of observed transitions.

We fit Markov regression models with transition intensities

$$q_{rs}(t|z) = q_{0rs}(t) exp(\beta_{rs} z), \quad (r, s) = (1, 2), (2, 1), (2, 3) \tag{3.13}$$

where the vector β_{rs} indicates the covariate effects for transitions r to s, and z is a covariate vector. We fitted time homogeneous and piecewise constant Markov models to both FARV1 and FARV2 groups using the msm package in R (Jackson 2011). The results for the time homogeneous Markov model fitted are not presented. Tables 3.5 and 3.6 show estimates and standard errors for the baseline transition intensities and covariate effects in the time inhomogeneous models. In these models, the cut time point is chosen to be 2 years. The msm software handles this by defining a time-dependent covariate which equals zero up to $t = 2$ and 1 for $t \geq 2$. So, the baseline intensities are for times $t < 2$ and the "time" regression coefficient is the log ratio of the transition intensities for $t \geq 2$ to those for $t < 2$.

Tables 3.5 and 3.6 show that individuals with age above 45 years tend to enter state 3 from state 2 at a lower rate than people below age 45. However, the directions of estimated regression coefficients for the older age people for moving between state 1 and 2 are different in the FARV1 and FARV2 groups. In the FARV2 group, the older age individuals are more likely to stay in state 2 once they have entered this state, whereas in the FARV1 group the opposite is true. Injection drug users have a higher rate of entry into state 3, particularly in the FARV2 group. Individuals with an unknown IDU status are similar to non-injection drug users. Regarding the cART covariate, people with PI initial treatment tend to move more from state 1 to state 2, and have less tendency to transition from state 2 to state 3. Finally, we can also see that time has a negative effect on different transitions. That is, individuals have less tendency to move between states after 2 years, especially from state 1 to state 2. One possibility is that this is a reflection of unobserved heterogeneity across individuals, with individuals more likely to move already in state 3 as time goes by.

Table 3.5: Estimated baseline transition intensities, regression coefficients and their standard errors (in brackets) for the piecewise constant Markov multistate model for the FARV1 group.

Transition	$log(q_{0rs})$	Age (> 45)	IDU (=Yes)	IDU (=NA)	cART (=PI)	Time
		vs ($\leq$ 45)	vs (No)	vs (No)	vs (NNRTI)	
1 → 2	-0.862	0.261	0.180	0.058	0.733*	-0.816*
	(0.138)	(0.124)	(0.135)	(0.161)	(0.116)	(0.117)
2 → 1	2.210	0.244	-0.487*	0.301	0.295*	-0.347*
	(0.165)	(0.151)	(0.176)	(0.178)	(0.142)	(0.136)
2 → 3	1.533	-0.420*	0.297*	0.145	-0.427*	-0.887*
	(0.153)	(0.136)	(0.149)	(0.169)	(0.127)	(0.136)

$*|\frac{\hat{\beta}}{se(\hat{\beta})}| > 1.96$

Table 3.6: Estimated baseline transition intensities, regression coefficients and their standard errors (in brackets) for the piecewise constant Markov multistate model for the FARV2 group.

Transition	$log(q_{0rs})$	Age ($>$ 45)	IDU (=Yes)	IDU (=NA)	cART (=PI)	Time
		vs ($\leq$ 45)	vs (No)	vs (No)	vs (NNRTI)	
1 → 2	-0.265	-0.085	0.003	0.112	0.426*	-0.860*
	(0.214)	(0.139)	(0.160)	(0.173)	(0.180)	(0.180)
2 → 1	2.653	-0.338*	-0.258	0.201	-0.068	-0.222
	(0.242)	(0.144)	(0.177)	(0.171)	(0.201)	(0.173)
2 → 3	0.523	-0.555*	0.808*	-0.214	-0.781*	0.137
	(0.257)	(0.196)	(236)	(0.264)	(0.204)	(0.268)

$*|\frac{\hat{\beta}}{se(\hat{\beta})}| > 1.96$

3.4.2 Effects of inter-visit times

We investigate here the effects of the frequency of visits on the precision of baseline transition intensities and covariate effects estimators in the FARV2 group. Individuals are nominally seen every three months. However, Table 3.2 shows that the median gap time is 0.189 years. We examine how the standard deviations of estimates change when the gap time between visits increases from 0.25 years (every three months) to 0.5 years (every 6 months). This exercise will illustrate the effect on estimation of having less frequent followup. We consider the estimated model represented in Table 3.6 as the true process, except that we assume the transition intensities do not change after 2 years. We then use the approach discussed in Section 3.2.2 to compute the asymptotic standard deviations for baseline transition intensities and covariate effects. We adjusted the total followup time for each individual as the closest quarter-year value, thus allowing for the different lengths of followup. Table 3.7 shows the results. As we see, the standard errors for baseline transition intensities and covariate effects are substantially larger when $\Delta t = 0.5$ than when it is 0.25 years. Table 3.8 shows the standard error estimates for prevalence probabilities. The estimates of probabilities are for an individual with z vector 0. The table indicates that the estimates of prevalence probabilities such as $P_{13}(1)$ and $P_{13}(2)$ have almost the same standard errors for both $\Delta t = 0.25$ and $\Delta t = 0.5$. Table 3.9 shows the estimated prevalence probabilities for the highest risk individuals (age$\leq$ 45, IDU=Yes, cART=NNRTI) and the lowest risk individuals (age$>$ 45, IDU=No, cART=PI). The $\hat{P}_{13}(t)$ at $t = 1$ and 2 for the highest risk individuals are 0.176 and 0.330, respectively. Similarly, the corresponding values for the lowest risk individuals are 0.043 and 0.089.

Table 3.7: Standard errors for estimators of baseline transition intensities, covariate effects if all gap times were 3 months ($\Delta t = 0.25$), or 6 months ($\Delta t = 0.5$).

Transition	Δt	$log(q_{0rs})$	Age (> 45)	IDU (=Yes)	IDU (=NA)	cART (=PI)
			vs (≤ 45)	vs (No)	vs (No)	vs (NNRTI)
$1 \rightarrow 2$	0.25	0.248	0.185	0.199	0.286	0.208
	0.5	0.548	0.430	0.507	0.709	0.442
$2 \rightarrow 1$	0.25	0.281	0.205	0.224	0.302	0.236
	0.5	0.623	0.479	0.578	0.758	0.505
$2 \rightarrow 3$	0.25	0.210	0.171	0.199	0.231	0.174
	0.5	0.226	0.179	0.211	0.238	0.189

Table 3.8: Standard errors for estimators of prevalence probabilities if all gap times were 3 months ($\Delta t = 0.25$), or 6 months ($\Delta t = 0.5$) with Z vector 0.

Δt	$P_{11}(1)$	$P_{11}(2)$	$P_{12}(1)$	$P_{12}(2)$	$P_{13}(1)$	$P_{13}(2)$
0.25	0.012	0.021	0.005	0.007	0.011	0.021
0.5	0.013	0.020	0.005	0.007	0.011	0.020

Table 3.9: Prevalence estimates for highest and lowest risk individuals.

Group	$\hat{P}_{11}(1)$	$\hat{P}_{11}(2)$	$\hat{P}_{12}(1)$	$\hat{P}_{12}(2)$	$\hat{P}_{13}(1)$	$\hat{P}_{13}(2)$
High risk	0.783	0.637	0.040	0.032	0.176	0.330
Low risk	0.865	0.820	0.091	0.087	0.043	0.089

3.5 Concluding remarks

In this chapter, we discussed the estimation of asymptotic covariance matrices for panel data and for continuous observations. We investigated the effect of gap times between visits on estimation in numerical studies, by considering the efficiency of maximum likelihood estimators of transition intensities and probabilities based on panel data, relative to the estimators obtained from continuous observation of individuals. The results show that in bi-directional models, the relative precision of transition intensity estimators drops off rapidly with increasing gap times between visits, but the decrease is smaller for transition probabilities. However, in progressive models, the loss of efficiency for transition intensities and probabilities is small. Therefore, we conclude that in bi-directional Markov models, transition intensities can not be estimated precisely when the observation times are too far apart.

The methods considered in this chapter can also be used to assess the relative precision of estimation for different gap times between visits, designing panel studies with the appropriate length and frequency of followup. For example, investigators can look at the effect of the frequency of visits on the precision of estimation of various model parameters, and design followup based on the precision of estimation and budgetary constraints.

Another use of asymptotic covariance matrices is to assess when msm or other software may encounter convergence problems when maximizing the likelihood functions. In these situations, we can compute the asymptotic covariance matrices for estimated transition intensities, and check the range of the corresponding eigenvalues. The existence of eigenvalues close to zero, indicates that certain combinations of the intensities can not be estimated precisely, and that the likelihood is flat in directions specified by the corresponding eigenvectors.

3.6 Appendix: Tables with Numerical Results on Relative Efficiency

Table 3.10: Ratios of asymptotic standard deviations (3.12) for estimators of transition intensities and probabilities in two-state Markov models with (1) $q_{12} = q_{21} = 1$ (M1a(i)), (2) $q_{12} = 1, q_{21} = 4$ (M1a(ii)), and (3) $q_{12} = 1, q_{21} = 0$ (M1b). Values of $P_{11}(t)$ in M1a(i) are $P_{11}(1) = 0.568$, $P_{11}(2) = 0.509$, $P_{11}(4) = 0.500$, in M1a(ii) they are $P_{11}(1) = 0.801$, $P_{11}(2) = 0.800$, $P_{11}(4) = 0.800$, and in M1b they are $P_{11}(1) = 0.368$, $P_{11}(2) = 0.135$, $P_{11}(4) = 0.018$. The total followup time is 4 years.

	Δt	q_{12}	q_{21}	$P_{11}(1)$	$P_{11}(2)$	$P_{11}(4)$
M1a(i)	0.25	0.89	0.87	0.64	0.64	0.64
	0.5	0.73	0.71	0.62	0.64	0.64
	1	0.40	0.40	0.50	0.62	0.60
	2	0.08	0.09	0.13	0.44	0.51
M1a(ii)	0.25	0.63	0.64	0.97	0.97	0.97
	0.5	0.30	0.30	0.87	0.88	0.88
	1	0.04	0.04	0.64	0.70	0.70
	2	4.2e-04	4.4e-04	0.01	0.52	0.54
M1b	0.25	0.99	-	0.99	0.99	0.99
	0.5	0.99	-	0.99	0.99	0.99
	1	0.96	-	0.96	0.96	0.96
	2	0.86	-	0.86	0.86	0.86

Table 3.11: Ratios of asymptotic standard deviations for estimators of transition intensities and probabilities in two-state Markov models with (1) $q_{12} = q_{21} = 1$ (M1a(i)), (2) $q_{12} = 1, q_{21} = 4$ (M1a(ii)), and (3) $q_{12} = 1, q_{21} = 0$ (M1b). Values of $P_{11}(t)$ in M1a(i) are $P_{11}(1) = 0.568$, $P_{11}(2) = 0.509$, $P_{11}(4) = 0.500$, $P_{11}(8) = 0.500$, in M1a(ii) they are $P_{11}(1) = 0.801$, $P_{11}(2) = 0.800$, $P_{11}(4) = 0.800$, $P_{11}(8) = 0.800$, and in M1b they are $P_{11}(1) = 0.368$, $P_{11}(2) = 0.013$, $P_{11}(4) = 0.002$, $P_{11}(8) = 3.3e - 04$. The total followup time is 10 years.

	Δt	q_{12}	q_{21}	$P_{11}(1)$	$P_{11}(2)$	$P_{11}(4)$	$P_{11}(8)$
M1a(i)	0.25	0.87	0.86	0.98	1.00	1.00	1.00
	0.5	0.70	0.69	0.92	0.98	0.98	0.98
	1	0.39	0.38	0.70	0.91	0.90	0.91
	2	0.08	0.08	0.19	0.19	0.58	0.75
M1a(ii)	0.25	0.62	0.64	0.95	0.95	0.95	0.95
	0.5	0.30	0.30	0.84	0.84	0.84	0.84
	1	0.04	0.04	0.59	0.66	0.66	0.66
	2	3.3e-04	3.4e-04	0.01	0.44	0.49	0.49
M1b	0.25	0.99	-	0.99	0.99	0.99	0.99
	0.5	0.99	-	0.99	0.99	0.99	0.99
	1	0.96	-	0.96	0.96	0.96	0.96
	2	0.85	-	0.85	0.85	0.85	0.85

Table 3.12: Ratios of asymptotic standard deviations for estimators of transition intensities and probabilities in three-state Markov models with (1) $q_{12} = 1, q_{21} = 2, q_{23} = 1$ (M2a) and (2) $q_{12} = 1, q_{21} = 0, q_{23} = 1$ (M2b). Values of $P_{11}(t)$ and $P_{13}(t)$ in M2a are $P_{11}(1) = 0.608$, $P_{11}(2) = 0.462$, $P_{11}(4) = 0.270$, $P_{13}(1) = 0.178$, $P_{13}(2) = 0.369$, $P_{13}(4) = 0.631$, and in M2b are $P_{11}(1) = 0.368$, $P_{11}(2) = 0.135$, $P_{11}(4) = 0.018$, $P_{13}(1) = 0.264$, $P_{13}(2) = 0.594$, $P_{13}(4) = 0.908$. The total followup time is 4 years.

	Δt	q_{12}	q_{21}	q_{23}	$P_{11}(1)$	$P_{11}(2)$	$P_{11}(4)$	$P_{13}(1)$	$P_{13}(2)$	$P_{13}(4)$
M2a	0.25	0.80	0.74	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	0.5	0.56	0.49	0.98	0.97	1.0	1.0	1.0	1.0	1.0
	1	0.27	0.23	0.90	0.85	0.99	1.0	0.91	1.0	1.0
	2	0.13	0.11	0.72	0.58	0.89	1.0	0.67	0.99	1.0
M2b	0.25	0.99	-	1.0	0.99	0.99	0.99	1.0	1.0	1.0
	0.5	0.99	-	0.99	0.99	0.99	0.99	0.99	0.99	0.99
	1	0.97	-	0.96	0.97	0.97	0.97	0.99	0.99	0.99
	2	0.87	-	0.84	0.87	0.87	0.87	0.95	0.95	0.95

Table 3.13: Ratios of asymptotic standard deviations for estimators of transition intensities and probabilities in three-state Markov models as in Figure1(b) with (1) $q_{12} = 1, q_{21} = 2, q_{23} = 1$ (M2a) and (2) $q_{12} = 1, q_{23} = 1$ M2b. Values of $P_{11}(t)$ and $P_{13}(t)$ in M2a are $P_{11}(2) = 0.462$, $P_{11}(4) = 0.270$. $P_{11}(8) = 0.092$, $P_{13}(2) = 0.369$, $P_{13}(4) = 0.631$, $P_{13}(8) = 0.874$ and in M2b are $P_{11}(2) = 0.013$, $P_{11}(4) = 0.002$. $P_{11}(8) = 7.7e-06$, $P_{13}(2) = 0.594$, $P_{13}(4) = 0.908$, $P_{13}(8) = 0.997$. The total followup time is 10 years.

	Δt	q_{12}	q_{21}	q_{23}	$P_{11}(2)$	$P_{11}(4)$	$P_{11}(8)$	$P_{13}(2)$	$P_{13}(4)$	$P_{13}(8)$
M2a	0.25	0.79	0.73	0.99	0.99	1.00	1.00	1.00	1.00	1.00
	0.5	0.55	0.48	0.95	0.98	0.99	1.00	1.00	1.00	1.00
	1	0.26	0.22	0.85	0.94	0.98	0.99	0.98	1.00	1.00
	2	0.13	0.10	0.66	0.81	0.95	0.98	0.90	1.00	1.00
M2b	0.25	0.99	-	0.99	0.99	0.99	0.99	0.99	0.99	0.99
	0.5	0.99	-	0.98	0.99	0.99	0.99	0.99	0.99	0.99
	1	0.96	-	0.94	0.96	0.96	0.96	0.98	0.98	0.98
	2	0.86	-	0.81	0.86	0.86	0.86	0.93	0.93	0.93

Table 3.14: Ratios of asymptotic standard deviations for estimators of transition intensities and regression coefficients in three-state Markov models with (1) $q_{12} = 1, q_{21} = 2, q_{23} = 1$, $\beta_{12} = -0.2$, $\beta_{21} = -0.5$, $\beta_{23} = -0.9$ (M3a) and (2) $q_{12} = 1, q_{21} = 0, q_{23} = 1$, $\beta_{12} = -0.2$, $\beta_{21} = 0$, $\beta_{23} = -0.9$ (M3b).The total followup time is 4 years.

	Δt	q_{12}	q_{21}	q_{23}	β_{12}	β_{21}	β_{23}
M3a	0.25	0.80	0.74	1.0	0.85	0.80	1.0
	0.5	0.56	0.49	0.98	0.66	0.59	1.0
	1	0.27	0.23	0.90	0.35	0.30	1.0
	2	0.13	0.11	0.72	0.15	0.14	0.90
M3b	0.25	0.99	-	1.0	1.0	-	1.0
	0.5	0.99	-	0.99	0.99	-	1.0
	1	0.97	-	0.96	0.98	-	1.0
	2	0.87	-	0.84	0.90	-	0.99

Table 3.15: Continuation of Table 3.14: Ratios of asymptotic standard deviations for estimators of transition probabilities in three-state Markov models when an individual has $Z = 1$. Values of P_{11} and P_{13} in M3a are $P_{11}(1) = 0.625$, $P_{11}(2) = 0.517$, $P_{11}(4) = 0.384$, $P_{13}(1) = 0.083$, $P_{13}(2) = 0.202$, $P_{13}(4) = 0.309$ and in M3b are $P_{11}(1) = 0.440$, $P_{11}(2) = 0.194$, $P_{11}(4) = 0.038$, $P_{13}(1) = 0.112$, $P_{13}(2) = 0.310$, $P_{13}(4) = 0.647$. The total followup time is 4 years.

	Δt	$P_{11}(1)$	$P_{11}(2)$	$P_{11}(4)$	$P_{13}(1)$	$P_{13}(2)$	$P_{13}(4)$
M3a	0.25	1.0	1.0	1.0	1.0	1.0	1.0
	0.5	0.98	1.0	1.0	1.0	1.0	1.0
	1	0.84	1.0	1.0	1.0	1.0	1.0
	2	0.47	0.86	1.0	1.0	1.0	1.0
M3b	0.25	0.99	0.99	0.99	1.0	1.0	1.0
	0.5	0.99	0.99	0.99	1.0	1.0	1.0
	1	0.97	0.97	0.97	1.0	1.0	1.0
	2	0.87	0.87	0.87	1.0	1.0	1.0

Table 3.16: Ratios of asymptotic standard deviations for estimators of transition intensities and regression coefficients in three-state Markov models with (1) $q_{12} = 1, q_{21} = 2, q_{23} = 1$, $\beta_{12} = -0.2$, $\beta_{21} = -0.5$, $\beta_{23} = -0.9$ (M3a) and (2) $q_{12} = 1, q_{21} = 0, q_{23} = 1$, $\beta_{12} = -0.2$, $\beta_{21} = 0$, $\beta_{23} = -0.9$ (M3b). The total followup time is 10 years.

	Δt	q_{12}	q_{21}	q_{23}	β_{12}	β_{21}	β_{23}
M3a	0.25	0.79	0.73	0.99	0.82	0.77	0.99
	0.5	0.55	0.48	0.95	0.60	0.54	0.98
	1	0.26	0.22	0.85	0.30	0.26	0.91
	2	0.13	0.10	0.66	0.14	0.12	0.76
M3b	0.25	0.99	-	0.99	0.99	-	0.99
	0.5	0.99	-	0.98	0.99	-	0.99
	1	0.96	-	0.94	0.97	-	0.97
	2	0.86	-	0.81	0.89	-	0.90

Table 3.17: Continuation of Table 3.16: Ratios of asymptotic standard deviations for estimators of transition probabilities in three-state Markov models with $Z = 1$. Values of P_{11} and P_{13} in M3a are $P_{11}(2) = 0.517$, $P_{11}(4) = 0.384$. $P_{11}(8) = 0.215$, $P_{13}(2) = 0.202$, $P_{13}(4) = 0.403$, $P_{13}(8) = 0.666$ and in M3b are $P_{11}(2) = 0.194$, $P_{11}(4) = 0.038$, $P_{11}(8) = 0.001$, $P_{13}(2) = 0.310$, $P_{13}(4) = 0.647$, $P_{13}(8) = 0.924$. The total followup time is 10 years.

	Δt	$P_{11}(2)$	$P_{11}(4)$	$P_{11}(8)$	$P_{13}(2)$	$P_{13}(4)$	$P_{13}(8)$
M3a	0.25	0.99	1.00	1.00	1.00	1.00	1.00
	0.5	0.99	1.00	1.00	1.00	1.00	1.00
	1	0.93	0.99	1.00	0.99	1.00	1.00
	2	0.76	0.93	1.00	0.85	1.00	1.00
M3b	0.25	0.99	0.99	0.99	1.00	1.00	1.00
	0.5	0.99	0.99	0.99	1.00	1.00	1.00
	1	0.97	0.97	0.97	0.99	0.99	0.99
	2	0.90	0.90	0.90	0.98	0.99	0.99

Table 3.18: Ratios of asymptotic standard deviations for estimators of transition intensities and probabilities in three-state Markov models with (1) $q_{12} = 1, q_{21} = 2, q_{13} = 0.25, q_{23} = 0.5$ (M4a) and (2) $q_{12} = 0.5, q_{21} = 0, q_{13} = 0.25, q_{23} = 0.5$ (M4b). Values of $P_{11}(t)$ and $P_{13}(t)$ in M4a are $P_{11}(1) = 0.515$, $P_{11}(2) = 0.364$. $P_{11}(4) = 0.188$, $P_{13}(1) = 0.263$, $P_{13}(2) = 0.469$, $P_{13}(4) = 0.725$ and in M4b are $P_{11}(1) = 0.472$, $P_{11}(2) = 0.223$. $P_{11}(4) = 0.045$, $P_{13}(1) = 0.259$, $P_{13}(2) = 0.487$, $P_{13}(4) = 0.779$. The total followup time is 4 years.

	Δt	q_{12}	q_{21}	q_{13}	q_{23}	$P_{11}(1)$	$P_{11}(2)$	$P_{11}(4)$	$P_{13}(1)$	$P_{13}(2)$	$P_{13}(4)$
M4a	0.25	0.75	0.74	0.83	0.80	0.99	1.0	1.0	1.0	1.0	1.0
	0.5	0.48	0.48	0.66	0.62	0.97	1.0	1.0	0.99	1.0	1.0
	1	0.14	0.15	0.40	0.37	0.80	0.96	1.0	0.93	1.0	1.0
	2	7e-03	8e-03	0.04	0.05	0.08	0.80	0.97	0.56	0.99	1.0
M4b	0.25	0.97	-	0.94	0.98	1.0	1.0	1.0	0.98	1.0	1.0
	0.5	0.93	-	0.88	0.95	1.0	1.0	1.0	0.96	1.0	1.0
	1	0.83	-	0.75	0.90	0.99	0.99	0.99	0.90	0.99	1.0
	2	0.60	-	0.51	0.76	0.93	0.93	0.93	0.70	0.90	1.0

Table 3.19: Continuation of Table 3.18: Values of $P_{22}(t)$ and $P_{23}(t)$ in M4a are $P_{22}(1) = 0.237$, $P_{22}(2) = 0.155$, $P_{22}(4) = 0.080$, $P_{23}(1) = 0.318$, $P_{23}(2) = 0.510$, $P_{23}(4) = 0.746$ and in M4b are $P_{22}(1) = 0.606$, $P_{22}(2) = 0.368$, $P_{22}(4) = 0.135$, $P_{23}(1) = 0.393$, $P_{23}(2) = 0.632$, $P_{23}(4) = 0.865$. The total followup time is 4 years.

	Δt	$P_{22}(1)$	$P_{22}(2)$	$P_{22}(4)$	$P_{23}(1)$	$P_{23}(2)$	$P_{23}(4)$
M4a	0.25	0.94	0.97	0.99	0.85	0.86	0.86
	0.5	0.81	0.91	0.96	0.68	0.70	0.70
	1	0.42	0.73	0.84	0.33	0.38	0.41
	2	0.03	0.26	0.58	0.02	0.03	0.04
M4b	0.25	0.98	0.98	0.98	0.98	0.98	0.98
	0.5	0.95	0.95	0.95	0.95	0.95	0.95
	1	0.90	0.90	0.90	0.90	0.90	0.90
	2	0.76	0.76	0.76	0.76	0.76	0.76

Table 3.20: Ratios of asymptotic standard deviations for estimators of transition intensities and probabilities in three-state Markov models with (1) $q_{12} = 1, q_{21} = 2, q_{13} = 0.25, q_{23} = 0.5$ (M4a) and (2) $q_{12} = 0.5, q_{21} = 0, q_{13} = 0.25, q_{23} = 0.5$ (M4b). Values of $P_{11}(t)$ and $P_{13}(t)$ in M4a are $P_{11}(2) = 0.364$, $P_{11}(4) = 0.188$. $P_{11}(8) = 0.050$, $P_{13}(2) = 0.469$, $P_{13}(4) = 0.725$, $P_{13}(8) = 0.926$ and in M4b are $P_{11}(2) = 0.223$, $P_{11}(4) = 0.500$. $P_{11}(8) = 0.002$, $P_{13}(2) = 0.487$, $P_{13}(4) = 0.779$, $P_{13}(8) = 0.966$. The total followup time is 10 years.

	Δt	q_{12}	q_{21}	q_{13}	q_{23}	$P_{11}(2)$	$P_{11}(4)$	$P_{11}(8)$	$P_{13}(2)$	$P_{13}(4)$	$P_{13}(8)$
M4a	0.25	0.74	0.74	0.82	0.79	0.99	0.99	0.99	0.99	1.00	1.00
	0.5	0.48	0.48	0.65	0.61	0.98	0.99	0.99	0.99	1.00	0.99
	1	0.14	0.15	0.39	0.36	0.92	0.98	0.99	0.98	1.00	0.99
	2	7.2e-03	8e-03	0.04	0.05	0.74	0.90	0.95	0.90	0.99	0.97
M4b	0.25	0.96	-	0.94	0.97	0.99	0.99	0.99	0.99	0.99	0.99
	0.5	0.92	-	0.87	0.93	0.99	0.99	0.99	0.98	0.99	0.98
	1	0.82	-	0.75	0.87	0.98	0.98	0.98	0.95	0.99	0.95
	2	0.60	-	0.52	0.72	0.91	0.91	0.91	0.85	0.97	0.89

Table 3.21: Continuation of Table 3.20: Values of $P_{22}(t)$ and $P_{23}(t)$ in M4a are $P_{22}(2) = 0.155$, $P_{22}(4) = 0.080$, $P_{22}(8) = 0.021$, $P_{23}(2) = 0.510$, $P_{23}(4) = 0.746$, $P_{23}(8) = 0.932$ and in M4b are $P_{22}(2) = 0.368$, $P_{22}(4) = 0.135$, $P_{22}(8) = 0.018$, $P_{23}(2) = 0.632$, $P_{23}(4) = 0.865$, $P_{23}(8) = 0.982$. The total followup time is 10 years.

	Δt	$P_{22}(2)$	$P_{22}(4)$	$P_{22}(8)$	$P_{23}(2)$	$P_{23}(4)$	$P_{23}(8)$
M4a	0.25	0.96	0.98	0.99	0.95	0.98	0.99
	0.5	0.89	0.94	0.97	0.89	0.95	0.98
	1	0.72	0.82	0.90	0.72	0.85	0.92
	2	0.26	0.56	0.71	0.43	0.60	0.75
M4b	0.25	0.97	0.97	0.97	0.97	0.97	0.97
	0.5	0.93	0.93	0.93	0.93	0.93	0.93
	1	0.87	0.87	0.87	0.87	0.87	0.87
	2	0.72	0.72	0.72	0.72	0.72	0.72

Chapter 4

Estimation of Prevalence Probabilities and Failure Time Distributions

4.1 Introduction

There has been considerable interest in the distribution of times to certain events in order to assess disease progression. As we know, in multistate models, the states are often based on underlying continuous measures, and failure times are often defined as the times of entry to certain states. Failure time distributions can also be estimated using survival analysis methods, and one way of assessing a multistate model's fit is to compare the estimates obtained by it and by a survival method. Similarly, state prevalence functions are also used to check model fit, as we discussed in Section 2.3.2. Titman and Sharples

(2010a) and Jackson (2011) mentioned such model checking methods for multistate models. However, when there is considerable variability of observation times across individuals, we have more chance of dependent observation times or dependent censoring. The estimation of failure time distributions and prevalence probabilities by empirical methods described in Section 2.3 can be biased in these situations.

Our purpose in this chapter is to discuss the estimation of marginal process features such as failure time distributions and prevalence probabilities, and to present a method to correct the bias of estimation in the presence of dependent observation times, Then, we apply the methodology presented to assess the fit of the multistate model described in Section 3.4.

The remainder of this chapter is as follows. Section 4.2 considers the estimation of failure time distribution and prevalence probabilities based on multistate models and survival methods. Section 4.3 discusses the effects of dependent observation times on estimation of a failure time distribution and prevalence functions and presents a novel method that adjusts the bias due to dependent observation. It also includes a simulation study for comparison of multistate and survival models. Section 4.4 gives applications of our methodology to CANOC data.

4.2 Estimation of prevalence probabilities

In this section, we describe the estimation of failure time distribution in both multistate and survival models for intermittently observed data. In addition, we consider the estimation of prevalence functions for multistate models when we have intermittent observation.

4.2.1 Multistate models

In multistate models, we define a prevalence function as $P_{y_0 r}(t) = P(Y(t) = r|Y(0) = y_0)$. However, for many disease processes, all individuals are in the same initial state, say state 1, at the time of entry into the study. So, the prevalence function can be denoted as $P_{1r}(t) = P(Y(t) = r|Y(0) = 1)$. The estimates $\hat{P}_{1r}(t)$ can be computed for Markov models using the maximum likelihood estimates discussed in the Section 2.2. However, if the processes involve internal covariates, direct calculation of prevalence probabilities is complicated.

Sometimes in multistate models, the distributions of time to certain events are of interest. For instance, we might be interested in the failure time T which is the time of first entry to a specific state. Then, in this case, we make that state absorbing, and define the failure time distribution $F(t|z) = P(T \leq t|z)$, for a given fixed covariate vector z. So, in this situation, the failure time distribution can be considered as a special case of prevalence functions. For example, we can use the model in Figure 3.1-b for analysis of viral load processes in the CANOC data, and consider $F(t) = P_{13}(t)$ for the time T to a viral rebound. In disease processes, sometimes multiple absorbing states can occur. This is often called competing risks.

4.2.2 Survival methods

Suppose that T_i is a survival time for individual i with the distribution function $P(T_i \leq t|z_i;\theta) = F(t|z_i;\theta)$, where θ is a vector of parameters. As in Section 2.1, suppose that individual i is seen at times $t_{ij}(j = 1, ..., m_i)$. We may then know that $T_i \in (L_i, R_i]$, where $L_i = t_{i,j-1}$ and $R_i = t_{ij}$. Then, the likelihood function for θ based on n independent individuals is, under suitable assumptions,

$$L(\theta) = \prod_{i=1}^{n} \prod_{j=1}^{m_i+1} \{F(t_{i,j}|Z_i;\theta) - F(t_{i,j-1}|Z_i;\theta)\}^{\delta_{ij}} = \prod_{i=1}^{n} \{F(R_i|z_i;\theta) - F(L_i|z_i;\theta)\}, \quad (4.1)$$

where $\delta_{ij} = I(t_{i,j-1} < T_i \leq t_{ij})$ and $t_{im_{i+1}} = \infty$. Failure time distribution $F(t|z;\theta)$ can be estimated by maximizing $L(\theta)$. Turnbull (1976) discussed the case where $F(t)$ is treated nonparametrically, where maximization of this likelihood produces the Turnbull estimator (Lawless (2003), Section 3.5.3).

This likelihood function (4.1) is not valid if observation times t_{ij} are not conditionally independent of T_i given z_i. When there is a biomarker process $\{Y_i(t), t > 0\}$ this assumption is for example violated when t_{ij} depends on $Y_i(t_{i,j-1})$ or previously observed process history, $\bar{Y}_i(t_{i,j-1}) = \{Y_i(t_{il}), l = 0, 1, ..., j-1\}$, and T_i is also related to the multistate process. Then, T_i and t_{ij} are not conditionally independent given z_i and that $T_i > t_{i,j-1}$. In Section 4.3, we discuss a method to adjust for the estimation bias that arises from the violation of the independent observation times assumption. We remark that in the setting just described, estimation for the multistate model would not be subject to bias in the case where t_{ij} depended on $Y_i(t_{i,j-1})$, as long as it is independent of $\{Y_i(s), s > t_{i,j-1}\}$ given $\bar{Y}_i(t_{i,j-1})$.

4.2.3 Empirical prevalence methods

Suppose that individual $i (i = 1, ..., n)$ is observed at times $t_{ij}, j = 1, ..., m_i$, and all individuals start from state 1 at time 0. For a pre-specified set of times $t_1 < t_2 < ... < t_m$, the nonparametric prevalence probability estimates (naive estimates), $\tilde{P}_{1r}(t)$, are obtained by dividing the number of individuals in state r at time t by the number at risk,

$$\tilde{P}_{1r}(t_j) = \sum_{i=1}^{n} I(Y_i(t_j) = r)\delta_i(t_j) / \sum_{i=1}^{n} \delta_i(t_j), \quad r = 1, ..., R, \tag{4.2}$$

where $\delta_i(t_j) = I(t_{im_i} \geq t_j)$ indicates that individual i was still being followed at time t_j. In the case of a regular observational scheme, when all individuals are observed at the same time points $t_1, ..., t_m$, we can compute the prevalence probabilities at these times. However, when the observation scheme is irregular, we can no longer choose a set of times that all individuals are observed. In this case, we can use one of the interpolation methods described in Section 2.3.2 of Chapter 2 to calculate the prevalences at times of interest. The msm package uses formula (4.2) to compute the empirical prevalences.

The empirical prevalence estimates can be used to assess the fit of multistate models by comparison of them with the corresponding model-based estimates, and the agreement of model-based estimates with the nonparametric estimates can be checked by constructing confidence intervals based on either type of estimates. For example, we can construct pointwise confidence intervals with the model based prevalence estimates by considering standard errors obtained from the bootstrap or by using asymptotic covariance matrix estimates discussed in Chapter 3. Bootstrap methods are very time-consuming, but can be implemented by the function prevalence.msm in the msm package.

For the prevalence estimates (4.2), we assume that the gap times $\Delta t_{ij} = t_{ij} - t_{i,j-1}$ are conditionally independent of $\{Y_i(t), t > t_{i,j-1}\}$ given $t_{i,j-1}$ and fixed covariate z_i. However, this assumption is often violated and the gap times might be related to the previous process history and this can produce a bias in estimating the prevalence probabilities. We propose a remedy for this issue in Section 4.3. There are some other problems in estimating the probabilities with intermittent observation that we will discuss later in Chapter 5.

4.3 Process-dependent observation times

In some studies, the gap times between visits depend on previous process history. For example, $\Delta t_{ij} = t_{ij} - t_{i,j-1}$ may depend on $Y_i(t_{i,j-1})$. This dependency does not have any effect on estimation in Markov models. However, it can produce bias in estimation for failure time and nonparametric prevalence methods that we discussed in Sections 4.2.2 and 4.2.3.

There are different approaches for addressing the problem of the bias in estimation of nonparametric prevalences in this context. One approach is to use a shared random effect model for the observation and response processes, and assume that the underlying response or event process is independent of the observation process given the unobserved random effects. Sun et al. (2004), Liang et al. (2009), and Zhu et al. (2011) discussed this method. Another approach is to use inverse intensity weights, described below. Lin et al. (2004), Buzkova and Lumley (2007, 2009) and Pullenayegum and Feldman (2013) considered this approach. This method provides unbiased estimating functions for marginal features of the process, and can be adapted to the estimation of prevalence probabilities $P_{1r}(t)$ and failure time distributions $F(t)$. In this section, we describe this approach, and compare different estimates under various scenarios of dependent observation times.

4.3.1 Inverse intensity weight estimation (IIW)

We define the counting process $\{N_i(t), t > 0\}$ for observation times; that is, $N_i(t)$ is the number of observation times t_{ij} by time t for individual i. We also define the intensity function $\lambda_i(t|\tilde{H}_i(t))$ for the observation process, where $\tilde{H}_i(t)$ includes the previous history of the multistate process and the observation process; $\tilde{H}_i(t) = \{\bar{N}_i(t^-), \bar{Y}_i(t^-)\}$, where

$\bar{N}_i(t) = \{N_i(s), 0 < s \leq t\}$ and $\bar{Y}_i(t) = \{Y_i(s), 0 < s \leq t\}$. External covariates are also understood to be part of $\tilde{H}_i(t)$. It is further assumed that $dN_i(t)$ is independent of $Y_i(t)$, given $\tilde{H}_i(t)$. The IIW method also requires the assumption that $\lambda_i(t|\tilde{H}_i(t)) = \lambda_i(t|\tilde{H}_i^{obs}(t))$, where $\tilde{H}_i^{obs}(t)$ is the observed history; suppressing any covariates for now, $\tilde{H}_i^{obs}(t) = \{t_{ij}, Y_i(t_{ij}), \text{ for } j = 1, ..., N_i(t^-)\}$. This assumption can be violated, especially when Δt_{ij}'s are large, but models that do not assume it require uncheckable assumptions. For convenience we write $\lambda_i(t)$ for $\lambda_i(t|\tilde{H}_i^{obs}(t))$. Finally, it is assumed that $\lambda_i(t)$ is either (a) positive for all $\tilde{H}_i^{obs}(t)$ or (b) zero for all $\tilde{H}_i^{obs}(t)$ at a given t. In the case of (b), we are not able to estimate $P_{1r}(t)$ for that value of t.

Assuming that we know the intensity $\lambda_i(t)$, an estimating function for $P_{1r}(t;\theta) = P(Y(t) = r;\theta)$ is

$$U_w(\theta) = \sum_{i=1}^{n} \int_0^{\tau} \lambda_i(t)^{-1}\{Y_{ir}(t) - P_{1r}(t;\theta)\} g_i(t,\theta) dN_i(t), \tag{4.3}$$

where $Y_{ir}(t) = I(Y_i(t) = r)$, $g_i(t;\theta)$ is a $p \times 1$ vector of functions of the same dimension as θ, and τ is an administrative upper limit on followup. Lin et al. (2004) showed that $E\{U_w(\theta)\} = 0$ by using iterated expectations under regularity conditions, as follows:

$$\begin{aligned} E\{U_w(\theta)\} &= E\{\sum_{i=1}^{n} \int_0^{\tau} E[g_i(t;\theta)\{Y_{ir}(t) - P_{1r}(t;\theta)\}\frac{dN_i(t)}{\lambda_i(t)}|\tilde{H}_i(t^-)]\} \qquad (4.4) \\ &= E\{\sum_{i=1}^{n} \int_0^{\tau} g_i(t;\theta) E\{Y_{ir}(t) - P_{1r}(t;\theta)|H_i(t^-)\}\frac{E[dN_i(t)|\tilde{H}_i(t^-)]}{\lambda_i(t)}\} \\ &= \sum_{i=1}^{n} \int_0^{\tau} g_i(t;\theta) E[E\{Y_{ir}(t) - P_{1r}(t;\theta)\}|\tilde{H}_i(t^-)] \\ &= \sum_{i=1}^{n} \int_0^{\tau} g_i(t;\theta) E\{Y_{ir}(t) - P_{1r}(t;\theta)\} = 0. \end{aligned}$$

Solving $U_w(\theta) = 0$ produces a consistent estimator of θ with asymptotic normal distribution, under mild regularity conditions (e.g. White (1982)).

An examination of (4.3) and (4.4) with $\lambda_i^{-1}(t)$ dropped shows why empirical estimators (4.2) can be biased. We note that (4.2) arises from equating (4.3) to zero and associating θ with the (non-countable) set of probabilities $P_{1r}(t)$, so when $\lambda_i^{-1}(t)$ is dropped from (4.3), the inner expectation in (4.4) produces $\lambda_i(t|\tilde{H}_i^{obs}(t))$, and $Y_{ir}(t)$ is not independent of this function of $\tilde{H}_i^{obs}(t)$ in general. Consequently, the expectation (4.4) is not zero, though it is in the case where $\lambda_i(t|\tilde{H}_i(t^-))$ does not involve $\tilde{H}_i(t^-)$.

An estimating function of the form (4.3) is given by the derivative of the weighted sum of squares, $S_w(\theta)$,

$$\begin{aligned} S_w(\theta) &= \sum_{i=1}^{n} \int_0^{\tau} w_i(t_{ij})\{Y_{ir}(t_{ij}) - P_{1r}(t_{ij};\theta)\}^2 dN_i(t) \qquad (4.5) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m_i} w_i(t_{ij})\{Y_{ir}(t_{ij}) - P_{1r}(t_{ij};\theta)\}^2 \end{aligned}$$

with respect to θ, where $w_i(t) = \lambda_i(t)^{-1}$. So one choice of $g_i(t;\theta)$ is $g_i(t,\theta) = \partial P_{1r}(t;\theta)/\partial\theta$. Another alternative is the binomial generalized linear model choice $g_i(t;\theta) = \frac{\partial P_{1r}(t;\theta)}{\partial\theta}/[P_{1r}(t;\theta)$ $(1 - P_{1r}(t;\theta))]$.

An important issue with using (4.3) is that, the observation process intensity $\lambda_i(t)$ is unknown and needed to be estimated via some family of models. We consider this next, where $Z_i(t)$ contains features of $\tilde{H}_i^{obs}(t)$.

4.3.2 Estimation of the observation process intensity

One approach is to model $\lambda_i(t)$ by the semiparametric Cox model: $\lambda_i(t|\tilde{H}_i^{obs}(t)) = \lambda_0(t)exp($ $\beta^T Z_i(t))$, Buzkova and Lumley (2007, 2009) used this model and showed that the baseline intensity $\lambda_0(t)$ can be dropped from the estimating function. However, with intermittently observed data, the observation intensities $\lambda_i(t)$ typically depend more on the time since the last visit, $t - t_{iN_i(t^-)}$, rather than observation time t. Thus, we use the Semi-Markov model

$$\lambda(t|\tilde{H}_i(t)) = \lambda_0(B_i(t))exp(\beta^T z_i(t)), \tag{4.6}$$

where $B_i(t) = t - t_{iN_i(t^-)}$, as our basic model. We assume that $z_i(t) = z_{ij} = z_i(t_{i,j-1})$, $t_{i,j-1} \leq t < t_{ij}$; that is, $z_i(t)$ only depends on the information up to the previous observation time. We let $b_{ij} = t_{ij} - t_{i,j-1}$ denote the observed gap times between visits and $b_{im_i+1} = \tau_i - t_{im_i}$, where τ_i is the end of followup time for individual i. When the b_{ij}s are different distinct values, the estimate of the observation process intensity at the observed gap time b_{ij} is

$$\hat{\lambda}_0(b_{ij}) = \frac{1}{\sum_{l=1}^{n}\sum_{r=1}^{m_l+1} I(b_{lr} \geq b_{ij})exp(\hat{\beta} z_{lr})}, \tag{4.7}$$

as discussed in Chapter 4 of Cook and Lawless (2007). For continuous gap time models, we should have distinct b_{ij}'s. However, a few ties can be handled by standard Cox model methods. Then, the weights in (4.5) are estimated as

$$\hat{w}_i(t_{ij}) = \{\lambda_0(\hat{b}_{ij})exp(\hat{\beta}^T z_{ij})\}^{-1}. \tag{4.8}$$

4.3.3 Nonparametric estimation of prevalence probabilities and failure time distributions

Suppose $s_1 < s_2 < ... < s_m$ are the distinct values among the set of observation times, $\{t_{ij}, j = 1, ..., m_i, i = 1, ..., n\}$. For a given r, we equate θ_{s_l} with $P_{1r}(s_l)$, and minimize (4.5) for θ. We can rewrite (4.5) as

$$S_w(\theta) = \sum_{l=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{m_i} I(t_{ij} = s_l)w_i(t_{ij})\{Y_{ir}(t_{ij}) - P_{1r}(s_l)\}^2, \tag{4.9}$$

and minimizing it gives nonparametric estimates $\tilde{P}_{1r}(s_l)$ as

$$\tilde{P}_{1r}(s_l) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i} I(t_{ij} = s_l) w_i(t_{ij}) Y_{ir}(t_{ij})}{\sum_{i=1}^{n}\sum_{j=1}^{m_i} I(t_{ij} = s_l) w_i(t_{ij})} = \frac{d_r(s_l)}{w_+(s_l)}. \tag{4.10}$$

However, there are problems concerning nonparametric prevalence estimates (4.10). The first issue is that, there are often few t_{ij} equal to any given s_l, and in extreme cases each s_l may only be associated with just one t_{ij} and (4.10) is 0 or 1. Thus, the estimates (4.10) are discrete and we need to use smoothing or grouping to obtain reasonable estimates. In order to overcome this issue, a simple approach is to break time into intervals and estimate prevalence probabilities for each interval. Suppose a grid of values a_l with $0 = a_0 < a_1 < ... < a_m$ partitions the time axis, and denote intervals $I_l = (a_{l-1}, a_l]$ for $l = 1, ..., m+1$. For each interval I_l, we let $t_{m_l} = 0.5(a_{l-1} + a_l)$ denote the midpoint of interval I_l. Now, we can define the nonparametric prevalence estimates for time interval I_l as

$$\tilde{P}_{1r}(t_{m_l}) \doteq \tilde{P}_{1r}(t \in I_l) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i} I(t_{ij} \in I_l) \hat{w}_i(t_{ij}) Y_{ir}(t_{ij})}{\sum_{i=1}^{n}\sum_{j=1}^{m_i} I(t_{ij} \in I_l) \hat{w}_i(t_{ij})}. \tag{4.11}$$

If we choose intervals I_l so that each interval contains a substantial number of observation times t_{ij}, then the estimates (4.11) are less variable, though it is better not to use intervals that are too wide. We may also want to smooth estimates (4.11) in some cases, using for example a standard procedure such as loess in R.

Another problem with estimates (4.10) or (4.11) is that sometimes the prevalence function $P_{1r}(t)$ needs to be monotonic, but it is not. For instance, when state r is an absorbing state, $P_{1r}(t)$ has to be nondecreasing with respect to time t. In this situation, we can equate $Y_{ir}(t_{ij})$ and $P_{1r}(s_l)$ in (4.9) with $Y_i(t) = I(T_i \leq t)$ and $F(t)$, respectively. Then, we can minimize (4.9) subject to the F_{s_l} being non-decreasing for $l = 1, 2, ...$ This is equivalent to minimizing $S_w^*(\theta)$ which is

$$S_w^*(\theta) = \sum_{l=1}^{m} w_+(s_l)\{\bar{y}_l - F(s_l)\}^2, \tag{4.12}$$

where $w_+(s_l) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} I(t_{ij} = s_l) w_i(t_{ij})$ and $\bar{y}_l = \sum_{i=1}^{n} \sum_{j=1}^{m_i} I(t_{ij} = s_l) w_i(t_{ij}) Y_i(t_{ij}) / w_+(s_l)$. Now, minimizing $S_w^*(\theta)$ subject to the constraint, yields the isotonic regression estimate, (e.g. Sun (2006), p. 210)

$$\tilde{F}(s_l) = max_{r \leq l} min_{u \geq l} \{ \frac{\sum_{v=r}^{u} d_r(s_v)}{\sum_{v=r}^{u} w_+(s_v)} \}. \tag{4.13}$$

Thus, we can get the raw estimates of (4.11) and then use the isotonic regression estimate of (4.13) as nonparametric estimates of $F(s_l)$. The isotonic regression estimates are easily computed utilizing the "isoreg" function in R.

In the discrete time framework, we assume that $s_1, ..., s_m$ are the actual times of visits which are prespecified. In the simulation studies of Section 4.3.4, we will use the discrete setting.

4.3.4 Simulation studies

Now we conduct some simulation studies to assess the performance of the proposed IIW method, and to evaluate its robustness with respect to different types of model misspecification for estimating weights. Our primary aim is to compare the IIW method with the naive method which does not incorporate weights to adjust for the process-dependent observation times. We do this for models that do not involve covariates, for simplicity. In Section 4.4 we analyze CANOC data and consider models with covariates. We show below that large biases in the prevalence estimates can result if the dependency is not accounted for when it should be. We also investigate the performance of the IIW prevalence probability estimates when the estimated weights are obtained from misspecified models.

Simulation 1

This simulation is motivated by the CANOC observational study, in which individuals observation times are related to previous states (viral loads). In Section 4.4.2 of this chapter, we will show that the states individuals were in at their previous observation times are highly associated with the gap times between visits. Briefly, we will consider states based on viral load values defined earlier in Section 3.4.1, and show that individuals who were in state 2 at their previous visit time tend to have shorter gap times compared to those whose last state was 1. The main objectives of the following simulation are, first, to examine bias induced by the state dependency of observation times and second, to assess the performance of the IIW method in the presence of process-dependent observation times.

In this simulation, we consider only one set of $n = 5000$ sample paths, and the bias and standard deviation are considered later in simulation 2. To keep complications to a minimum, we assume that individuals can be seen only at discrete times $t = 0, 0.1, 0.2, 0.3, ..., 4$. We considered the multistate model in Figure 3.1(b) with transition intensities $\lambda_{12} = 1, \lambda_{21} = 2$, and $\lambda_{23} = 1$, and we assumed that all individuals are in state 1 at time 0. For the process-dependent observation times, we supposed that the discrete gap times Δt_{ij} between visits depend on the states individuals were in at their previous observation times. We generated $\Delta t_{ij} = b_{ij}|Y_i(t_{ij-1}) = 1$ and $\Delta t_{ij} = b_{ij}|Y_i(t_{ij-1}) = 2$ from geometric distrubutions $(1-p)^{b-1}p, \quad b = 1, 2, ..$ with $p = 1/6$ and $1/3$ respectively. Here, for simplicity of notation $b = 1, 2, 3, ...$ correspond to $0.1, 0.2, 0.3, ...$ Note that an individual is in state 1 at a visit, then the mean time to their next visit is 0.6, whereas it is 0.3 if they are in state 2.

In all the simulation studies of this section, we assume that the visit time processes are discrete; that is the actual visit times are prespecified at time $0.1, 0.2, 0.3, ..., 4$. In order to compute the IIW prevalence estimates in this discrete time framework, we assume all the possible discrete visit times $0.1, 0.2, 0.3, ..., 4$ as $s_1, s_2, s_3..., s_m$, and use equation (4.10) to estimate the IIW prevalence probabilities $\hat{P}_{11}(t)$ and $\hat{P}_{12}(t)$.

To estimate the required weights for the IIW methods, we fitted geometric distributions and more general semi-Markov models to both $\Delta t_{ij}|Y_i(t_{ij-1}) = 1$ and $\Delta t_{ij}|Y_i(t_{ij-1}) = 2$. We denote $P(b_{ij} = b|\Delta t_{ij} \geq b, Y_i(t_{i,j-1}) = r)$ for $r = 1, 2$ by $\alpha_r(b)$. The gap time intensities $\alpha_r(b)$ are estimated by $\hat{\alpha}_r(b) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i} I(b_{ij}=b)I(Y_i(t_{i,j-1})=r)}{\sum_{i=1}^{n}\sum_{j=1}^{m_i} I(b_{ij}\geq b)I(Y_i(t_{i,j-1})=r)}$ for $r = 1, 2$. For geometric distributions, $\alpha_r(b) = p_r, (r = 1, 2)$, and it is estimated by $\frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i} I(Y_i(t_{i,j-1})=r)}{\sum_{i=1}^{n}\sum_{j=1}^{m_i} I(Y_i(t_{i,j-1})=r)(b_{ij}/0.1)}$ for $r = 1, 2$. The estimated standard deviations for the semi-Markov and geometric intensity estimators are $SE(\hat{\alpha_r}(b)) = \sqrt{\frac{\hat{\alpha}_r(b)(1-\hat{\alpha}_r(b))}{n_r(b)}}$, $r = 1, 2$, and $SE(\hat{p}_r) = \sqrt{\frac{\hat{p}_r(1-\hat{p}_r)}{n_r}}$, $r = 1, 2$, respectively, where $n_r(b)$ is $\sum_{i=1}^{n}\sum_{j=1}^{m_i} I(b_{ij} \geq b)I(Y_i(t_{i,j-1}) = r)$, and n_r is $\sum_{i=1}^{n}\sum_{j=1}^{m_i} I(Y_i(t_{i,j-1}) = r)$. The estimated geometric probabilities were $\hat{p}_1 = 0.184$ and $\hat{p}_2 = 0.348$ respectively for the sample of 5000 processes generated, and the corresponding standard deviation estimates were 0.003 and 0.006. For semi-Markov models, the estimated gap time intensities $\hat{\alpha}_r(b)$ for $r = 1$ and 2 are shown in Figure 4.1, along with the geometric estimates. In addition, Tables 4.1 and 4.2 show the gap time intensity estimates and their standard errors for the both geometric and more general semi-Markov models. For those people who entered state 3, we kept them in this state at every time point after their entry to this state and took $w_i(t) = 1$ for each subsequent visit time t. Then, we applied the IIW method using the two choices of weights obtained from the two fitted models. In the geometric models, the IIW method can be applied by assigning weights as $w_i(t) = 1/\hat{p}_r$, when $Y_{obs}(t^-) = 1$, 2, and 1 if $Y_{obs}(t^-) = 3$. In the semi-Markov models, the IIW prevalence estimates are estimated using weights $w_i(t)$ as $1/\hat{\alpha}_r(b)$ when $r = 1$, 2, and 1 if $Y_{obs}(t^-) = 3$. After estimating prevalence probabilities $\hat{P}_{11}(t)$ and $\hat{P}_{12}(t)$ using equation (4.10), we used

Table 4.1: Estimated gap time intensities $\hat{\alpha}_1(b)$ and their standard errors for the first simulation study, when gap times are generated from the Geometric distribution. The true gap time intensity estimate is 0.167. The estimated Geometric gap time intensity and its estimated standard error are 0.184 and 0.003, respectively.

	b	0.5	1	1.5	2	2.5	3	3.5
SM	$\hat{\alpha}_1(b)$	0.167	0.177	0.169	0.198	0.150	0.217	0.182
	$SE(\hat{\alpha}_1(b))$	0.004	0.006	0.010	0.018	0.028	0.061	0.116

Table 4.2: Estimated gap time intensities $\hat{\alpha}_2(b)$ and their standard errors for the first simulation study, when gap times are generated from the Geometric distribution. The true gap time intensity estimate is 0.333. The estimated Geometric gap time intensity and its estimated standard error are 0.348 and 0.006 respectively.

	b	0.3	0.6	0.9	1.2	1.5	1.8	2.1
SM	$\hat{\alpha}_2(b)$	0.336	0.331	0.311	0.350	0.250	0.091	0.143
	$SE(\hat{\alpha}_2(b))$	0.008	0.016	0.028	0.053	0.088	0.087	0.132

the loess() function in R with span=0.25 to smooth the estimates. A span is the window which contains the percentage of cases in each local regression. A span that is too small captures more trends and produces a curve with a lot of noise, while a large span will over smooth the regression. Here, we decided to set the span to 0.25, after considering numbers of observation times that fall within this window. For estimating $\hat{P}_{13}(t)$, we computed the raw estimates $\hat{P}_{13}(t) = 1 - (\hat{P}_{11}(t) + \hat{P}_{12}(t))$ obtained from equation (4.10) as the input for standard isotonic regression, iso.reg() in R, to provide monotone $\hat{P}_{13}(t)$ estimates.

The prevalence probability estimates $\tilde{P}_{1r}(t), r = 1, 2, 3$ for the fitted multistate model

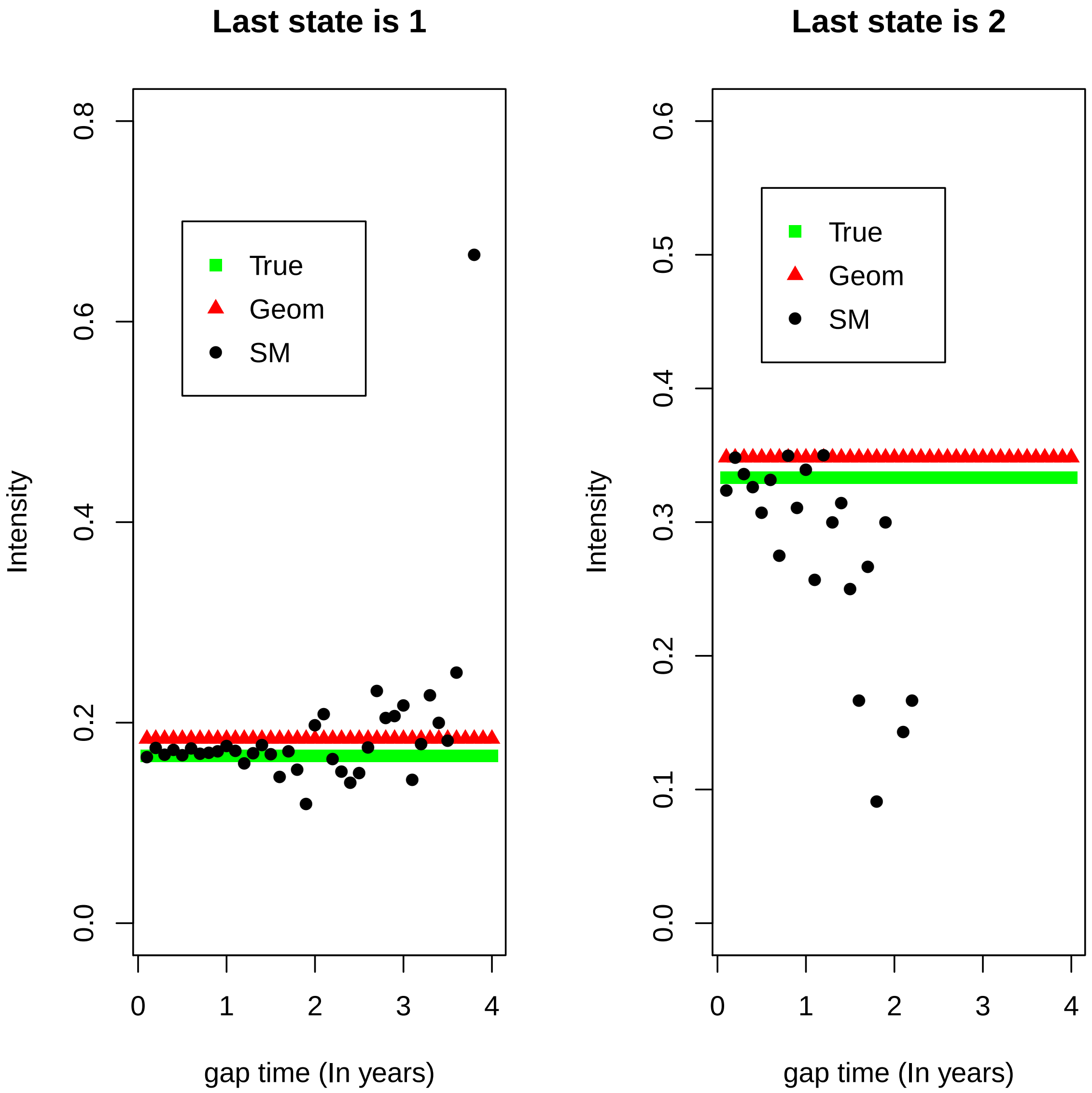

Figure 4.1: The estimates of intensities for gap times between visits along with the true values from the first simulation study, when gap times are generated from the Geometric distribution.

along with the curves obtained from the two IIW methods are presented in Figure 4.2. Empirical prevalences (Naive) as well as the true probability values are also plotted. The Naive prevalences were estimated at times $t = 0, 0.1, 0.2, ..., 4$, using (4.2). For individual i who was seen at two consecutive times t_{j-1} and t_j, where $t_{j-1} \leq t < t_j$, the state at time t was assumed to be $y(t_{j-1})$ if $t \leq \frac{t_{j-1}+t_j}{2}$, otherwise $y(t_j)$. The estimated prevalence probabilities based on geometric weights and semi-Markov weights are referred as "IIW-Geometric" and "IIW-SM" respectively. We can see that the both IIW estimates are in good agreement with the prevalence estimates obtained from the multistate model and the true probability values. In contrast, the biases of the naive prevalence estimates are remarkable.

Figure 4.1 shows plots of the gap times intensity estimates for the fitted geometric distributions and semi-Markov models along with the true values. It can be seen that the estimates obtained from the both geometric and semi-Markov models are in good agreement with the true intensity values when the last state is 1, except for the final very large gap time. In contrast, the semi-Markov estimates differ from the geometric estimates and true values when the last state is 2. This is due to the fact that few gap times are larger than about $b = 1$ in this case, and so estimates of $\alpha_2(b)$ for $b > 1$ are very imprecise. We have also estimated the survival functions and their standard deviations by $\hat{S}_r(b) = \prod_{u<b}\{1 - \hat{\alpha}_r(b)\}$ and $SE(\hat{S}_r(b)) = \sqrt{\hat{S}_r^2(b)\sum_{u<b}(\frac{\hat{\alpha}_r(u)}{n_r(u)(1-\hat{\alpha}_r(u))})}$, respectively. Figure 4.3 contains plots of the survival functions of the gap times for the two IIW methods along with the true survival probabilities when the previous states are 1 and 2. The plots reveal that the estimates obtained from the general semi-Markov models are almost identical to the true survival functions, and, the survival probabilities estimated from the geometric distributions are slightly below the true curves. Tables 4.3 and 4.4 show the estimates of the gap time survival functions and their standard errors for the both geometric distributions and semi-

Table 4.3: Estimated gap time survivals $\hat{S}_1(b)$ and their standard errors for the first simulation study, when gap times are generated from the Geometric distribution.

	b	0.5	1	1.5	2	2.5	3	3.5
True	$S_1(b)$	0.402	0.161	0.065	0.026	0.010	0.004	0.002
Geometric	$\hat{S}_1(b)$	0.362	0.131	0.047	0.017	0.006	0.002	0.00
SM	$\hat{S}_1(b)$	0.394	0.153	0.061	0.026	0.010	0.003	0.001
	$SE(\hat{S}_1(b))$	0.003	0.002	0.002	0.001	0.00	0.00	0.00

Note: $SE = 0.00$ means that it is less than 0.001.

Table 4.4: Estimated gap time survivals $\hat{S}_2(b)$ and their standard errors for the first simulation study, when gap times are generated from the Geometric distribution.

	b	0.3	0.6	0.9	1.2	1.5	1.8	2.1
True	$S_2(b)$	0.296	0.088	0.026	0.008	0.003	0.001	0.00
Geometric	$\hat{S}_2(b)$	0.277	0.077	0.021	0.006	0.002	0.00	0.00
SM	$\hat{S}_2(b)$	0.293	0.091	0.030	0.009	0.003	0.002	0.001
	$SE(\hat{S}_2(b))$	0.005	0.003	0.002	0.001	0.00	0.00	0.00

Note: $SE = 0.00$ means that it is less than 0.001.

Markov models.

Simulation 2

It is also of interest to explore how the IIW prevalence estimates may be influenced by an improper choice of model for the gap times. In the second simulation study, we also look at variance of the estimators, in addition to the effects of misspecification. In this simulation study, we first considered a sample of $n = 5000$ individuals, and we generated $\Delta t_{ij} = b_{ij}|Y_i(t_{ij-1}) = 1$ and $\Delta t_{ij} = b_{ij}|Y_i(t_{ij-1}) = 2$ from $0.1 \times [Poisson(\mu) + 1]$ with $\mu = 5$

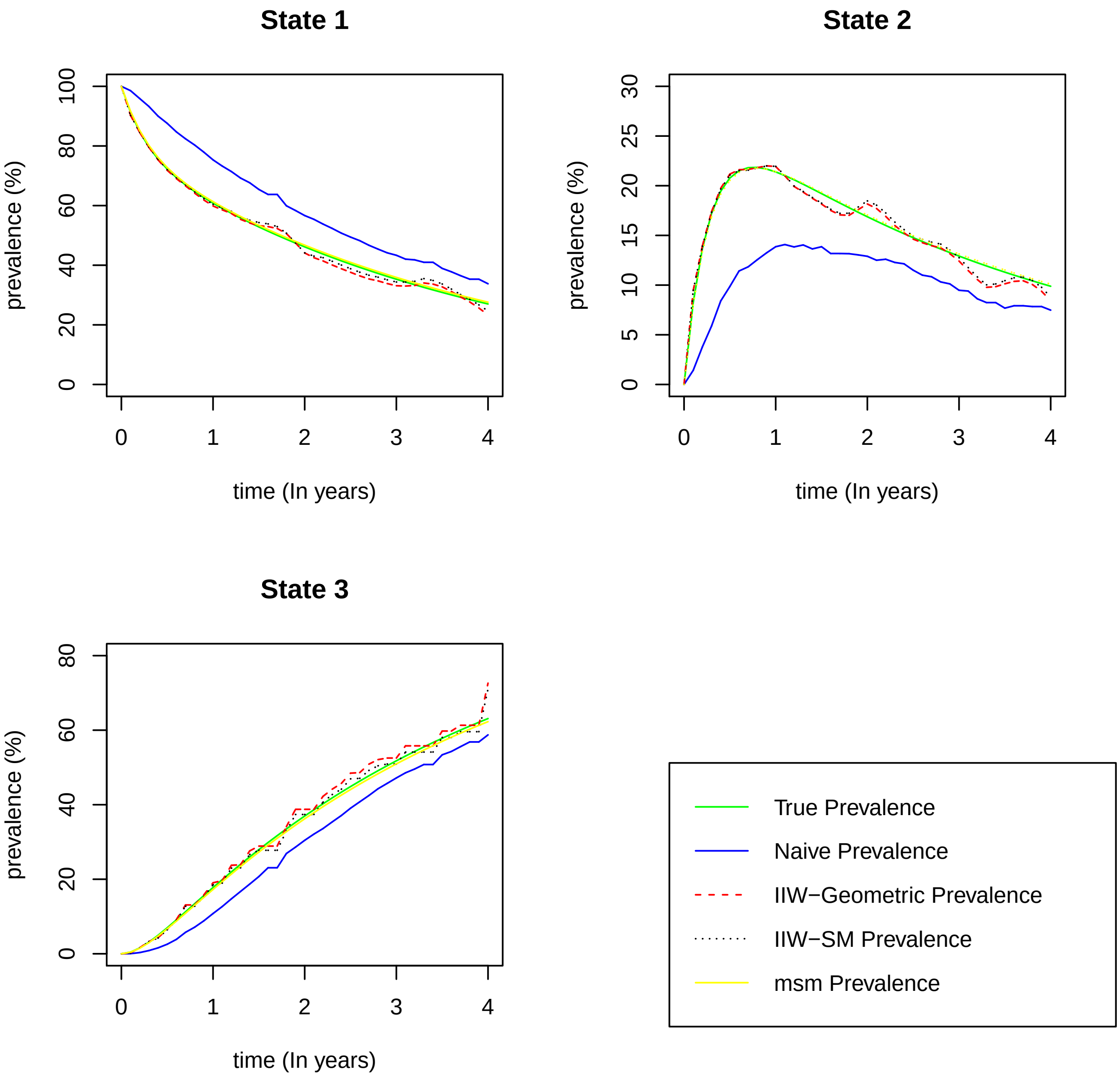

Figure 4.2: Comparison of true, naive, msm, and IIW prevalence estimates for the first simulation study, when gap times are generated from the Geometric distribution.

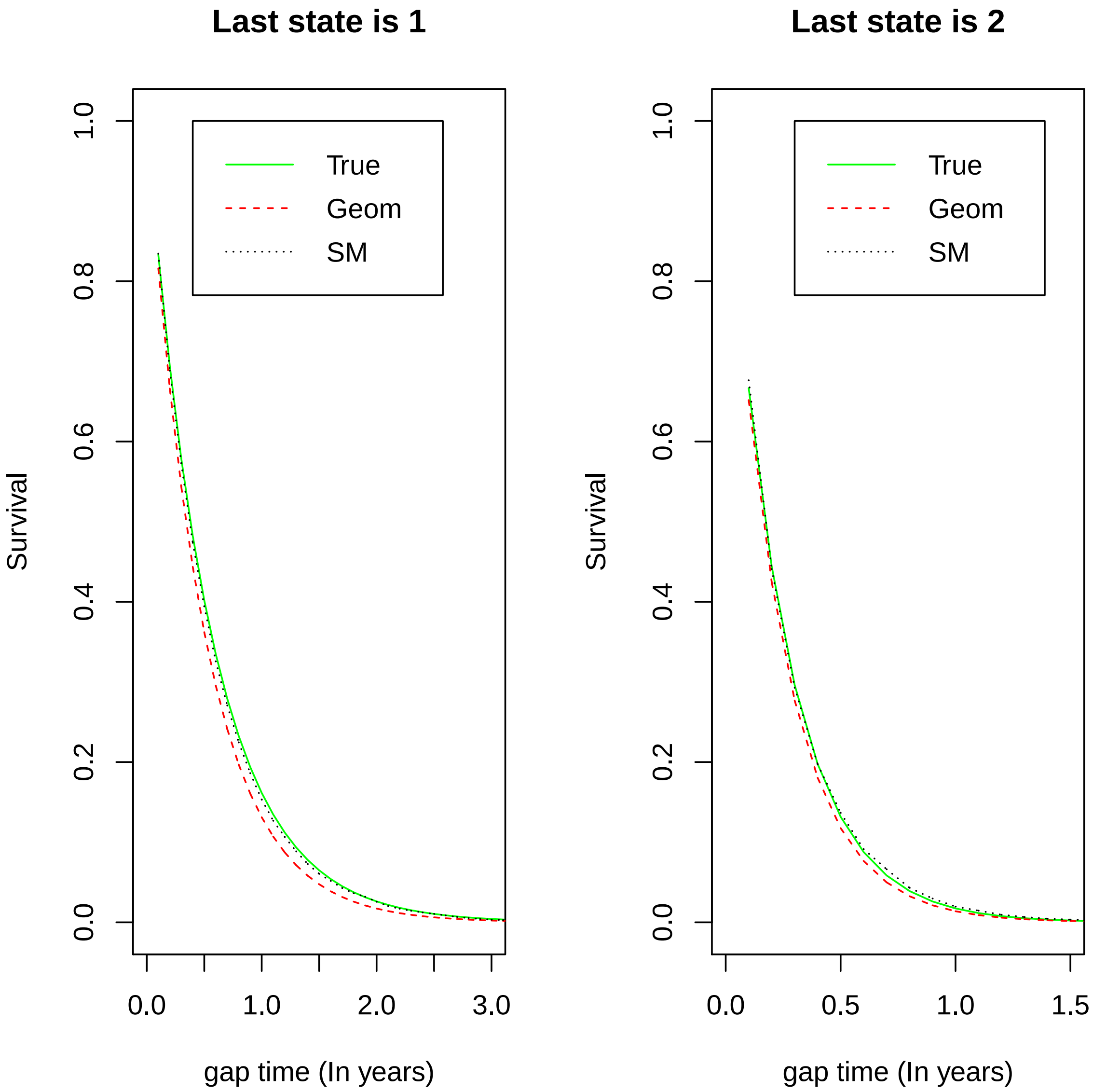

Figure 4.3: The estimates of survival functions for gap times between visits along with the true values from the first simulation study, when gap times are generated from the Geometric distribution.

and 2 respectively. In this case the visit process (gap time) model is semi-Markov, but the gap times do not follow a geometric distribution. Similar to the first simulation study, we fitted geometric distributions and semi-Markov models to gap times. The estimated geometric probabilities for gap times are $\hat{p}_1 = 0.169$ and $\hat{p}_2 = 0.342$, and the estimated gap time intensities for it and for the semi-Markov model are presented in Figure 4.4. It can be seen that the estimated gap times intensities obtained from the general semi-Markov models are quite close to the corresponding true values except, once again, for large values of b having small probability. The geometric intensity estimates are in this case far off the true values. In addition, Tables 4.5 and 4.6 show the gap times intensity estimates and their standard errors for the general semi-Markov models. The estimated standard errors for the geometric intensity estimates are $SE(\hat{p}_1) = 0.003$ and $SE(\hat{p}_2) = 0.005$. We have also shown the estimated survival functions of the gap times along with the true values in Figure 4.5. This shows that the estimates obtained from fitting geometric distributions differ from the semi-Markov estimates, which are close to the true survival functions. The estimates for the gap times survival functions and their standard errors are also presented in Tables 4.7 and 4.8.

The prevalence probability estimates are presented in Figure 4.6. It can be seen that in each plot, the naive prevalence curve is farther from the true curve than the two IIW curves, once again showing bias. The two IIW methods with different choices of weights yield satisfactory estimates with small biases, but surprisingly the IIW-Geometric curve is closer to the true curve, especially as time increases. In addition, we notice that the IIW-SM estimates in this simulation are a bit more variable than in the first simulation. This may be due to 1) imperfect fit of semi-Markov models for the gap times, or 2) on insufficient number of observations at each time t to estimate precisely the IIW-SM prevalence estimates $P_{1r}(t), r = 1, 2, 3$. These factors are especially pronounced for longer

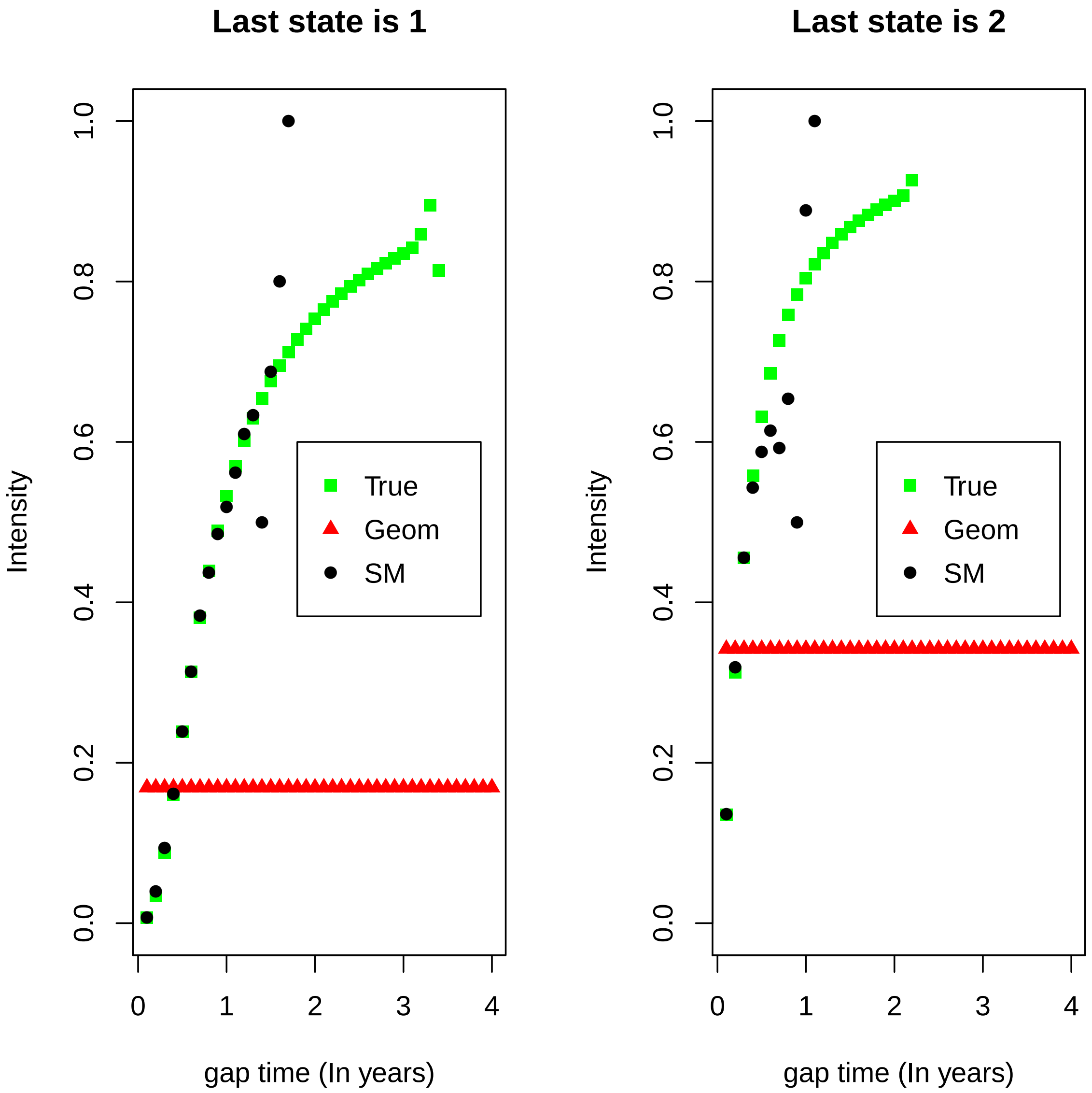

Figure 4.4: The estimates of intensities for gap times between visits along with the true values from the second simulation study, when gap times are generated from the Poisson distribution.

Table 4.5: Estimated gap time intensities $\hat{\alpha}_1(b)$ and their standard errors for the second simulation study, when gap times are generated from the Poisson distribution. The estimated Geometric gap time intensity and its estimated standard error are 0.169 and 0.003, respectively.

	b	0.2	0.4	0.6	0.8	1	1.2	1.4	1.6
True	$\alpha_1(b)$	0.034	0.160	0.314	0.439	0.532	0.602	0.654	0.695
SM	$\hat{\alpha}_1(b)$	0.040	0.161	0.314	0.437	0.519	0.610	0.500	0.800
	$SE(\hat{\alpha}_1(b))$	0.001	0.003	0.004	0.007	0.014	0.031	0.088	0.179

Table 4.6: Estimated gap time intensities $\hat{\alpha}_2(b)$ and their standard errors for the second simulation study, when gap times are generated from the Poisson distribution. The estimated Geometric gap time intensity and its estimated standard error are 0.342 and 0.005, respectively.

	b	0.2	0.4	0.6	0.8	1
True	$\alpha_2(b)$	0.313	0.558	0.685	0.758	0.804
SM	$\hat{\alpha}_2(b)$	0.040	0.161	0.314	0.437	0.519
	$SE(\hat{\alpha}_2(b))$	0.001	0.003	0.004	0.007	0.014

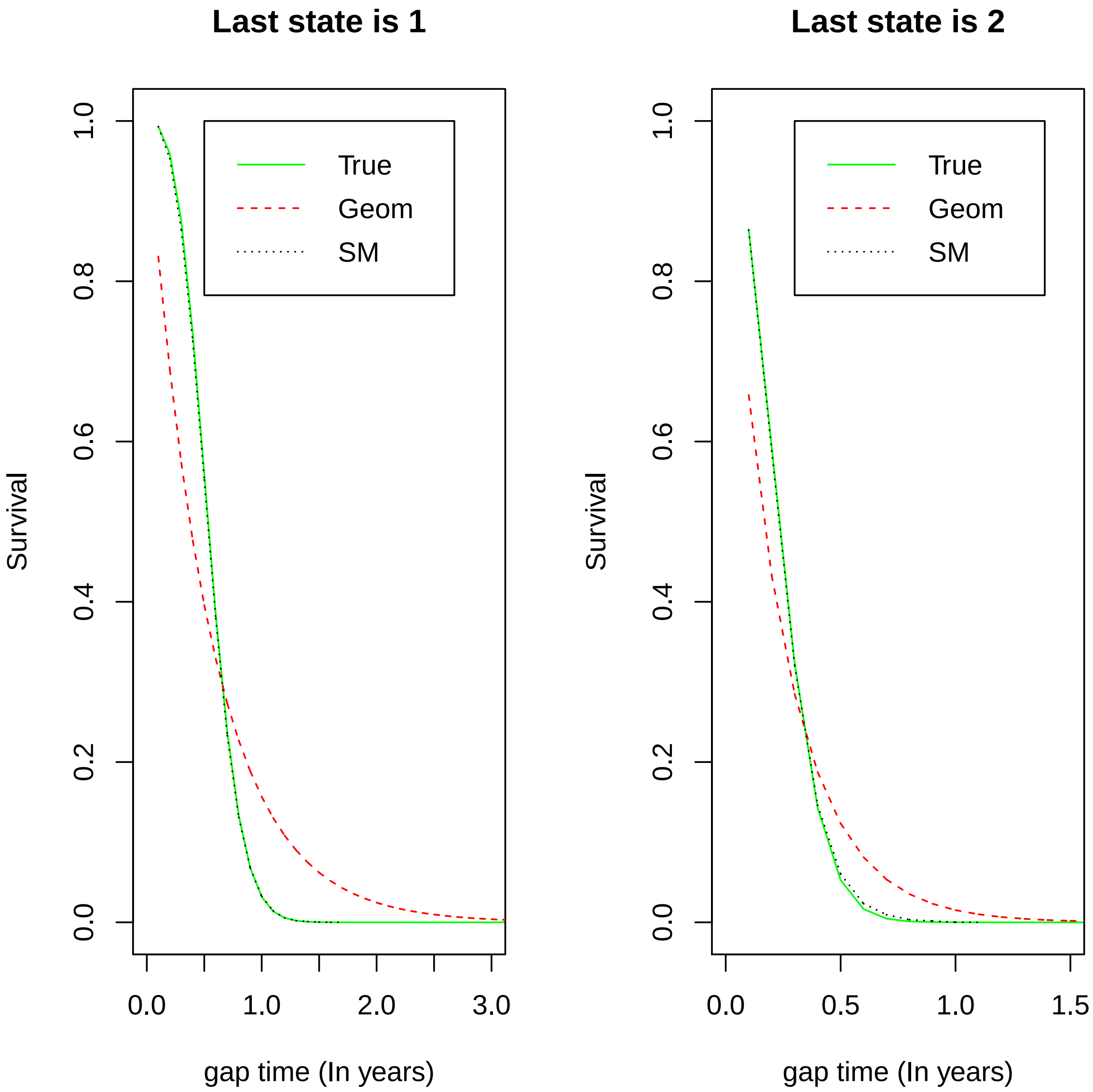

Figure 4.5: The estimates of survival functions for gap times between visits along with the true values from the second simulation study, when gap times are generated from the Poisson distribution.

Table 4.7: Estimated gap time survivals $\hat{S}_1(b)$ and their standard errors for the second simulation study, when gap times are generated from the Poisson distribution.

	b	0.2	0.4	0.6	0.8	1	1.2	1.4	1.6
True	$S_1(b)$	0.959	0.735	0.384	0.133	0.032	0.005	0.001	0.00
Geometric	$\hat{S}_1(b)$	0.690	0.476	0.329	0.227	0.157	0.108	0.075	0.051
SM	$\hat{S}_1(b)$	0.953	0.724	0.378	0.131	0.032	0.005	0.001	0.00
	$SE(\hat{S}_1(b))$	0.001	0.003	0.003	0.002	0.001	0.00	0.00	0.00

Table 4.8: Estimated gap time survivals $\hat{S}_2(b)$ and their standard errors for the second simulation study, when gap times are generated from the Poisson distribution.

	b	0.2	0.4	0.6	0.8	1
True	$S_2(b)$	0.594	0.143	0.016	0.001	0.00
Geometric	$\hat{S}_2(b)$	0.433	0.187	0.081	0.035	0.015
SM	$\hat{S}_2(b)$	0.588	0.146	0.023	0.003	0.00
	$SE(\hat{S}_2(b))$	0.006	0.004	0.002	0.00	0.00

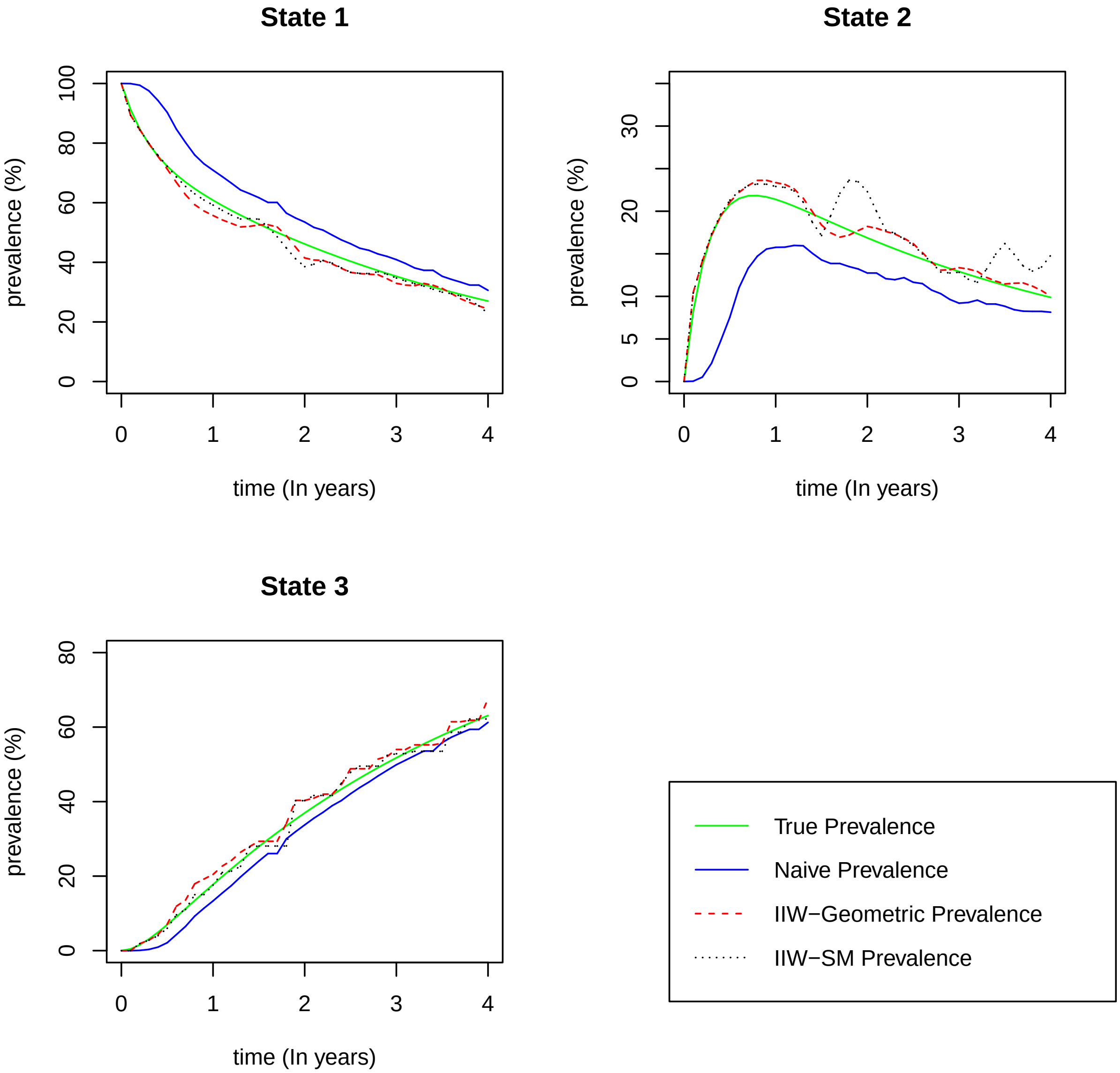

Figure 4.6: Comparison of true, naive, and IIW prevalence estimates for the second simulation study, when gap times are generated from the Poisson distribution.

values of t. In order to investigate these reasons, we computed weights based on the true Poisson distributions, and then calculated the IIW prevalences with these weights. Figure 4.7 shows that these IIW prevalences are very close to the IIW-SM prevalences, which points out that the choice of semi-Markov models for the gap times was reasonable.

We also did the simulation with a much bigger sample size of 50,000 individuals to see whether we get similar results, since imprecision of intensity estimates will then be much less of a problem. In addition, by splitting the 50,000 individuals into (a) 10 samples of size 5000, and (b) 50 samples of size 1000, we can study the variation in each estimator, and estimate their variance for these specific sample sizes. In this simulation, the observation times were generated from the Poisson distributions with the same parameter values as the previous simulation study. The prevalence probability estimates are shown in Figure 4.8. We can see that the IIW-SM curves are now in better agreement with the true curves compared to the IIW-Geometric estimates, and the amount of variability in the IIW-SM estimates has been greatly reduced. However, the estimates obtained from IIW-Geometric method are still performing well. In order to compute the standard deviations of estimators, we partitioned the realizations for the 50,000 individuals into (a) 10 samples of size 5000 and (b) 50 samples of size 1000. These partions were made to relate the results of replications to the simulation of 5000 individuals realizations and the real data sample size. The sample means and 95% confidence intervals of the two IIW estimators for the both cases (a) and (b) are presented in Figures 4.9 and 4.10, respectively. In addition, Tables 4.9, 4.10, and 4.11 show the samples means and standard deviations of the IIW prevalence estimators as well as the naive and msm estimators.

The variability of the IIW-SM estimates in Figure 4.6 seems to be due to having insufficient observations for some time points t. It should be noted that the IIW prevalences at time t are solely estimated on the basis of the observations at time t. Hence, although a sample of 5000 individuals has been simulated, the number of observed individuals at each time point t is much less than the total number of individuals. In addition, as t increases, fewer individuals remain in states 1 or 2. Therefore, some observations with very small probabilities of occurrence can lead to extremely large weight values, which may result in high variability of the IIW estimates. Thus, we also explored the variability of the estimated IIW-SM prevalence probabilities in more detail. We considered the distribution of the estimated weights obtained from the fitted semi-Markov models, and noticed that there were 149 observations with weights 137.86. These weights were related to gap times equal to 0.1 when the last state is 1. Hence, to obtain less variable prevalence estimates in the smaller sample size, we adjusted these extreme weights by truncating them. The weights were truncated at 25.1, which is the corresponding weight for the gap time of 0.2 when the last state is 1. The prevalence estimates obtained by this adjustment are shown in Figure 4.11. We see that the variability in IIW-SM prevalence estimates has been reduced substantially by truncating large weights. As a second investigation, we revised the second simulation by generating gap times from Poisson distributions that are truncated at $b = 1$; that is, with probability function $\frac{\mu^{b-1}e^{-\mu}/(b-1)!}{(1-e^{-\mu})}, b = 2, 3, ...$ This model does not have extremely small intensities for any value of b. The smallest $\hat{\alpha}_r(b)$'s are 0.034 and 0.137 for $r = 1$ and 2, respectively. Figure 4.12 shows prevalence probability curves obtained from this simulation. The plots reveal that the estimated IIW-SM prevalences have even less variability than the similar estimates in Figure 4.11.

From the simulation results presented in this subsection, we can conclude the following. First, in the presence of the process-dependent observation times, the IIW method yields prevalence estimates that are close to the true values, while the naive method that ignores this dependency yields estimates that are substantially far from the true values. Second, simulations demonstrate that when there are sufficient number of observations, the IIW

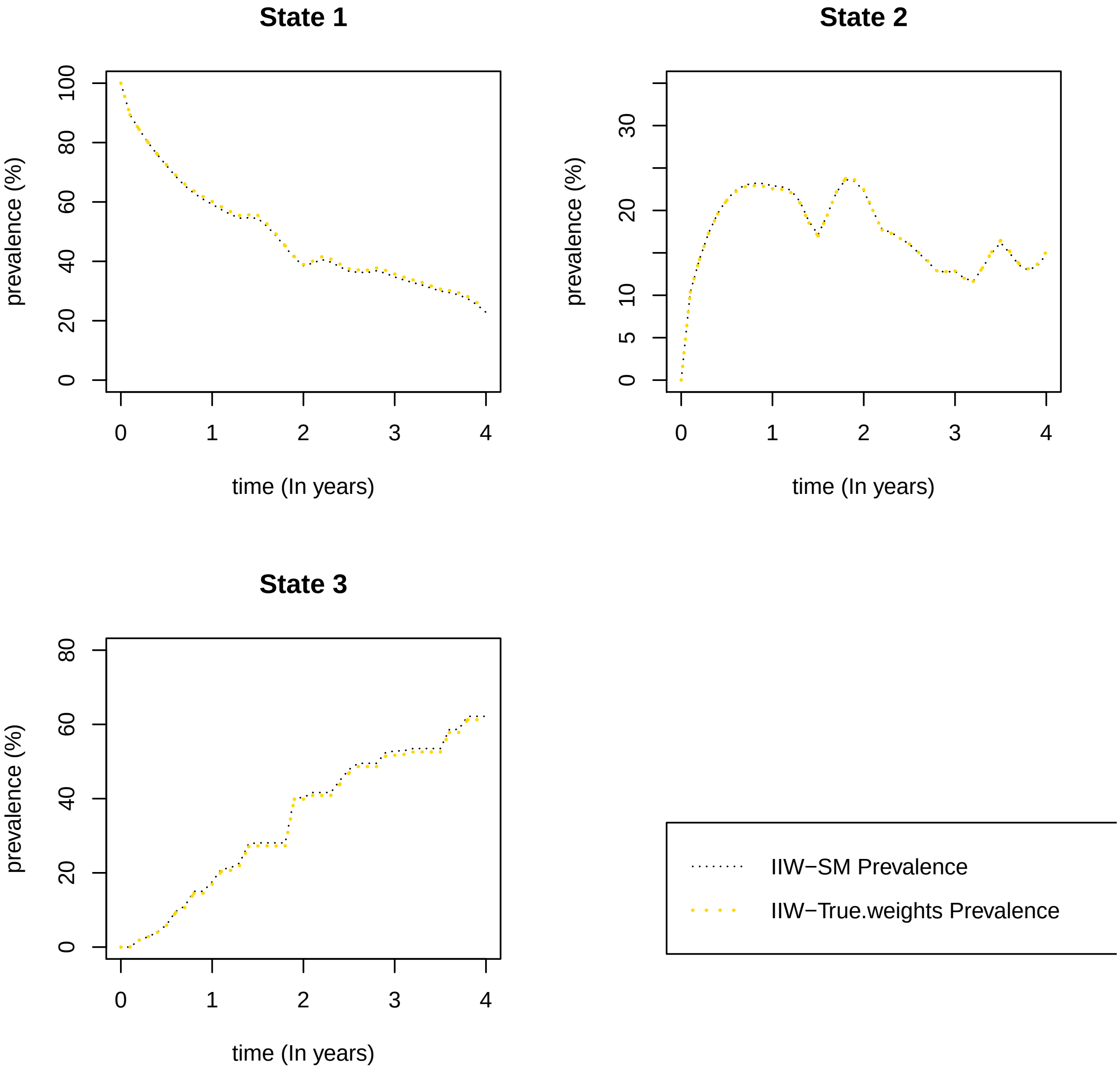

Figure 4.7: Comparison of true and IIW prevalence estimates for the second simulation study, when gap times are generated from the Poisson distribution.

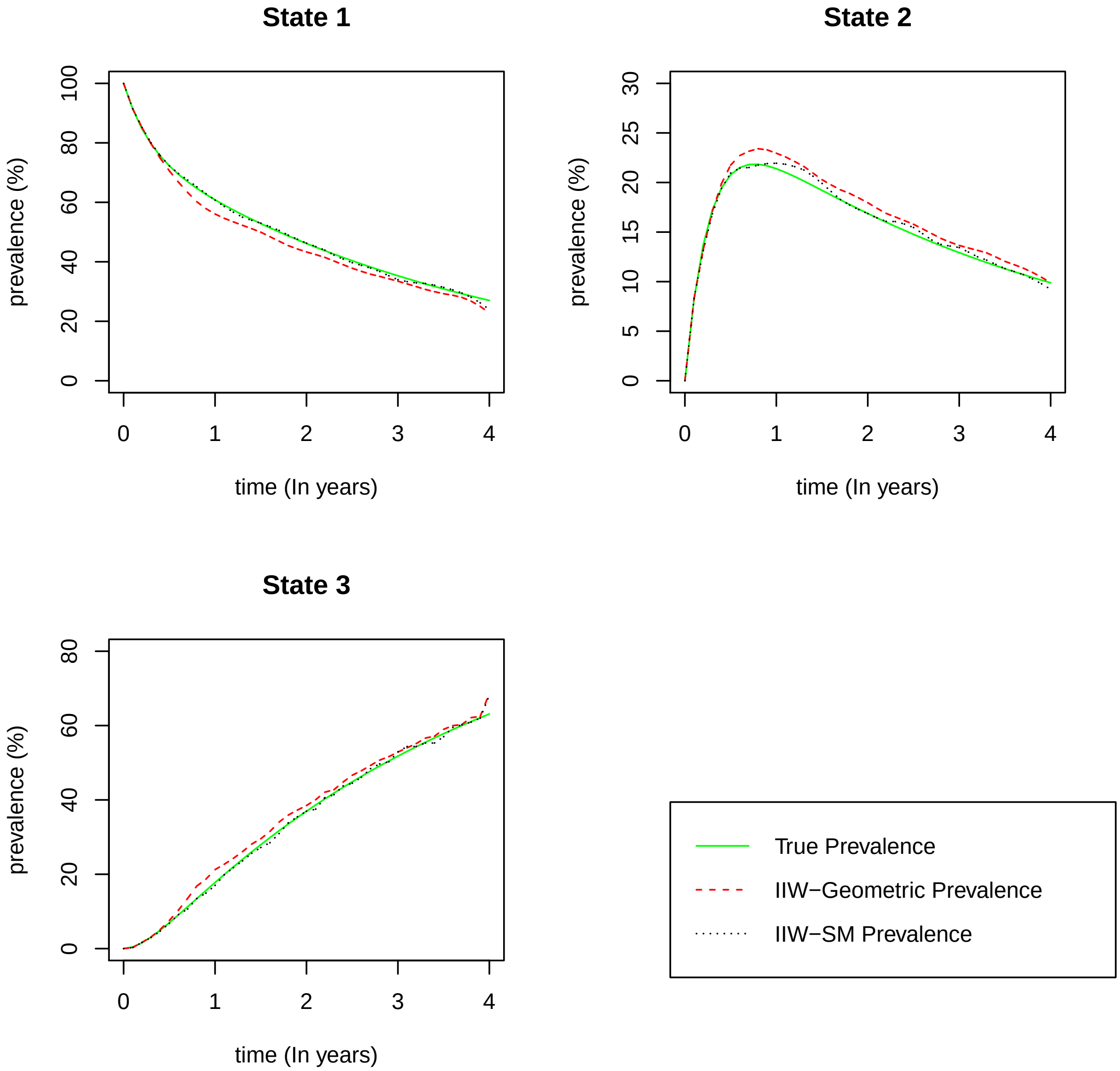

Figure 4.8: Comparison of true and IIW prevalences for the second simulation study, when we increased the sample size to $n = 50,000$.

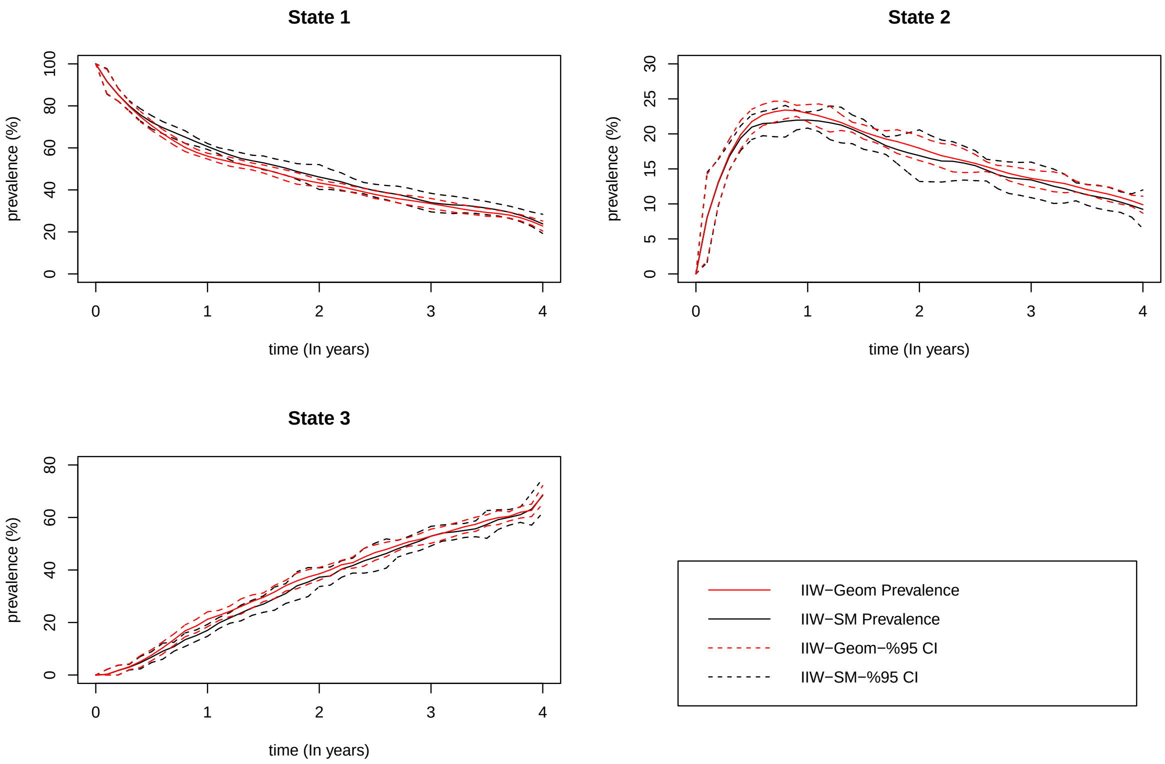

Figure 4.9: The two IIW prevalence means with their 95% confidence intervals, when we partitioned the sample size of $n = 50,000$ in the second simulation study into 10 samples of size 5000.

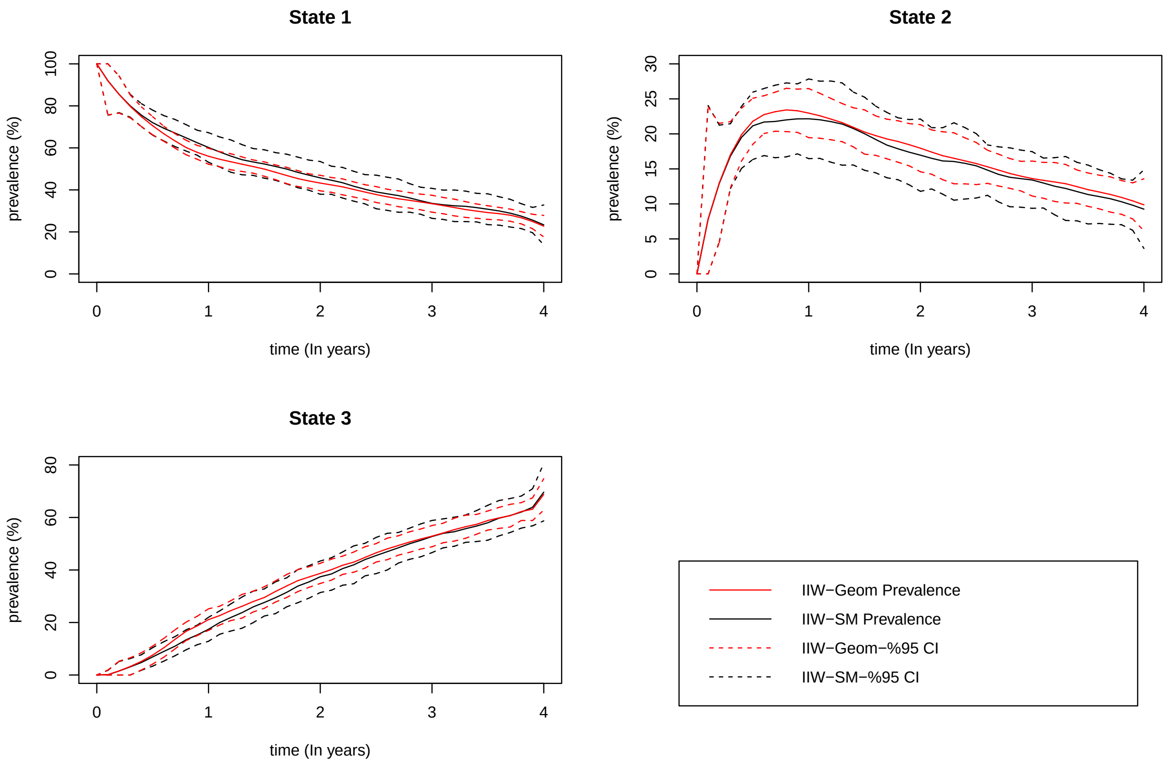

Figure 4.10: The two IIW prevalence means with their 95% confidence intervals, when we partitioned the sample size of $n = 50,000$ in the second simulation study into 50 samples of size 1000.

Table 4.9: The results of the samples means and standard deviations of the IIW, naive, and msm estimates for $P_{11}(t)$ when we partitioned n=50,000 sample into (a) 10 samples of size 5000 and (b) 50 samples of size 1000.

		t	1	2	3	4
		True	0.608	0.462	0.353	0.270
(a) 50 samples of size 1000	Geometric	Mean	0.560	0.432	0.334	0.227
		SE	0.019	0.018	0.020	0.025
	SM	Mean	0.602	0.457	0.336	0.233
		SE	0.035	0.039	0.035	0.047
	msm	Mean	0.607	0.461	0.353	0.270
		SE	0.008	0.011	0.012	0.012
	Naive	Mean	0.710	0.536	0.410	0.312
		SE	0.014	0.013	0.016	0.015
(b) 10 samples of size 5000	Geometric	Mean	0.561	0.433	0.335	0.228
		SE	0.006	0.007	0.010	0.010
	SM	Mean	0.607	0.461	0.339	0.238
		SE	0.006	0.026	0.019	0.020
	msm	Mean	0.607	0.461	0.353	0.270
		SE	0.004	0.006	0.006	0.006
	Naive	Mean	0.710	0.536	0.410	0.312
		SE	0.005	0.007	0.009	0.008

Table 4.10: The results of the samples means and standard deviations of the IIW, naive, and msm estimates for $P_{12}(t)$ when we partitioned n=50,000 sample into (a) 10 samples of size 5000 and (b) 50 samples of size 1000.

		t	1	2	3	4
		True	0.214	0.169	0.129	0.099
(a) 50 samples of size 1000	Geometric	Mean	0.230	0.179	0.136	0.099
		SE	0.017	0.017	0.012	0.012
	SM	Mean	0.221	0.169	0.134	0.092
		SE	0.028	0.026	0.020	0.028
	msm	Mean	0.216	0.170	0.131	0.100
		SE	0.005	0.004	0.004	0.004
	Naive	Mean	0.158	0.130	0.098	0.079
		SE	0.012	0.008	0.007	0.008
(b) 10 samples of size 5000	Geometric	Mean	0.229	0.179	0.136	0.099
		SE	0.005	0.008	0.005	0.005
	SM	Mean	0.220	0.169	0.134	0.092
		SE	0.005	0.016	0.011	0.012
	msm	Mean	0.216	0.170	0.131	0.100
		SE	0.002	0.002	0.002	0.002
	Naive	Mean	0.158	0.130	0.098	0.079
		SE	0.002	0.005	0.003	0.003

Table 4.11: The results of the samples means and standard deviations of the IIW, naive, and msm estimates for $P_{13}(t)$ when we partitioned n=50,000 sample into (a) 10 samples of size 5000 and (b) 50 samples of size 1000.

		t	1	2	3	4
		True	0.178	0.370	0.518	0.631
(a) 50 samples of size 1000	Geometric	Mean	0.211	0.387	0.529	0.688
		SE	0.020	0.019	0.020	0.030
	SM	Mean	0.174	0.374	0.527	0.696
		SE	0.023	0.030	0.030	0.054
	msm	Mean	0.177	0.369	0.516	0.630
		SE	0.006	0.012	0.014	0.015
	Naive	Mean	0.132	0.334	0.492	0.608
		SE	0.010	0.014	0.015	0.016
(b) 10 samples of size 5000	Geometric	Mean	0.212	0.385	0.529	0.687
		SE	0.012	0.011	0.012	0.015
	SM	Mean	0.212	0.385	0.529	0.687
		SE	0.010	0.016	0.016	0.029
	msm	Mean	0.177	0.368	0.516	0.630
		SE	0.003	0.006	0.007	0.008
	Naive	Mean	0.132	0.334	0.492	0.608
		SE	0.005	0.007	0.007	0.009

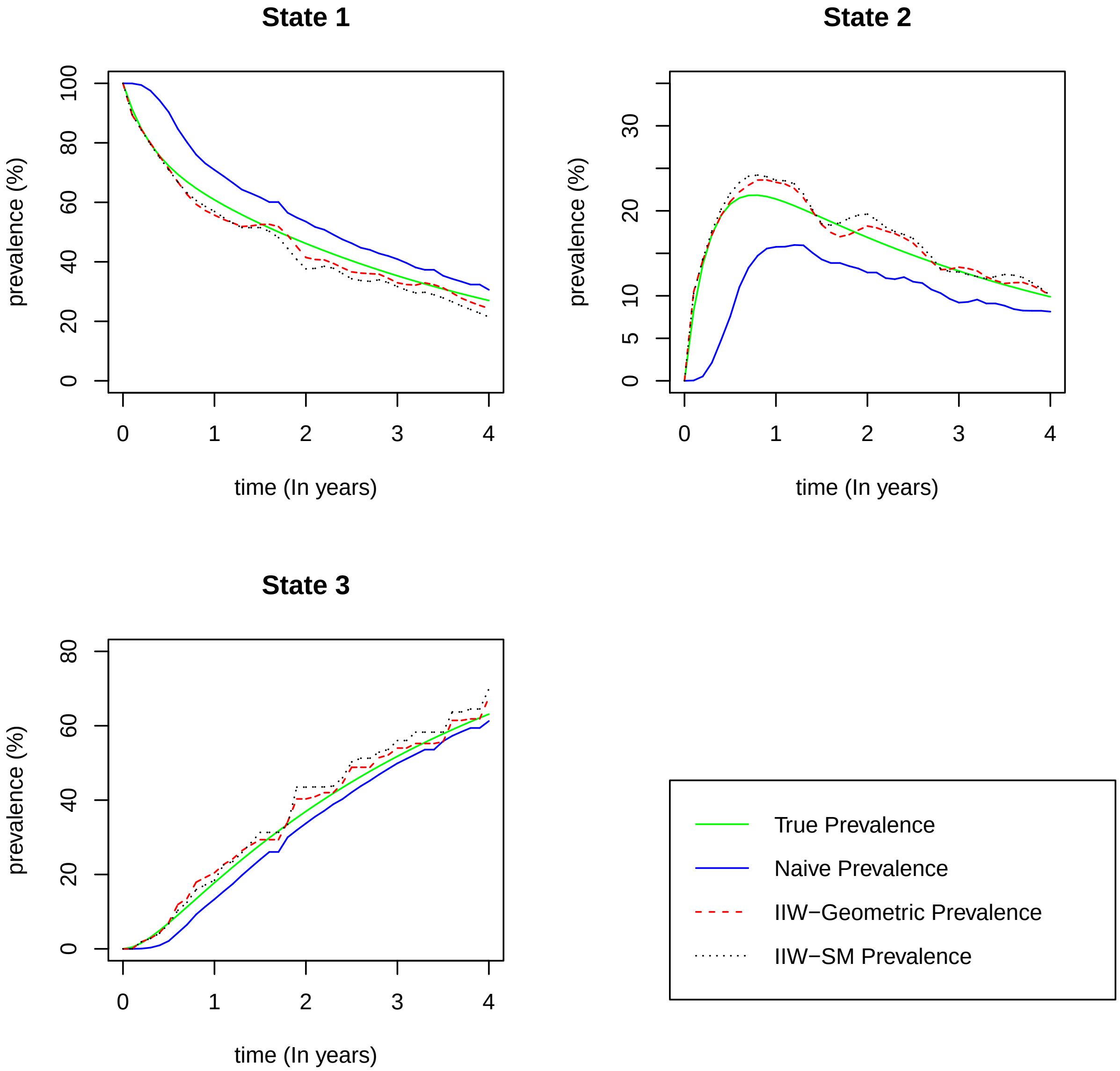

Figure 4.11: Comparison of true, naive, and IIW prevalence estimates for the second simulation study, when we truncated the semi-Markov weights at 25.1.

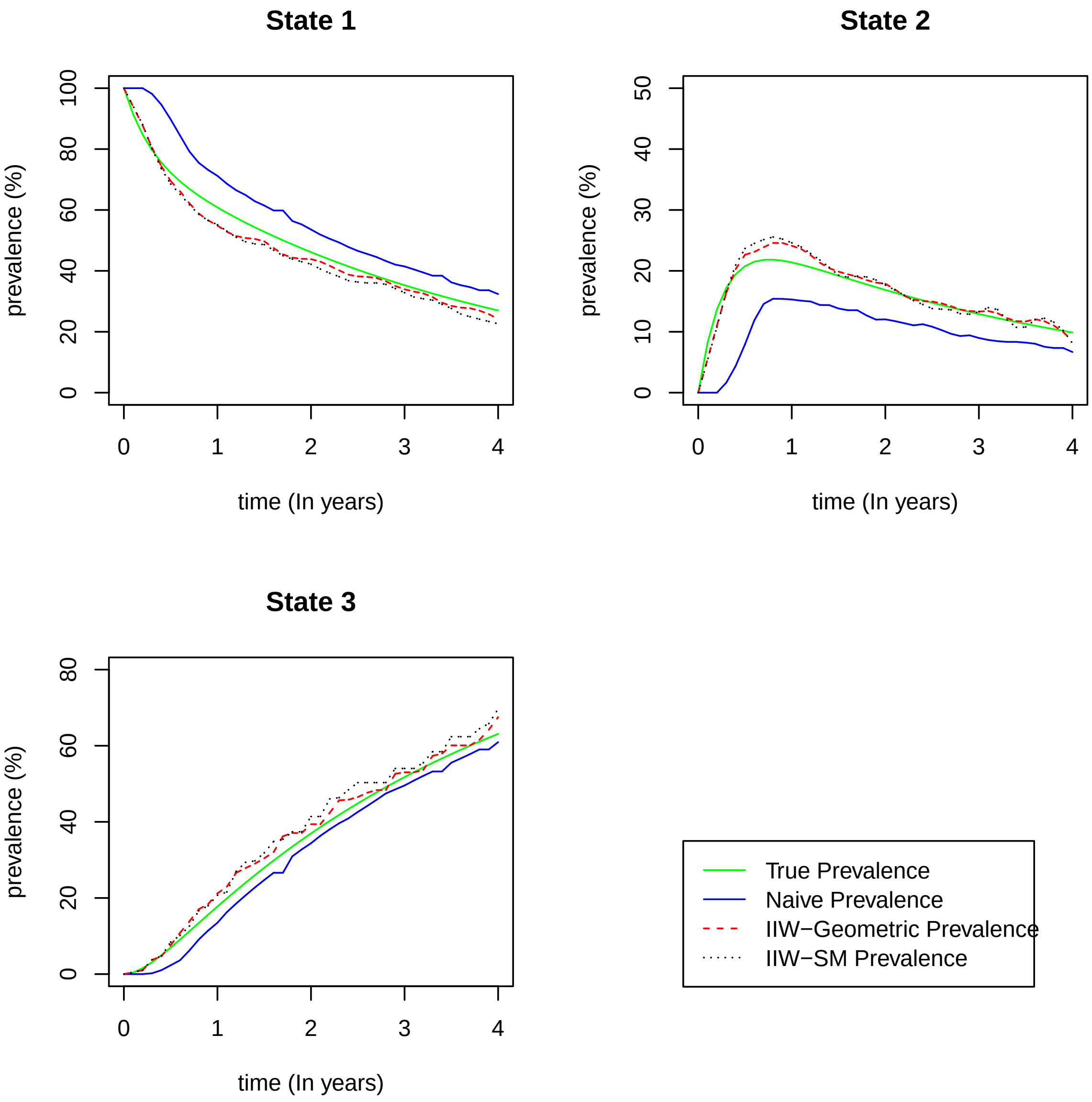

Figure 4.12: Comparison of true, naive, and IIW prevalence estimates for the second simulation study, when observations are generated from the truncated Poisson distribution.

method performs well, but when the number is not sufficiently large, high variability in the weights can produce high variability in the prevalence estimates. In situations where there is not enough observations, we should either use models that do not produce extreme weights, or consider truncation of weights to adjust for influential observations obtained from the large weights.

In practice, it is usually the case that the number of processes under observation is smaller than the 5000 we chose for the simulation studies. For example, in the CANOC studies a sample of about 1000 individuals is more common. In this situation, proper IIW estimation of prevalences requires taking the existence of extreme weights into consideration. We recommend the use of truncation, as well as, the use of intervals in (4.11) that each contain a substantial number of gap times b_{ij}. We also recommend the use of parametric models to estimate the gap time intensities for each interval. They can be chosen for flexibility (e.g. piecewise-constant intensities) and to avoid extreme weight values, by effectively grouping gap times that are less frequently observed. In Section 4.4.2, we consider the CANOC data, and we will find that there are no extreme estimated weights that require adjustment.

4.4 Application to CANOC data

In this section, we apply the methodology in this chapter to the CANOC data which we described in Section 1.6. First, we fit Cox survival model to time to the first viral rebound, and then compare the covariate effects obtained from this approach with those estimated by multistate models. Second, we show that the observation times t_{ij} are dependent on the previous process history, in CANOC data. Then, we consider prevalence estimation and apply the IIW method to adjust for bias due to the dependent observation times.

4.4.1 Survival models for the time to first viral rebound

The distribution of time T_i to first viral rebound is estimated from the multistate model 3.1(b) by the probability of entry to state 3. Another way to analyze times to viral rebounds is through ordinary failure time methodology that we discussed in Section 4.2.2. Here, we consider the use of Cox models as in Grennan et al. (2012). We fitted the Cox model with the intensity for T_i of the form $\lambda(t|z_i) = \lambda_0(t)exp(\beta^T z_i), i = 1, ..., n$, to the times of viral rebounds, and considered covariates age, IDU, and cART as the baseline covariate vector z_i. We assumed "T_i" to be t_{ij} when t_{ij} is the first time a person is observed in state 3. Tables 4.12 and 4.13 show the estimated covariate effects and their standard errors. From the Tables 4.12 and 4.13, we can conclude that injection drug users, and individuals older than 45 years have a higher, and lower probability for time to viral rebound, respectively. Similarly, from the Tables 3.5 and 3.6 in Section 3.4 for multistate models, we see that injection drug users tend to have more transitions from state 1 to state 2, and from state 2 to state 3, whereas they show less tendency to move from state 2 to state 1. This leads to having larger probability of progression to viral rebound among injection drug users. Persons over 45 tend to have less transitions from state 2 to state 3. In addition, Tables 4.12 and 4.13 show that individuals with unknown IDU status, and those taking PI treatment are similar to non-injection drug users, and persons with NNRTI treatment, respectively. It should be noted, however, that a closer picture of time to viral rebound estimates is obtained if we consider prevalence estimates $P_{13}(t)$ from the multistate models, rather than just transition intensities. We do this next.

The comparison of multistate and survival models in the presence of internal covariates will be discussed later in Chapter 5.

4.4.2 Prevalence estimation and viral rebounds

In earlier sections, we noted that sometimes the observation times may depend on previous process history. In the CANOC data, we now examine the dependency of gap times between visits on the previous viral load state by fitting Cox models to the gap times, as discussed in Section 4.3.2. We fit Cox models to the inter-visit gap times, $b_{ij} = t_{ij} - t_{i,j-1}$, and examined the effects of different covariates. Our baseline covariates include age, IDU, and cART. We also define the time-varying covariate last.state, which is the corresponding viral load state of the individual at the last visit. This covariate has two levels of 1 and 2; the reference level is taken to be 1.

Tables 4.14 and 4.15 show the results of the fitted Cox models for both the FARV1 and FARV2 groups. From these tables, we can see that injection drug users have reduced hazard function for the next visit, so the gap times between visits tend to be longer. However, $age > 45$, PI treatment, and having last.state=2 increases the hazard, and thus the gap times are shorter. In particular, the covariate last.state is highly significant in connection with gap times. The estimated hazard of the gap times among individuals with the last.state equal to 2 is 1.488 (1.443) times that of those with the last.state equal to 1 in FARV1 (FARV2). This shows that the observation times are state dependent. Since this dependency can cause bias in nonparametric prevalence estimates as discussed in Section 4.3, we will use the IIW approach to obtain nonparametric estimates. These will be compared with estimates based on the Markov multistate model.

We apply the IIW method by assigning weights based on Cox models for the gap times. The Cox models were fitted using the "coxph" function in R, and the estimated covariate

Table 4.12: The results of the fitted Cox model for time to viral rebound in the FARV1 group.

Covariates	$\hat{\beta}$	$se(\hat{\beta})$	Relative Risk (RR)	95% CI for RR	P-value
Age (> 45)	-0.350	0.121	0.705	(0.556, 0.894)	0.004*
IDU (=Yes)	0.704	0.134	2.021	(1.554, 2.629)	1.55e-07*
IDU (=NA)	-0.102	0.151	0.903	(0.672, 1.213)	0.499
cART (=PI)	0.155	0.112	1.167	(0.938, 1.453)	0.166

$*|\frac{\hat{\beta}}{se(\hat{\beta})}| > 1.96$

Table 4.13: The results of the fitted Cox model for time to viral rebound in the FARV2 group.

Covariates	$\hat{\beta}$	$se(\hat{\beta})$	Relative Risk (RR)	95% CI for RR	P-value
Age (> 45)	-0.320	0.181	0.726	(0.509, 1.036)	0.077
IDU (=Yes)	0.891	0.216	2.438	(1.595, 3.725)	3.82e-05*
IDU (=NA)	-0.274	0.245	0.760	(0.470, 1.230)	0.265
cART (=PI)	-0.296	0.182	0.744	(0.521, 1.062)	0.103

$*|\frac{\hat{\beta}}{se(\hat{\beta})}| > 1.96$

Table 4.14: The results of the fitted Cox model for gap times between visits in the FARV1 group.

Covariates	$\hat{\beta}$	$se(\hat{\beta})$	Relative Risk (RR)	95% CI for RR	P-value
Age (> 45)	0.228	0.016	1.256	(1.216, 1.298)	$< 2e-16*$
IDU (=Yes)	-0.184	0.021	0.832	(0.798, 0.867)	$< 2e-16*$
IDU (=NA)	-0.221	0.019	0.802	(0.772, 0.832)	$< 2e-16*$
cART (=PI)	0.122	0.042	1.130	(1.094, 1.167)	1.02e-13*
Last.state (=2)	0.398	0.016	1.488	(1.371, 1.615)	$< 2e-16*$

$*|\frac{\hat{\beta}}{se(\hat{\beta})}| > 1.96$

Table 4.15: The results of the fitted Cox model for gap times between visits in the FARV2 group.

Covariates	$\hat{\beta}$	$se(\hat{\beta})$	Relative Risk (RR)	95% CI for RR	P-value
Age (> 45)	0.081	0.021	1.084	(1.041, 1.129)	9.71e-05*
IDU (=Yes)	-0.265	0.027	0.767	(0.727, 0.809)	$< 2e-16*$
IDU (=NA)	-0.333	0.024	0.717	(0.684, 0.751)	$< 2e-16*$
cART (=PI)	0.117	0.022	1.124	(1.076, 1.175)	$1.42e-07*$
Last.state (=2)	0.367	0.042	1.443	(1.330, 1.566)	$< 2e-16*$

$*|\frac{\hat{\beta}}{se(\hat{\beta})}| > 1.96$

effects, $\hat{\beta}$, were used to estimate the baseline intensity function in (4.7). We then estimated weights for the observation times t_{ij} by taking the inverse of the estimated Cox model intensities for corresponding gap times b_{ij} as in (4.8). The estimated weights in the FARV1 and FARV2 groups were within ranges $0.45 - 5.65$ and $0.27 - 4.09$ respectively, and 99.8% of the observation times t_{ij} had weight values inside intervals $0.45 - 1$ and $0.27 - 1.4$ correspondingly. Hence, in this case, we did not face the extreme weight issue that we discussed at the end of Section 4.3.4, and no truncation of weights was used.

The IIW prevalence estimates were calculated via (4.11) and (4.13). We took a set of grid values $a_l = 0, 0.25, 0.5, 0.75, ...$, and made associated intervals I_l as in Section 4.3.3. The IIW prevalence estimates were computed at midpoints $t_{m_l} = 0.125, 0.375, 0.625, ...$ Figures 4.13 to 4.18 show prevalence probability estimates $\hat{P}_{1r}(t), r = 1, 2, 3$ including estimates obtained from fitting multistate models, empirical, and IIW methods. The empirical prevalence estimates (denoted "Naive Prevalence") were calculated using formula (4.2), and the model based estimates (denoted "msm Prevalence") are based on the fitted multistate models associated with Tables 3.5 and 3.6. The nonparametric prevalence estimates were computed at points $t = 0, 0.125, 0.375, 0.625, ...$ Figures 4.13 and 4.16 show aggregated estimates obtained from averaging the estimated probabilities for all individuals (averaging over covariate values) at each time point t. Figures 4.14, 4.15, 4.17, and 4.18 represent the combined probability estimates for injection drug users or non users. In Figures 4.13 and 4.16, we have also shown the empirical prevalence estimates given by the msm package (denoted "Empirical-msm Prevalence"). In this case, suppose that individual i was seen at two consecutive times t_{j-1} and t_j, where $t_{j-1} \leq t < t_j$. Then, the msm software assumes that the state of this individual at time t is $y(t_{j-1})$ if $t \leq \frac{t_{j-1}+t_j}{2}$, otherwise $y(t_j)$. These empirical prevalences given by the msm software are very similar to the naive estimates of (4.2).

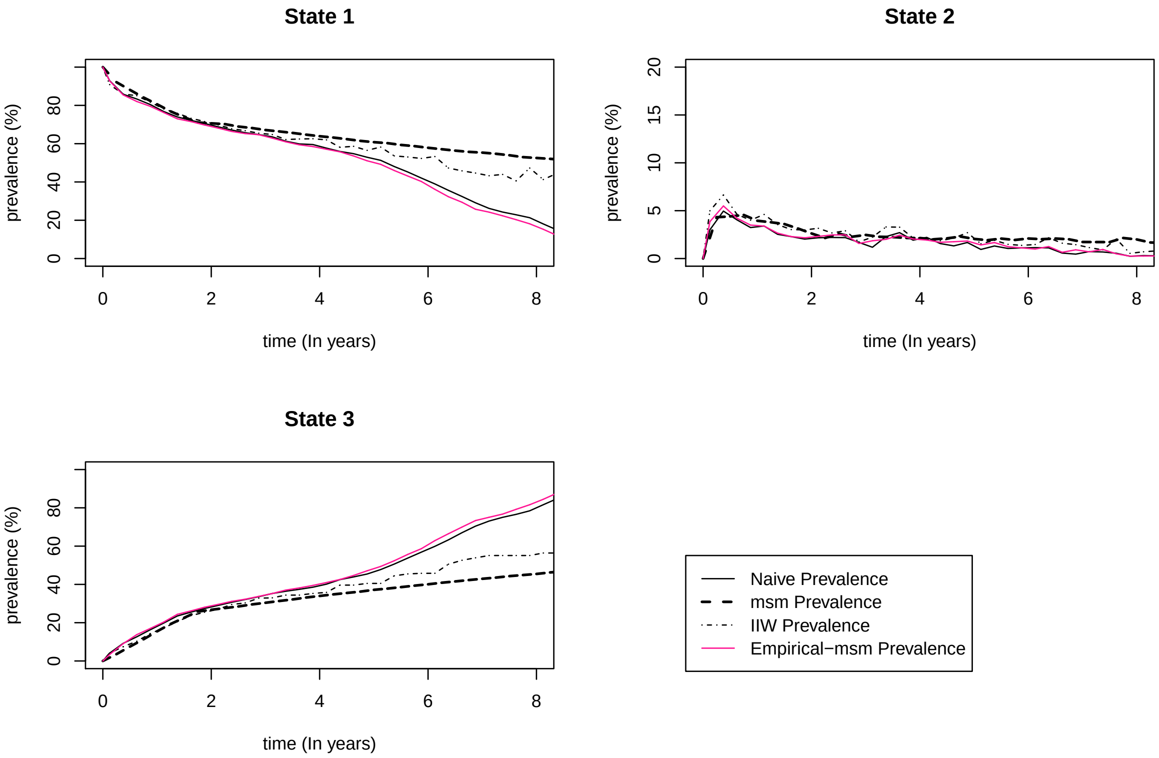

Figure 4.13: Comparison of naive, msm, empirical-msm, and IIW prevalences for all individuals in the FARV1 group.

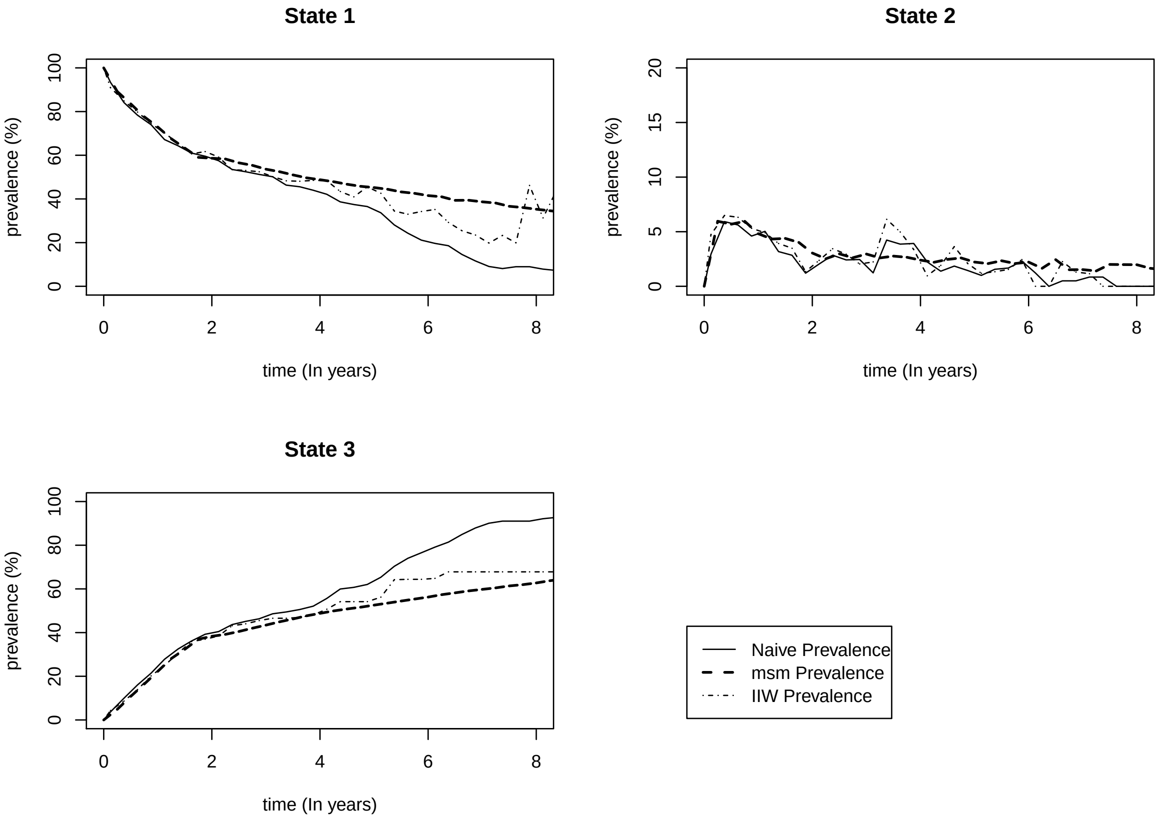

Figure 4.14: Comparison of naive, msm, and IIW prevalences for injection drug users in the FARV1 group.

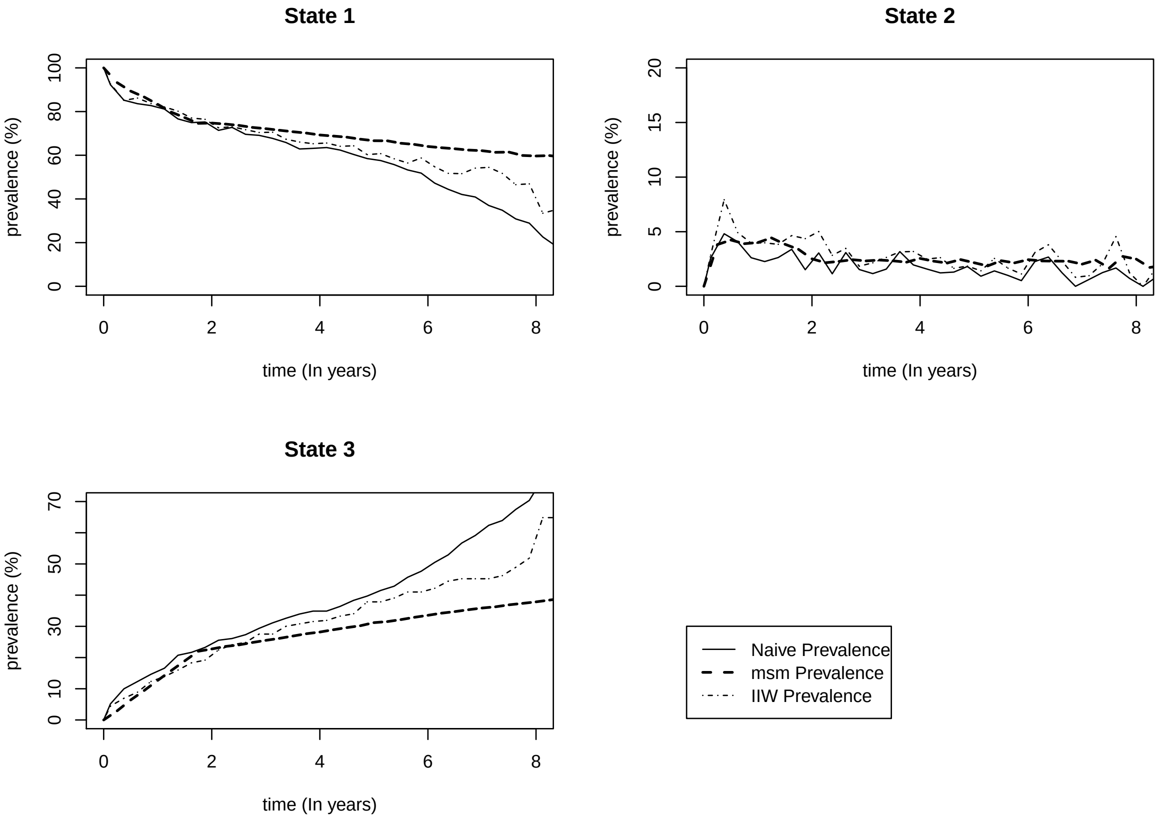

Figure 4.15: Comparison of naive, msm, and IIW prevalences for non-injection drug users in the FARV1 group.

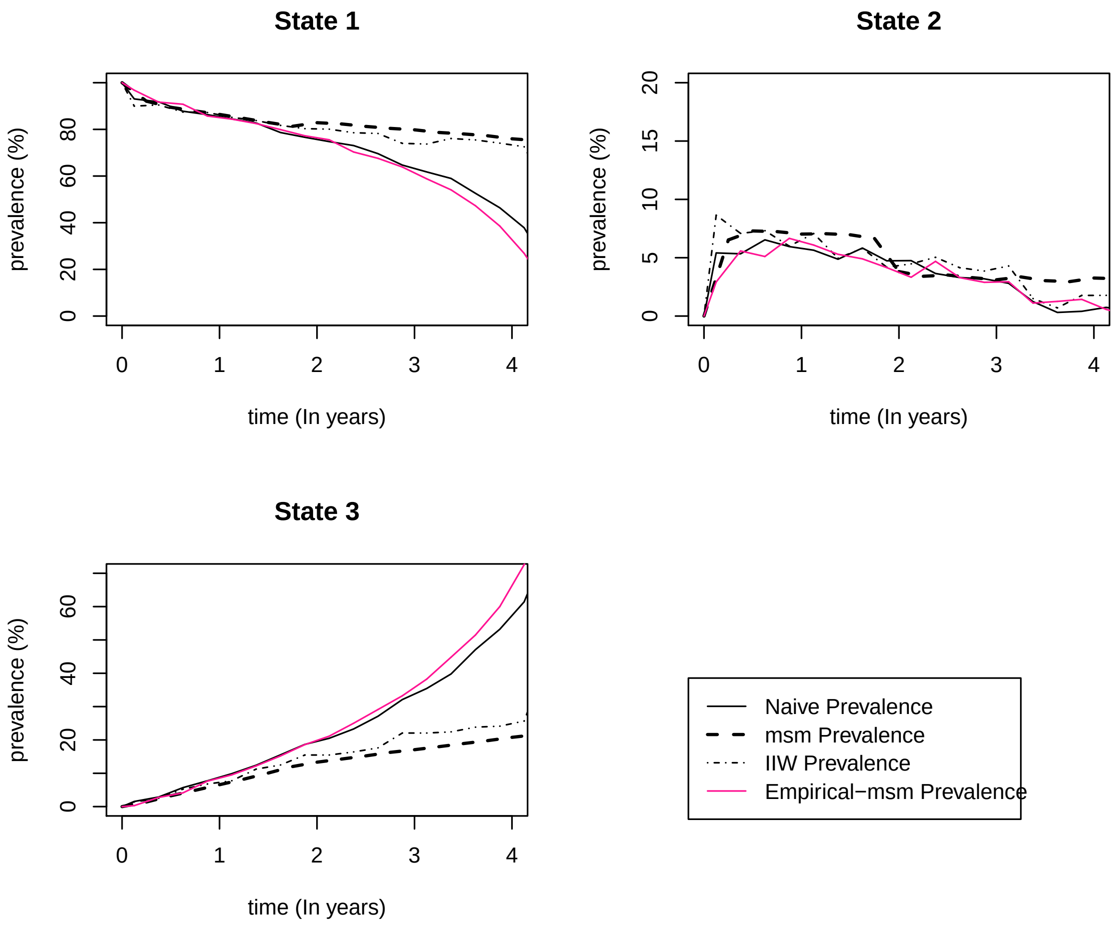

Figure 4.16: Comparison of naive, msm, empirical-msm, and IIW prevalences for all individuals in the FARV2 group.

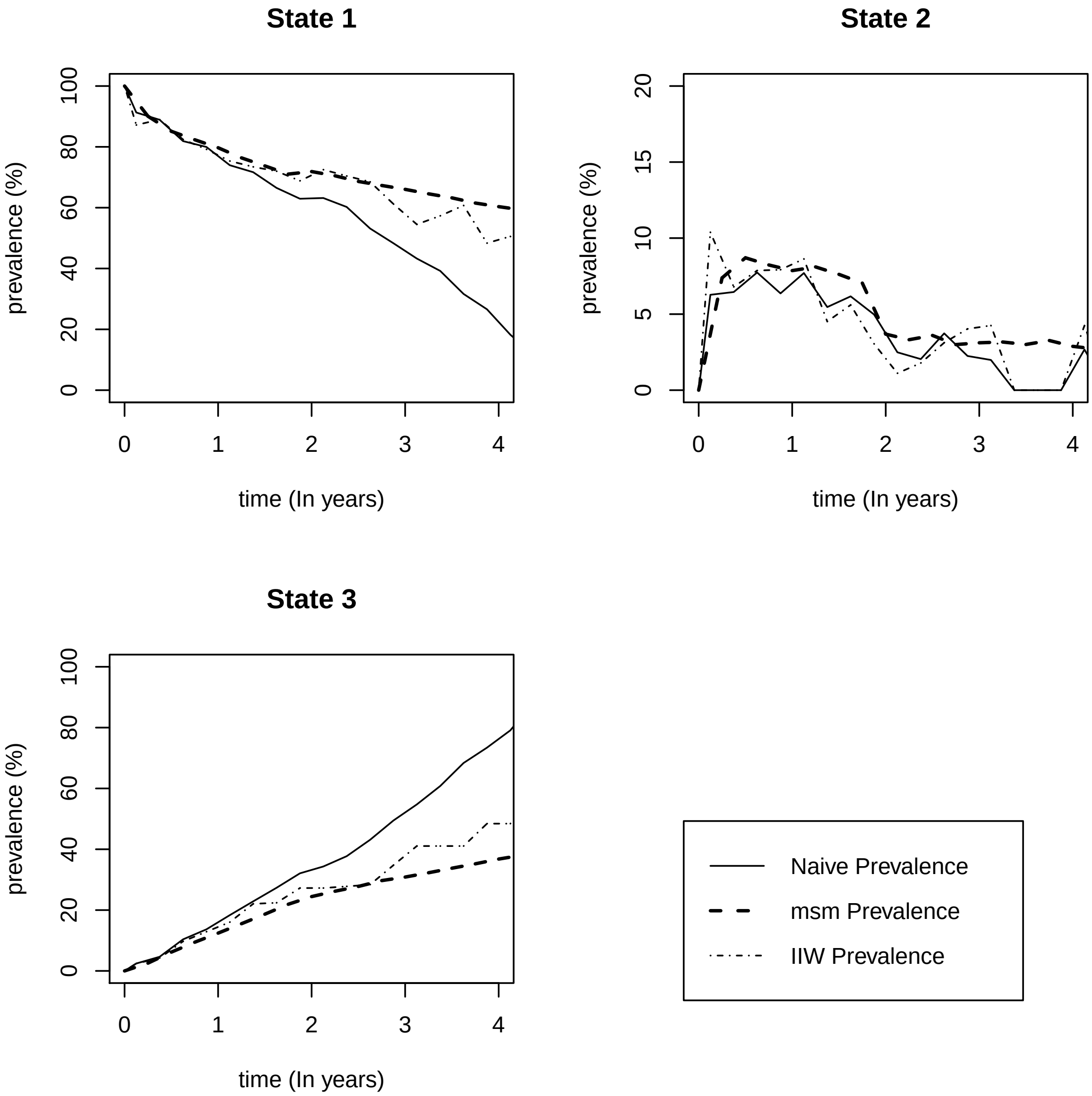

Figure 4.17: Comparison of naive, msm, and IIW prevalences for injection drug users in the FARV2 group.

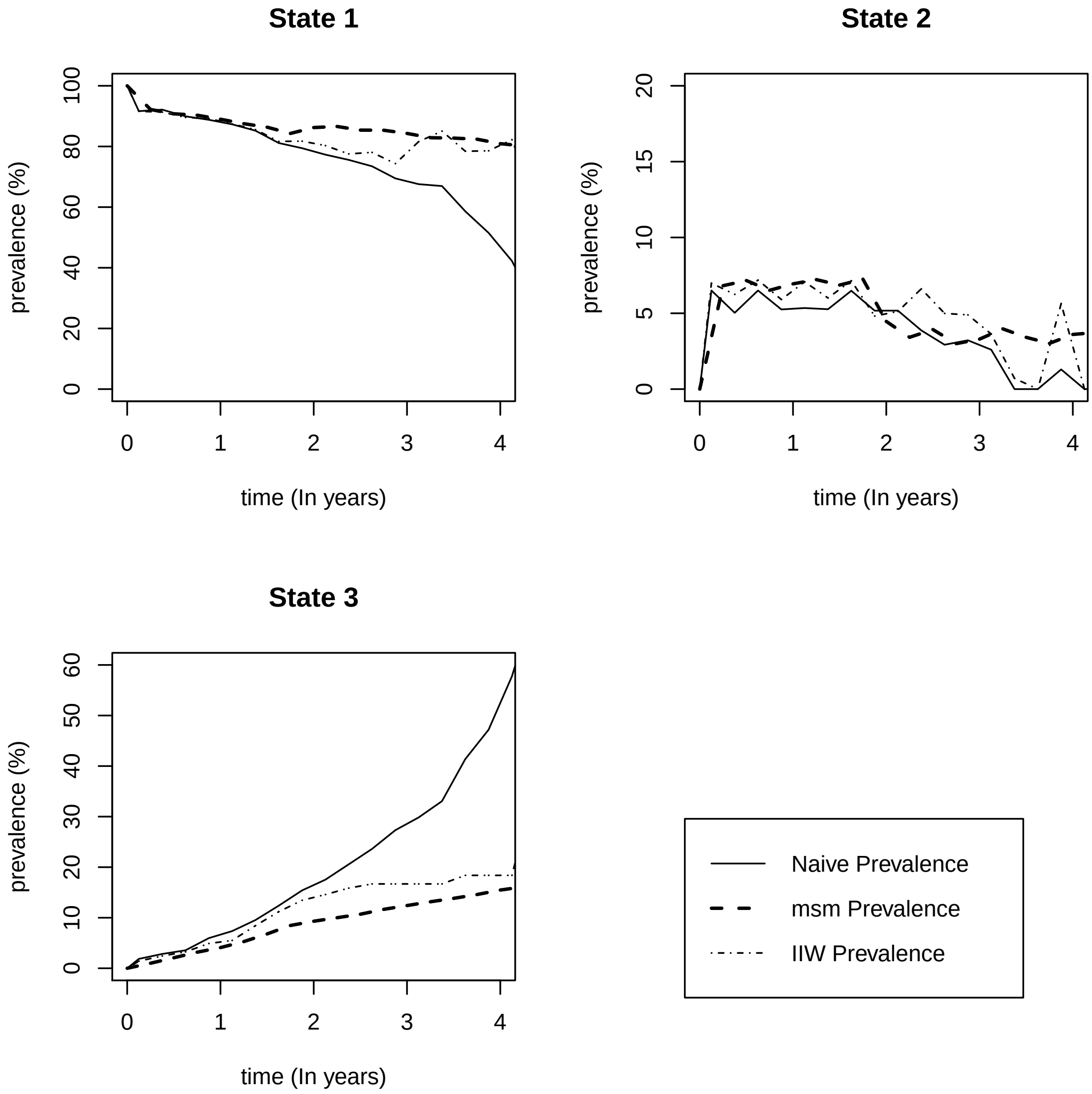

Figure 4.18: Comparison of naive, msm, and IIW prevalences for non-injection drug users in the FARV2 group.

The figures show that there is good agreement between the msm and IIW prevalence estimates when we have sufficient number of individuals in the three states at different times t. However, as time increases, the number of individuals in states 1 and 2 decreases, and the plots show that the msm and IIW prevalence curves get farther apart for the large t values. Tables 4.16 and 4.17 show the number of individuals at different times t. The good agreement of the msm and IIW estimates provides support for the msm models. However, the naive curves do not agree with these two estimates in many cases, and they would lead us to believe erroneously that the multistate model did not estimate prevalences well.

We note that Figures 4.13 to 4.18 provide considerable information on time to viral rebound. In particular, injection drug users, as might be expected, experience viral rebounds sooner and more frequently than non-users. The individuals in FARV1 also experience higher rate of viral rebounds; this could be due to improvements in cART over the years since 2000, or differences in the distribution of covariate values. We note that estimates can also be based on the Cox model of Section 4.4.1. However, these estimates can also be biased due to the dependent observation times. Figures 4.19 to 4.22 show the Cox models probability estimates for time to viral rebound along with the corresponding probability estimates obtained from fitting the multistate models for the injection drug users and non users in the both FARV1 and FARV2 groups. The"msm probability" estimates are similar to the $\hat{P}_{13}(t)$ estimates in Figures 4.14, 4.15, 4.17, and 4.18. The figures show that there is good agreement between the two estimates, and the dependent observation times did not have that much effect on the estimation of the Cox model probabilities. Finally, we remark that the IIW prevalence estimates could be smoothed; this is a topic for further development.

Table 4.16: The number of individuals in different groups of the FARV1 at specified time points (time is in years).

	Time								
Group	0	1	2	3	4	5	6	7	8
FARV1	834	820	800	781	753	685	560	459	412
Non-injection drug users	272	272	271	264	258	232	196	159	103
Injection drug users	272	262	253	247	235	220	186	167	163

Table 4.17: The number of individuals in different groups of the FARV2 at specified time points (time is in years).

	Time								
Group	0	0.5	1	1.5	2	2.5	3	3.5	4
FARV2	1035	1011	876	724	480	361	292	225	156
Non-injection drug users	323	318	284	243	191	156	120	98	71
Injection drug users	287	279	235	194	156	137	107	90	78

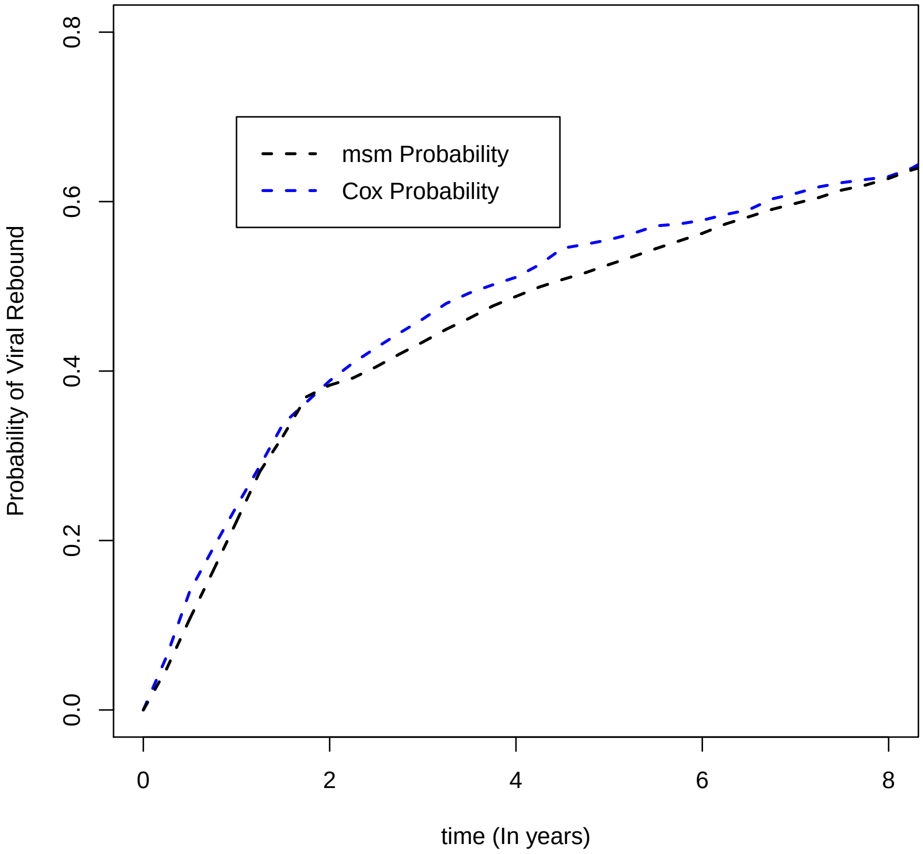

Figure 4.19: The msm and Cox model probability estimates of time to viral rebound for injection drug users in the FARV1 group.

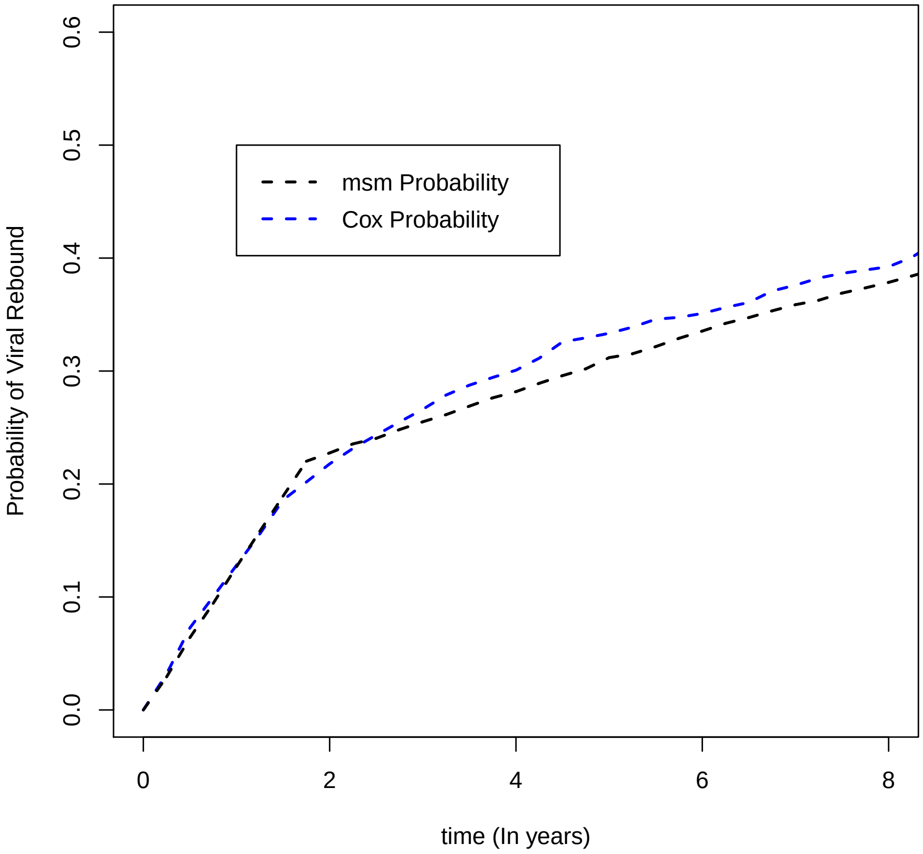

Figure 4.20: The msm and Cox model probability estimates of time to viral rebound for non injection drug users in the FARV1 group.

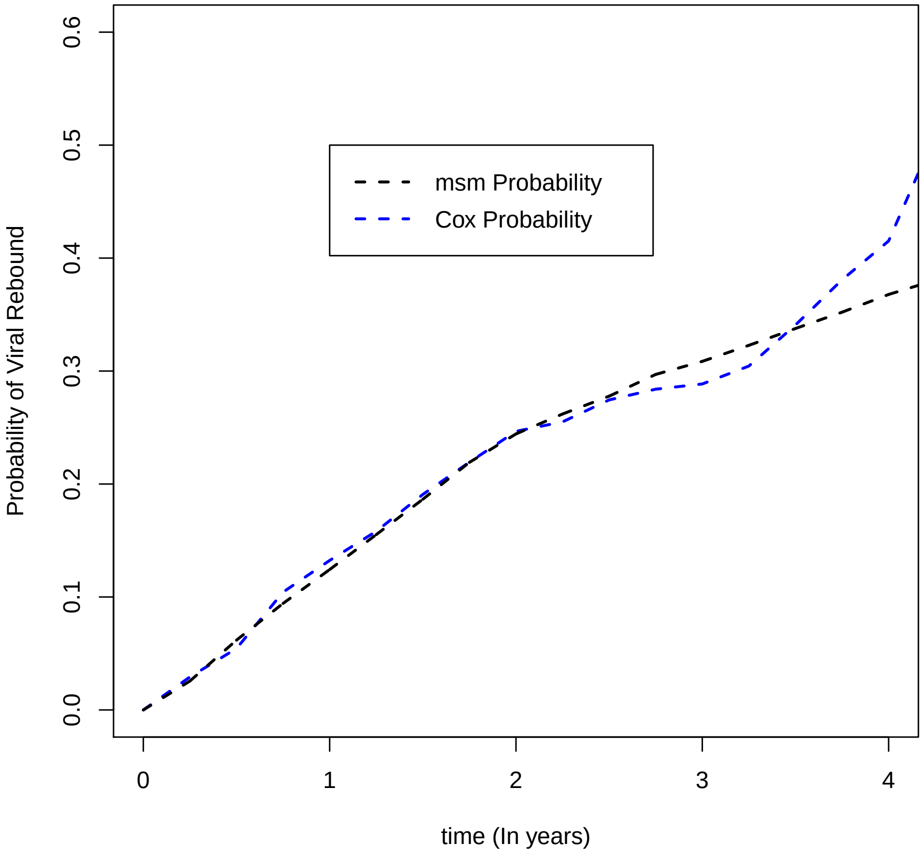

Figure 4.21: The msm and Cox model probability estimates of time to viral rebound for injection drug users in the FARV2 group

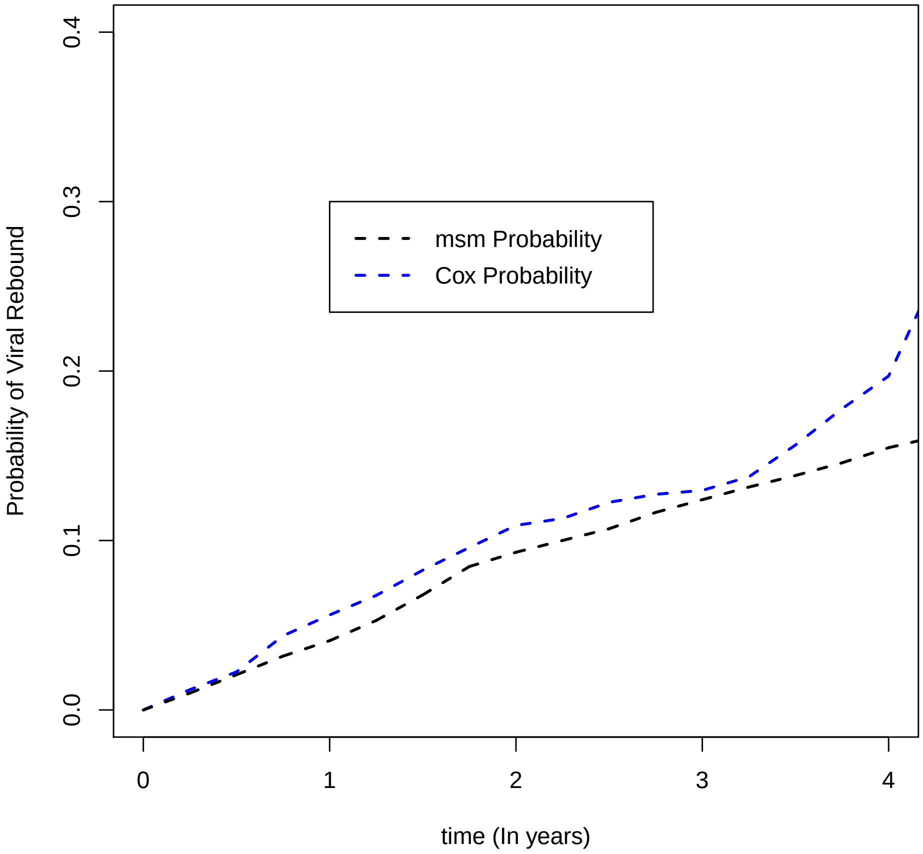

Figure 4.22: The msm and Cox model probability estimates of time to viral rebound for non injection drug users in the FARV2 group.

4.5 Concluding Remarks

In this chapter, we considered process-dependent observation times, and the importance of accounting for it when estimating marginal process features such as failure time distributions and prevalence probabilities was discussed. We proposed the IIW estimation method which incorporates the association between the gap times between visits and previous process history. Simulations were conducted to evaluate the bias-correction performance of the proposed estimates. The simulation results demonstrate that the IIW prevalence estimates are much less biased than naive estimates that ignore the dependency between inter-visit times and previous process history. In observational studies like CANOC, where times between visits vary considerably across individuals, there is a strong possibility that this irregularity of followup visits may be induced by their dependency on previous process history. The IIW method was used here to correct bias due to this dependency. However, in order to compare the IIW prevalence estimates and corresponding estimates obtained from Markov models more precisely, it is required to have confidence intervals based on the IIW estimates. One approach for computing the variance estimates is to assume a parametric model for the observation process intensity, and then apply White (1982) results on the estimating function for the observation process parameters and (4.3) estimating function to obtain the asymptotic covariance matrix for the observation process parameter estimates as well as the IIW prevalence estimates. An alternative approach for computing variances for the IIW prevalence estimates is to use bootstrap. We can generate B samples from the fitted observation model and compute the IIW prevalence estimates for each sample. Then, we can get the empirical standard deviation for each estimator. The development of variance estimates and confidence intervals is a needed area of study, and will be considered as future work.

Chapter 5

Model Assessment with Intermittent Observation

5.1 Introduction

The assessment of bi-directional multistate models with irregular intermittent observation times is challenging. Kalbfleisch and Lawless (1985) considered model checking and tests for Markov models with regular intermittent observations. Titman and Sharples (2010a) reviewed methods for assessing Markov models with irregular observation times and the msm package in R provides model checking approaches for models with piecewise-constant intensities. The methods in the msm package involve the comparison of empirical and model-based estimates such as prevalence functions and transition probabilities. The msm package also provides Pearson-type tests of fit that are discussed by Titman (2009, 2010) and Titman and Sharples (2010a). There are several problems with methods that are based on the comparison of empirical and model-based estimates. We will discuss these issues in

Section 5.2.

Another difficulty in the assessment of multistate models with intermittent observation is the dependency of transition intensities of Markov models on internal covariates that are related to the previous process history. In studies with intermittent observations, the values of time-dependent covariates are only known at observation times, and their values are unknown at times between visits. In this situation, likelihood functions based on observed data can be computed by assuming that covariate values for a given individual are constant between visit times. Then, the transition intensity estimates can be obtained by maximizing the likelihood function. However, the assumption that the covariate values are constant between observation times can be problematic when visit times are far apart, and the direct estimation of prevalence probability estimates is complicated because of various internal covariate paths for each individual.

This chapter has several objectives. One is to present a method for computing prevalence probabilities in multistate models in the presence of internal covariates. We also discuss how the prevalence estimates obtained from fitting flexible parametric models to intermittently observed data are robust for non-Markov processes, assuming that the observation times are independent of process history. The chapter then provides a discussion on difficulties which arise for model assessment tests based on comparison of empirical and model-based estimates. Then, the use of likelihood ratio tests within the Markov process family is suggested, and methods for estimating the power of likelihood ratio tests of fit are proposed. Another objective of the chapter is to propose a method for comparing models based on different outcome spaces in terms of prediction. Finally, we apply the methodology presented here to the assessment and prediction of viral rebounds for individuals in the CANOC study.

The remainder of this chapter is as follows. Section 5.2 discusses the estimation of prevalence probabilities with Markov models including internal covariates. Section 5.3 considers the effect of misspecification of the Markov assumption on estimation of prevalence probabilities based on flexible parametric models. Section 5.4 discusses difficulties for model assessment and proposes the use of likelihood ratio tests. In Section 5.5, we propose a method for comparison of different models in terms of prediction. Section 5.6 examines viral rebounds for individuals in the CANOC study using the methods discussed in this chapter.

5.2 Estimation of prevalence probabilities based on multistate models involving internal covariates

We discussed in Section 2.2 the likelihood functions (2.3)-(2.5) based on intermittently observed processes that had Markov multistate models. The computation of transition probabilities for processes that depend on fixed or external time-dependent covariates is straightforward, provided covariates are fixed between visit times. Prevalence estimates are also readily computed in these cases. However, if the process involves internal covariates, we can not compute prevalence probability estimates directly, because of many internal covariate paths that can occur for each individual. Thus, we propose simulating a large number of sample paths for each individual in order to approximate prevalence probability estimates. In many analyses, fixed covariates take on a small number of distinct values, and all individuals start from the same initial state, and then we can simulate paths for different combinations of fixed covariates.

We assume that the process is of modulated Markov form, with transition intensities

$$q_{rs}(t|H(t),Y(t^-)=r)=q_{0rs}(t)exp(\beta^T z(t)),\quad r\neq s, \tag{5.1}$$

and that the baseline intensities are piecewise-constant. We assume that an internal covariate is a deterministic function of previous observed state-history. For simplicity, we suppose there is only a single time-dependent internal covariate $z(t)$ in the model. We simulate the process $\{Y(t), t > 0\}$ at a discrete set of times $t_j (j = 1, ..., m)$, and assume that $z(t) = z_j$ is fixed over $(t_{j-1}, t_j]$. Thus, we have a time-homogeneous Markov process over $(t_{j-1}, t_j]$ with the transition intensity matrix Q_j having elements $q^{(j)}_{rs} exp(\beta^T z_j), r \neq s$. By knowing $Y(t_0) = y_0$, parameter values $q^{(j)}_{rs}$, β, and that the state $Y(t_j)$ has a multinomial distribution given $Y(t_{j-1}) = r$, an algorithm for simulating $\{Y(t_j), j = 1, ..., m\}$ is as follows:

1. Set $Y(t_0) = y_0$, $Z_1 = 0$. Then, for $j = 1, ..., m$,
2. Compute $Q_j = (q^{(j)}_{rs} exp(\beta' Z_j))$ and $P^{(j)} = exp(Q_j(t_j - t_{j-1}))$.
3. Generate $Y(t_j) = y_j$ from multinom $(1; P^{(j)}_{y_{j-1},s} \quad (s = 1, ..., K))$.
4. If $j = m$, stop, otherwise compute Z_{j+1} based on $\{Y(t_0) = y_0, ..., Y(t_j) = y_j\}$, and return to 2.

By simulating a large number (B) of sample paths, we can estimate prevalence probabilities of interest. In particular, if $y_0 = 1$ then

$$\hat{P}_{1r}(t_j)=\frac{1}{B}\sum_{b=1}^{B} I(Y_b(t_j)=r),\quad r=1,...,K, \tag{5.2}$$

where $\{Y_b(t), t > 0\}$ is the b'th simulated sample path. We illustrate the use of this procedure in Section 5.6.

5.3 Misspecification of the Markov assumption

Aalen et al. (2001) and Datta and Satten (2001) showed that the Markov (Aalen-Johansen) nonparametric estimates of prevalence probabilities based on continuous observation are robust for non-Markov processes, assuming that the loss to followup is independent of the process history. This suggests that flexible parametric models such as piecewise-constant models with sufficient number of pieces would also be fairly robust in the case where observation is intermittent and the observation times are not too far apart, and independent of the process history. To examine this, we considered the three-state model in Figure 3.1-(b). We assumed that the true process is semi-Markov, with transition intensities of Weibull form,

$$q_{rs}(t|H(t)) = a_{rs}B(t)^{b_{rs}}, \quad r \neq s, \tag{5.3}$$

where $B(t)$ is the time since entry to the current state. For the model parameters, we used the values $a_{12} = a_{23} = 1.5, a_{21} = 0.707, b_{12} = b_{23} = 0.5, b_{21} = -0.5$. We considered a single sample of $10,000$ individuals with all individuals starting from state 1 at time $t = 0$. Individuals were followed over time interval $(0, 4]$, and were observed with regular inter-visit times, $\Delta t = 0.5, 1$, or 2. Tables 5.1 and 5.2 show estimates of prevalence probabilities at times $t = 1, 2, 3, 4$ obtained by fitting three different piecewise-constant Markov models. In the tables, PC(0,4), PC(0,2,4), and PC(0,1,2,3,4) stand for models with 1, 2, and 4 pieces. For instance, in PC(0,2,4) the two pieces are $t \in [0, 2]$ and $t \in (2, 4]$; the one-piece model is time-homogeneous. The “Empirical values” in the tables are the estimates of the true prevalence probabilities, and should be close to the true values, since they are based on a sample of 10,000 processes.

As we can see from the Tables 5.1 and 5.2, piecewise-constant models can provide prevalence probability estimates with small biases. The prevalence estimates from the model with 4 pieces and visit gap time of 0.5 are close to the true prevalence values, except for $t = 1$, and the estimates are reasonably close when models with 1 or 2 pieces are used. However, msm was not able to fit the 4 piece model when inter-visit times were 1, and only the one-piece model could be fitted with gaps of size 2. The simulation thus shows that the prevalence estimates obtained from Markov models with several pieces are quite robust here. This implies that prevalence estimates will not be able to detect non-Markov behavior of true processes, since they are robust in terms of violation of the Markov assumption. In addition, prevalence function estimates can be adjusted for dependent irregular visit times, as discussed in Chapter 4. Therefore, we can conclude that tests of fit based on prevalence functions are not good options for detecting non-Markov behaviour. They can, however, detect departures from assumed forms of time-dependence in transition intensities for Markov models.

5.4 Likelihood ratio test

There are different problems with assessing the model fit on the basis of comparison of model-based and empirical estimates. Here, we review these difficulties, and suggest the use of the likelihood ratio test in order to overcome some of these issues. First, as we discussed in Section 5.3, prevalence estimates based on fitted Markov piecewise-constant models tend to be robust and so are not good at detecting non-Markov behaviour. Second, we discussed in Chapter 4 that the empirical estimates of prevalences are biased when we have dependent observation times, and we proposed nonparametric estimates that adjust for this bias. However, this adjustment affects the distribution of test statistics and consequently obtaining reliable p-values of formal tests of fit can be challenging. A third problem is

Table 5.1: Estimates $\hat{P}_{11}(t)$ from fitting piecewise-constant Markov models with 1, 2 and 4 pieces when the true model is semi-Markov and gap time between visits is $\Delta t = 0.5, 1, 2$.

		$t=1$	$t=2$	$t=3$	$t=4$
	Empirical values	0.663	0.472	0.341	0.252
$\Delta t = 0.5$	PC(0,4) estimates	0.616	0.461	0.354	0.273
	PC(0,2,4) estimates	0.638	0.4962	0.350	0.255
	PC(0,1,2,3,4) estimates	0.687	0.480	0.344	0.253
$\Delta t = 1$	PC(0,4) estimates	0.625	0.461	0.349	0.267
	PC(0,2,4) estimates	0.637	0.483	0.351	0.256
	PC(0,1,2,3,4) estimates	—	—	—	—
$\Delta t = 2$	PC(0,4) estimates	0.642	0.462	0.343	0.257
	PC(0,2,4) estimates	—	—	—	—
	PC(0,1,2,3,4) estimates	—	—	—	—

Table 5.2: Estimates $\hat{P}_{12}(t)$ from fitting piecewise-constant Markov models with 1, 2 and 4 pieces when the true model is semi-Markov and gap time between visits is $\Delta t = 0.5, 1, 2$.

		$t = 1$	$t = 2$	$t = 3$	$t = 4$
	Empirical values	0.248	0.225	0.169	0.121
$\Delta t = 0.5$	PC(0,4) estimates	0.234	0.200	0.155	0.120
	PC(0,2,4) estimates	0.231	0.203	0.168	0.124
	PC(0,1,2,3,4) estimates	0.217	0.221	0.172	0.123
$\Delta t = 1$	PC(0,4) estimates	0.232	0.205	0.160	0.123
	PC(0,2,4) estimates	0.238	0.217	0.163	0.120
	PC(0,1,2,3,4) estimates	—	—	—	—
$\Delta t = 2$	PC(0,4) estimates	0.225	0.211	0.166	0.126
	PC(0,2,4) estimates	—	—	—	—
	PC(0,1,2,3,4) estimates	—	—	—	—

when we have irregular intermittent observation, in order to compute empirical prevalence or transition counts as in msm, we need to interpolate and impute the states of individuals at common set of times, or else to group individuals. However, the formal tests of fit do not handle this well with regard to calculation of p-values. Moreover, test statistics based on prevalence or transition counts involve a mixture of multinomial variables, and limiting distributions of test statistics depend on unknown parameter values.

Considering the above mentioned difficulties, we suggest using likelihood ratio tests based on nested models within the family of Markov processes. The likelihood ratio tests have a number of advantages over the other model checking methods that we mention here. These tests are easily implemented and are asymptotically equivalent to the Pearson tests based on transition counts when individuals are observed at a common set of times (Section 2.4.2). In addition, we can use asymptotic chi-squared approximations for obtaining reliable p-values of likelihood ratio tests. Another advantage is that the presence of dependent observation times of the type considered in Chapter 4 does not have an effect on transition probability estimates, and consequently the likelihood ratio tests would not be affected by this issue. A further advantage of these tests is that the assessment of models involving fixed or time-dependent covariates can be carried out easily through computing likelihood functions of null and expanded models. However, the problem with recognizing the non-Markov property still remains, and this should be taken into account when considering internal covariates which can make models non-Markov.

Finally, it is usually of interest to know the power for discriminating between different Markov models when planning studies or followup. In this section, we present different approaches for estimating the power of likelihood ratio tests. A nested multistate model

M_0 can be tested against a model M_1 that includes it via a likelihood ratio statistic

$$\Lambda = 2logL(\hat{\theta}_1) - 2logL(\hat{\theta}_0), \tag{5.4}$$

where θ_0 and θ_1 are parameters corresponding to models M_0 and M_1, respectively. Λ is asymptotically chi-squared under the null model M_0, assuming that the observation times are independent of the process history. The degrees of freedom d for this asymptotic chi-squared distribution is $dim(\theta_1) - dim(\theta_0)$, which is the difference in the dimensions of the parameter spaces for models M_0 and M_1. Under model M_1, Λ is approximately non-central chi-squared with degrees of freedom d and the non-centrality parameter δ. This can be used for approximating the power for an alternative model by assuming that for a sample of size n, the likelihood ratio statistic Λ_n is approximately chi-squared with non-centrality parameter $\delta_n = n\gamma$. Now, we can estimate γ by the fact that

$$E(\Lambda_n) \doteq E(\chi^2(d, \delta_n)) = d + \delta_n. \tag{5.5}$$

One approach for estimating γ is to generate different samples of n individuals, and to compute Λ for each sample. Since d is known, we can obtain δ_n via estimating $E\Lambda_n$ by $\bar{\Lambda}_n$. So, $\hat{\delta}_n = \bar{\Lambda}_n - d$, and $\hat{\gamma} = \frac{\bar{\Lambda}_n - d}{n}$. If n is large, we can also estimate power directly by the proportion of samples giving Λ_n larger than the critical value under M_0. However, this approach is computationally time consuming, because of the need to generate many different samples of size n, and then fitting multistate models M_0 and M_1 to each.

Another way to estimate γ is to generate a single large sample of size N and compute the observed likelihood ratio statistic Λ_N. Then, we can estimate $\gamma = \frac{\delta_N}{N}$ as $\hat{\gamma} = \frac{\Lambda_N - d}{N}$. We can then estimate δ_n for an arbitrary sample size n by $\hat{\delta}_n = n\hat{\gamma}$.

A third approach is related to asymptotic results for maximum likelihood. Suppose we write $\theta = (\alpha, \beta)$ such that the null hypothesis is $\beta = 0$. Then, γ is estimated by $\hat{\beta}'\hat{V}^{-1}\hat{\beta}$, where $\hat{V}$ is N times the estimated covariance matrix for $\hat{\beta}$ for the sample of size N. We will discuss here the details of this approach. We know that under a sequence of alternative models M_1 that approach the null model at rate $1/\sqrt{n}$, the asymptotic distribution for Λ is $\chi^2(d, \delta)$. So, we can test the null hypothesis $\beta = 0$ for a sequence of alternatives $\beta = \frac{\phi}{\sqrt{n}}$, and it can be shown that as n approaches ∞, the limiting distribution of Λ_n under the alternative hypothesis $\phi = 0$ is $\chi^2(d, \delta)$, where $\delta = \lim\limits_{n \to \infty} \hat{\phi}'(nv\hat{a}r(\hat{\beta}))^{-1}\hat{\phi}$. To show this, we use the fact that $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, nv\hat{a}r(\hat{\beta}))$ as $n \to \infty$, and equivalently $\sqrt{n}(\hat{\beta} - \frac{\phi}{\sqrt{n}}) \xrightarrow{d} N(0, nvar(\hat{\beta}))$ as $n \to \infty$. Thus, $\sqrt{n}\hat{\beta} \xrightarrow{d} N(\phi, nvar(\hat{\beta}))$ as $n \to \infty$, and consequently $n\hat{\beta}'(nv\hat{a}r(\hat{\beta}))^{-1}\hat{\beta} \xrightarrow{d} \chi^2(d, \phi'(nvar(\beta))^{-1}\phi)$ as $n \to \infty$. Therefore, $\hat{\beta}'(v\hat{a}r(\hat{\beta}))^{-1}\hat{\beta} \xrightarrow{d} \chi^2(d, \beta'(var(\beta))^{-1}\beta)$ as $n \to \infty$, and γ can be estimated by $\frac{1}{N}\hat{\beta}'(v\hat{a}r(\hat{\beta}))^{-1}\hat{\beta}$, where N is large.

For illustration, we used the second method of estimating the power of likelihood ratio tests. We considered testing the time homogeneity of the Markov model in Figure 3.1(b). We assumed that the alternative M_1 is a piecewise-constant model with 4 pieces (0, 1], (1, 2], (2, 3], and (3, 4]. The corresponding transition intensities (q_{12}, q_{21}, q_{23}) for these pieces are (0.44, 2.13, 0.86), (0.90, 1.57, 0.95), (1.19, 1.21, 1.13) and (1.30, 1.07, 1.46), respectively, and the prevalence probability values are given in Table 5.3. We generated a single large sample of $N = 10,000$ processes from M_1, so that each of the processes was observed with gap times $\Delta t = 0.25, 0.5$, or 1 over time interval (0, 4]. We then fitted a time-homogeneous model (M_0) and a 4-piece model M_1 and obtained the likelihood ratio statistic (5.4), and thus $\hat{\gamma}$. The results on power are given in Table 5.4. Powers were computed for a test of size 0.05, as $P(\chi^2(q, \widehat{n\gamma}) > \chi^2_{0.95}(q))$, where $\chi^2_{0.95}(q)$ is the 0.95 quantile for $\chi^2(q)$. We have also computed confidence intervals for the power of tests, by using the fact that $\chi^2(d, \delta)$ has variance $2(d + 2\delta)$.

Table 5.3: The prevalence probability values for the simulation study of Section 5.4.

t	$P_{11}(t)$	$P_{12}(t)$	$P_{13}(t)$
1	0.804	0.118	0.078
2	0.533	0.205	0.262
3	0.311	0.187	0.502
4	0.169	0.112	0.718

The estimated powers in Table 5.4 are all high. However, they are based on asymptotic power formulas, so might not be highly accurate for smaller values of n. In order to check the accuracy of the power estimates for a typical case $n = 100$ and $\Delta t = 0.5$ presented in Table 5.4, we conducted a simulation which includes $B = 1000$ samples of $n = 100$ processes. We simulated the data under the model M_1, and fitted both models M_0 and M_1 to each sample to compute the corresponding likelihood ratio statistic Λ. Then, we estimated the power for the test with size $\alpha = 0.05$ by taking the proportion of the B samples for which the likelihood ratio statistic exceeds the $\chi^2_{0.95}(9)$. The simulation gave 972 samples, all of which had $\Lambda > \chi^2_{0.95}(9)$. However, for the remaining 28 samples, msm was not able to fit the M_1 model. This was due to not having any observed transitions for at least one of the time intervals in the 4-piece model M_1. We nevertheless can conclude here that the estimated power of 0.996 in Table 5.4 is not very inaccurate.

5.5 Prediction and comparison of models

We may assess models in terms of their model fit or predictive performance. The former is usually evaluated by checking how closely models fit the observed data while assessment of the latter requires new data or cross-validation. In Chapter 4 and earlier sections of

Table 5.4: Estimates of the power of likelihood ratio tests with size 0.05.

Δt	$\hat{\gamma}$ (%95 CI)	n	Power of test (%95 CI)
0.25	0.343 (0.320, 0.366)	50	0.843 (0.811, 0.870)
		100	0.995 (0.991, 0.997)
		200	1.00 (1.00, 1.00)
		500	1.00 (1.00, 1.00)
		1000	1.00 (1.00, 1.00)
0.5	0.354 (0.330, 0.377)	50	0.856 (0.826, 0.881)
		100	0.996 (0.993, 0.998)
		200	1.00 (1.00, 1.00)
		500	1.00 (1.00, 1.00)
		1000	1.00 (1.00, 1.00)
1	0.270 (0.249, 0.290)	50	0.724 (0.682, 0.762)
		100	0.974 (0.961, 0.983)
		200	1.00 (1.00, 1.00)
		500	1.00 (1.00, 1.00)
		1000	1.00 (1.00, 1.00)

this chapter, we discussed the assessment of Markov multistate models based on the data set from which they were derived. In this section, we describe an approach to assess the predictive performance of these models. In addition, this method will enable us to compare families of models with different structures in terms of their ability to predict specific outcomes. For instance, this allows us to compare models based on failure time distributions and multistate models. Moreover, the comparison of multistate models with different state spaces becomes possible.

Van Houwelingen and Putter (2012, Chapter 3), discussed three ways for measuring the predictive performance of models. They mentioned 1) the absolute error, 2) the Brier score, and 3) Kullback-Leibler (KL) score. However, because of some unpleasant properties of the absolute error approach, they restricted their attention to the Brier and KL scores. Here, we consider the KL measure, since it has a close connection to maximum likelihood estimation, which we use. Our method is similar to Liquet and Commenges (2011), where they used the KL score to compare the illness-death model with a survival model in terms of their ability to predict the survival probability.

Suppose that we want to compare two models in terms of their ability to predict the probability for the time T of a specific event, given that it has not occured by time t_0^*, and still under observation between times t_0^* and t^*. We also have a vector of covariates $z(t_0^*)$. In this context, our predictive probability is

$$P(T > t^* | T > t_0^*, Z(t_0^*)) \quad t^* > t_0^*. \tag{5.6}$$

The KL measure based on a predictive probability model for (5.6) can be written as

$$\begin{aligned} KL(M) &= V(t^*) log P(T > t^* | T > t_0^*, z(t_0^*)) \qquad (5.7) \\ &+ (1 - V(t^*)) log\{1 - P(T > t^* | T > t_0^*, z(t_0^*))\}, \end{aligned}$$

where $V(t^*) = I(T > t^*)$, and “M” is the model for (5.6). In the comparison of two models based on the KL measure, the model that has larger KL(M) value is preferred.

KL(M) can be extended to consider a set of times $t_1^* < t_2^* < ... < t_p^*$. In this case, the KL measure compares models based on observed data $\{t_{ij}, Z_i(t_{ij}), T_i; j = 1, ..., m_i\}$ for individuals $i = 1, ..., n$. We can write the extended KL measure as

$$\begin{aligned} KL(M) = \sum_{i=1}^{n}\sum_{j=1}^{m_i^T} & \{ \quad R_{ij} logPr(T_i \leq t_{ij} | T_i > t_{i,j-1}, Z_{ij}) \\ & + \quad (1 - R_{ij}) logPr(T_i > t_{ij} | T_i > t_{i,j-1}, Z_{ij})\}, \end{aligned} \tag{5.8}$$

where covariate vector $Z_{ij} = Z_i(t)$ is fixed over $t_{i,j-1} < t \leq t_{ij}$, $R_{ij} = I(T_i \in (t_{i,j-1}, t_{ij}])$ and $m_i^T = min\{max_j(j : T_i \leq t_{ij}), m_i\}$. In order to compare the KLs of two models, we use cross-validation. That is, we randomly split the n individuals into L groups of approximately the same size. At each stage, we drop group l from the data, and fit model M to the remaining groups; then, we calculate the average $KL_l(M)$ for the l'th group as the observed value (5.8), summing only over i in group l. The measure of predictive performance is then given by $KL = \sum_{l=1}^{L}(1/L)KL_l(M)$. We illustrate this method in Section 5.6, where we compare the Cox model and multistate models in terms of the probability of viral rebounds.

5.6 Application to CANOC data

The importance of viral fluctuations on the occurrence of viral rebound has been raised in clinical research related to HIV (Grennan et al.(2012)). In this section, we define a time-dependent covariate “Blip” as an occasion in which a viral load value of $50 - 999$ copies/mL at a visit is preceded and followed by observed viral load values below 50 copies/mL. Here, we investigate the association between Blip and risk of viral rebounds, considering Blip as an internal covariate. First, we fit Markov multistate models similar to those in Section 3.4.1. Second, we compare the multistate and Cox models in terms of predicting the probability of viral rebounds, using the KL measure. Third, we check the fit of multistate models by comparison of the nonparametric prevalence probability estimates with the model-based estimates.

5.6.1 Multistate models with covariate Blip

We fit Markov regression models with transition intensities

$$q_{rs}(t|z(t)) = q_{0rs}(t) exp(\beta_{rs}^T z(t)), \quad r, s = 1, 2, 3. \tag{5.9}$$

In these models $z(t)$ includes fixed covariates and two time-varying covariates, (i) $z_1(t) = I(t > 2)$, and (ii) $z_2(t) = I(\text{Blip has been observed before time } t)$. Piecewise-constant Markov models for both FARV1 and FARV2 groups were fitted using the msm package in R. Tables 5.5 and 5.6 show estimates and standard errors for the baseline intensities and covariate effects for two-piece models with cut point 2 years.

The estimated covariate effects associated with the Tables 5.5 and 5.6 are similar to the ones that we had earlier in Tables 3.5 and 3.6, except for the new covariate Blip. From Tables 5.5 and 5.6, we can see that individuals who have experienced a Blip are more likely to move from state 1 to state 2, and to remain longer in this state. However, they are also less likely to make the transition into the viral rebound state 3. We remark that Grennan et al. (2012) found in analysis for the entire CANOC cohort that Blips involving a viral load of over 500 were associated with higher risk of viral rebound. Our analysis does not support this.

Table 5.5: The estimated regression coefficients and standard errors (in brackets) of piecewise constant Markov multistate model for the FARV1 group.

Transition	$log(q_{0rs})$	Age (> 45)	IDU (=Yes)	IDU (=NA)	cART (=PI)	Blip	Time
		vs ($\leq$ 45)	vs (No)	vs (No)	vs (NNRTI)		
1 → 2	-0.927	0.237	0.176	0.035	0.638*	0.784*	-0.884*
	(0.138)	(0.124)	(0.134)	(0.161)	(0.117)	(0.141)	(0.125)
2 → 1	2.161	0.277	-0.544*	0.248	0.359*	-0.160	-0.159
	(0.170)	(0.151)	(0.174)	(0.177)	(0.143)	(0.158)	(0.151)
2 → 3	1.709	-0.429*	0.165	-0.006	-0.342*	-0.915*	-0.645*
	(0.152)	(0.138)	(0.148)	(0.169)	(0.129)	(0.187)	(0.143)

$*|\frac{\hat{\beta}}{se(\hat{\beta})}| > 1.96$

Table 5.6: The estimated regression coefficients and standard errors (in brackets) of piecewise constant Markov multistate model for the FARV2 group.

Transition	$log(q_{0rs})$	Age (> 45)	IDU (=Yes)	IDU (=NA)	cART (=PI)	Blip	Time
		vs ($\leq$ 45)	vs (No)	vs (No)	vs (NNRTI)		
1 → 2	-0.537	-0.089	0.171	0.212	0.555*	0.476*	-0.837*
	(0.201)	(0.138)	(0.161)	(0.173)	(0.170)	(0.174)	(0.184)
2 → 1	2.384	-0.304*	-0.034	0.330	0.127	-0.060	-0.210
	(0.237)	(0.143)	(0.178)	(0.173)	(0.194)	(0.174)	(0.182)
2 → 3	0.764	-0.556*	0.820*	-0.239	-0.994*	-0.247	-0.172
	(0.252)	(0.197)	(0.238)	(0.266)	(0.202)	(0.259)	(0.297)

$*|\frac{\hat{\beta}}{se(\hat{\beta})}| > 1.96$

Table 5.7: -2 log likelihood values for models without covariate Blip.

Model	FARV1	FARV2
Time-homogeneous	7550.006	5697.425
Two-piece	7417.230	5669.010

Table 5.8: -2 log likelihood values for models with covariate Blip included.

Model	FARV1	FARV2
Time-homogeneous	7470.755	5677.601
Two-piece	7332.435	5641.255

Tables 5.7 and 5.8 show the -2loglikelihood values for time homogenous and two-piece models. These models include covariates age, IDU, and cART. The covariate Blip was also included in models considered in Table 5.8. As we can see, the likelihood ratio tests of models with Blip against models without Blip show significant Blip effects. In addition, the two-piece models have substantially smaller -2loglikelihood values compared to the time homogeneous models.

5.6.2 KL of msm and Cox models for time to viral rebound

We next compare Cox and multistate models in terms of their ability to predict the probability of viral rebounds. We use two-piece multistate models, and consider covariates age, IDU, cART, and Blip in fitting both multistate and Cox models. For comparison of the KL values of two models, we use $L = 10$ fold cross-validation method. Tables 5.9 and 5.10 show the results of the fitted Cox model for time to viral rebound in the FARV1 and FARV2 groups. The estimated covariate effects associated with these tables are similar to the ones that we had earlier in Tables 4.12 and 4.13, except for the new covariate Blip. We can also compare the covariate effect estimates of Blip in Tables 5.9 and 5.10 with the covariate effect estimates obtained from multistate models in Tables 5.5 and 5.6. All tables do not show any significant association between occurrence of Blip and having viral rebound.

Tables 5.11 and 5.12 show the KL values of fitted Cox and multistate models for the lth, ($l = 1, 2, ..., 10$) subgroup. As we can see from these tables, the multistate model has somewhat larger KL values than the Cox model in the FARV1 group. However, in the FARV2 group total, KL values of the multistate and Cox models are very close to each other.

5.6.3 Prevalence probability estimates

For checking the fit of multistate models, we estimate prevalence using the Naive, IIW, and msm prevalence probabilities, and compare them in plots similar to those presented in Section 4.4. However, since the process involves the internal covariate Blip, we are not able to compute the msm prevalence estimates directly. Therefore, we use the Section 5.2 simulation approach to approximate them. We simulate $B = 1000$ sample paths for each combination of age, IDU, and cART categories. Each path is simulated at a discrete set of times $t_j = 0, 0.25, 0.5, 0.75, ...$, assuming that the covariate Blip has fixed values over $(t_{j-1}, t_j]$.

We also computed Naive prevalence estimates following the same approach presented in Section 4.4.2. The IIW method was applied by assigning weights based on the Cox model results given in Tables 4.14 and 4.15. Figures 5.1 and 5.4 show aggregated estimates of the prevalence probabilities $P_{1r}(t), r = 1, 2, 3$ for all individuals in the FARV1 and FARV2

Table 5.9: The results of the fitted Cox model for time to viral rebound in the FARV1 group.

Covariates	$\hat{\beta}$	$se(\hat{\beta})$	Relative Risk (RR)	95% CI for RR	P-value
Age ($>$ 45)	-0.354	0.122	0.701	(0.553, 0.890)	0.003*
IDU (=Yes)	0.704	0.134	2.021	(1.554, 2.629)	1.54e-07*
IDU (=NA)	-0.100	0.151	0.905	(0.674, 1.216)	0.508
cART (=PI)	0.144	0.114	1.155	(0.925, 1.443)	0.204
Blip	0.093	0.174	1.097	(0.779, 1.545)	0.595

$*|\frac{\hat{\beta}}{se(\hat{\beta})}| > 1.96$

Table 5.10: The results of the fitted Cox model for time to viral rebound in the FARV2 group.

Covariates	$\hat{\beta}$	$se(\hat{\beta})$	Relative Risk (RR)	95% CI for RR	P-value
Age ($>$ 45)	-0.320	0.181	0.726	(0.510, 1.036)	0.078
IDU (=Yes)	0.893	0.216	2.443	(1.598, 3.733)	3.68e-05*
IDU (=NA)	-0.279	0.246	0.756	(0.467, 1.224)	0.256
cART (=PI)	-0.311	0.184	0.733	(0.511, 1.051)	0.091
Blip	0.147	0.254	1.158	(0.704, 1.904)	0.563

$*|\frac{\hat{\beta}}{se(\hat{\beta})}| > 1.96$

Table 5.11: KL comparison of msm and Cox models for time to viral rebound in the FARV1 group.

Subgroups	multistate model	Cox model (F-H)
1	-130.106	-136.169
2	-166.254	-169.781
3	-118.617	-126.848
4	-134.207	-136.758
5	-175.594	-171.924
6	-123.198	-123.643
7	-142.557	-143.917
8	-130.179	-127.907
9	-148.163	-148.579
10	-159.737	-160.495
Total	-1428.612	-1446.021

Table 5.12: KL comparison of msm and Cox models for time to viral rebound in the FARV2 group.

Subgroups	multistate model	Cox model (F-H)
1	-60.161	-63.420
2	-43.101	-46.586
3	-45.742	-46.186
4	-70.312	-73.492
5	-89.888	-86.038
6	-87.619	-84.288
7	-56.097	-58.136
8	-56.084	-56.453
9	-76.333	-70.274
10	-29.607	-32.849
Total	-614.944	-617.722

Table 5.13: The msm prevalence estimates for Figures 5.1 and 4.13.

Time		1	2	3	4	5	6	7	8
$P_{11}(t)$	without Blip	0.806	0.706	0.671	0.639	0.608	0.578	0.553	0.525
	with Blip	0.807	0.704	0.665	0.631	0.559	0.569	0.537	0.514
$P_{12}(t)$	without Blip	0.039	0.026	0.025	0.021	0.021	0.021	0.017	0.020
	with Blip	0.041	0.023	0.024	0.022	0.022	0.022	0.024	0.019
$P_{13}(t)$	without Blip	0.155	0.267	0.304	0.340	0.371	0.400	0.429	0.455
	with Blip	0.152	0.273	0.311	0.347	0.379	0.409	0.439	0.466

groups. Figures 5.2, 5.3, 5.5, and 5.6 represent the aggregated probabilities for injection drug users or non users. The msm prevalence estimates are based on the fitted multistate models associated with Tables 5.5 and 5.6. As shown in Figures 5.1 to 5.6, the model-based prevalence estimates are close to the estimates obtained from the IIW method. The good agreement of the msm and IIW estimates provides support for the msm models. However, the naive curves do not agree with these two estimates in many cases, and they would lead us to believe erroneously that the multistate model did not estimate prevalences well.

Comparing the msm prevalence estimates in Figures 5.1 to 5.6 with the msm estimates in Figures 4.13 to 4.18 shows there are differences in the msm prevalence estimates obtained from models with and without Blip.In order to see the differences, a couple of prevalence estimate values for the both models are given in Tables 5.13 and 5.14. However, considering the variability of the IIW estimates, we are not able to conclude which model outperforms the other based on the prevalence estimates.

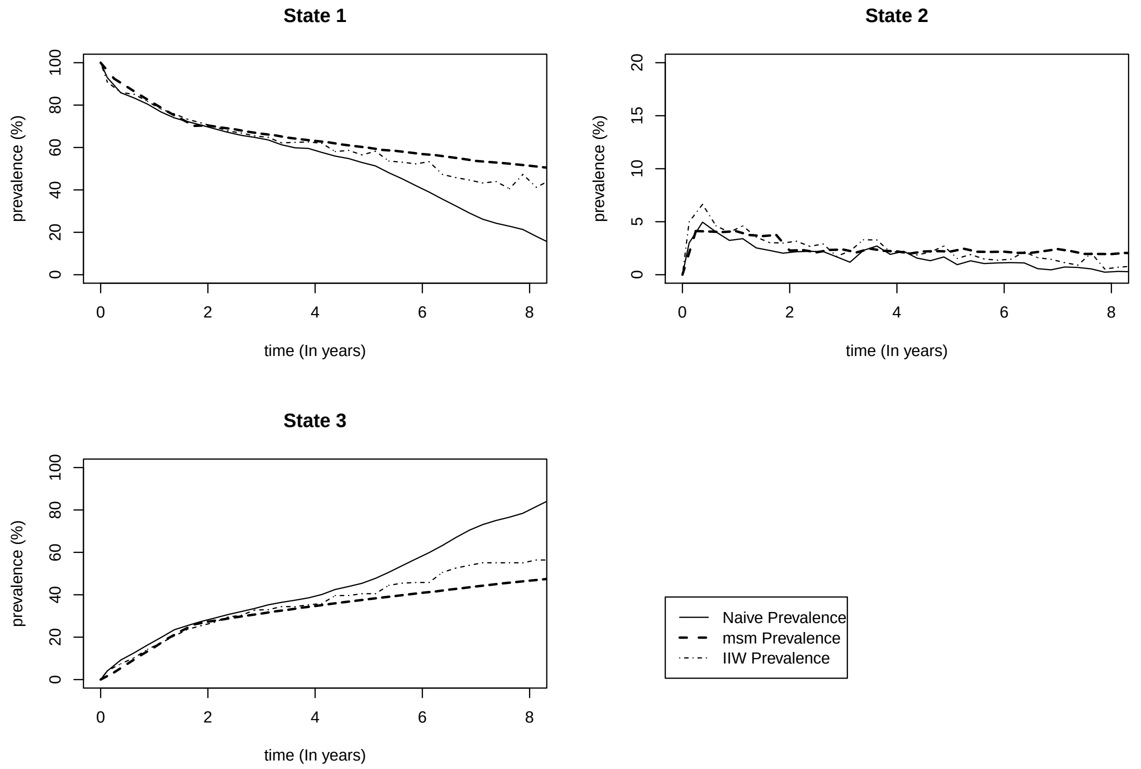

Figure 5.1: Comparison of Naive, msm, and IIW prevalences for all individuals in the FARV1 group, considering the BLIP covariate

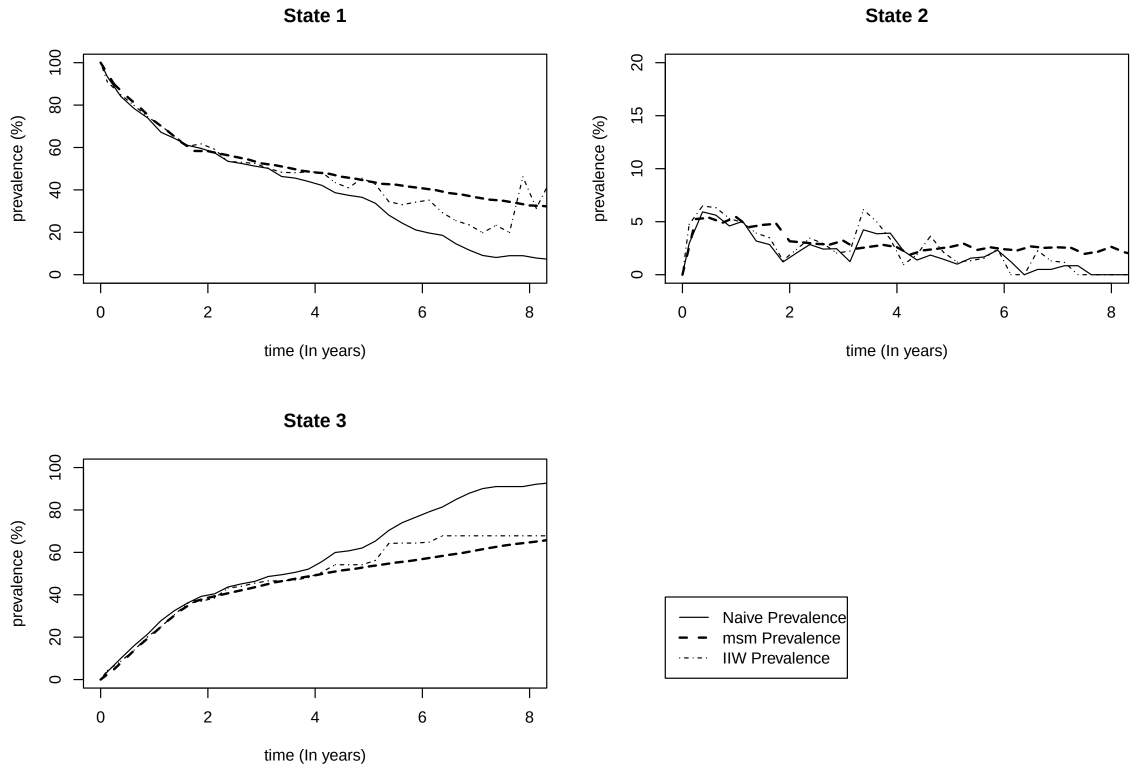

Figure 5.2: Comparison of Naive, msm, and IIW prevalences for injection drug users in the FARV1 group, considering the BLIP covariate

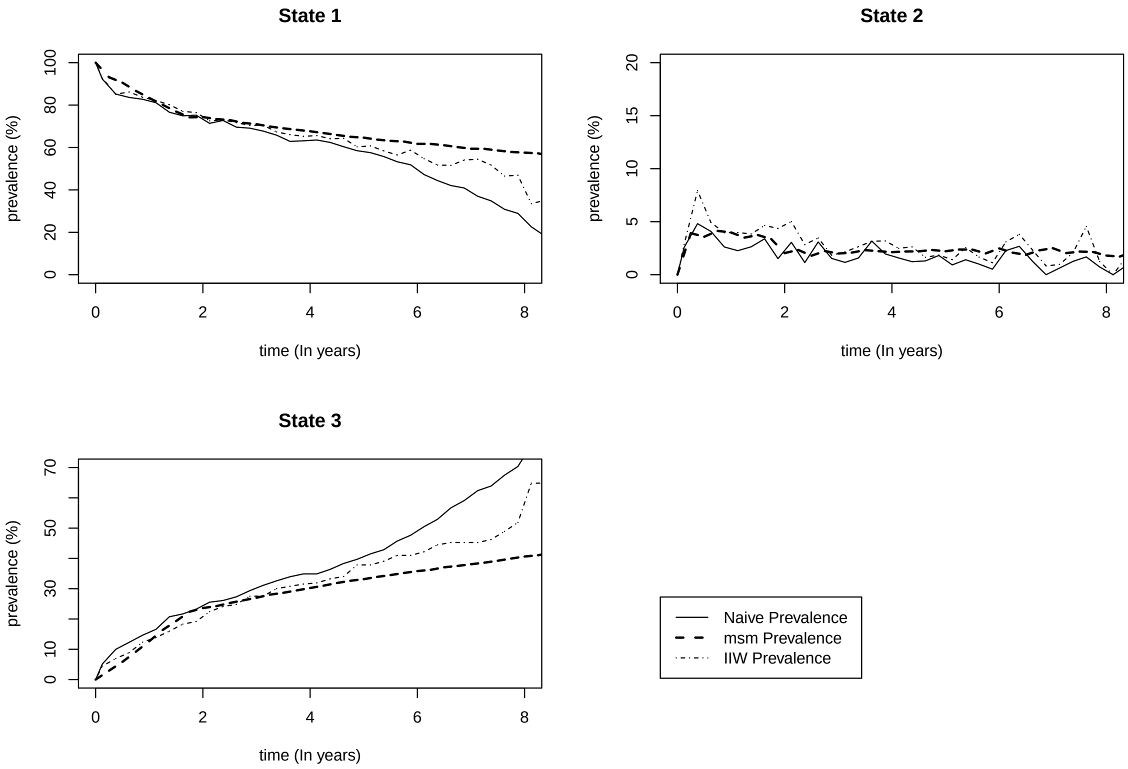

Figure 5.3: Comparison of Naive, msm, and IIW prevalences for non-injection drug users in the FARV1 group, considering the BLIP covariate

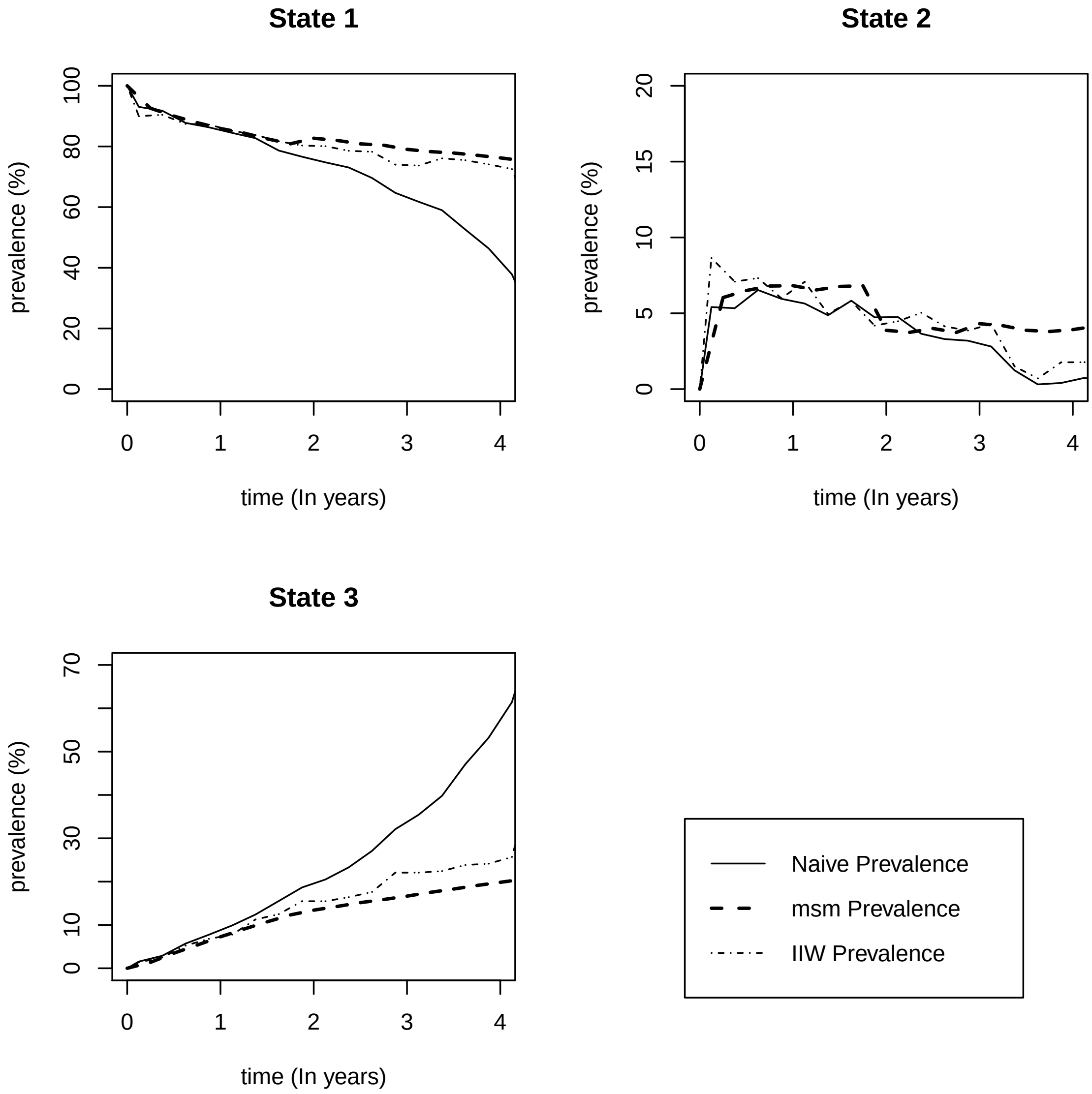

Figure 5.4: Comparison of Naive, msm, and IIW prevalences for all individuals in the FARV2 group, considering the BLIP covariate

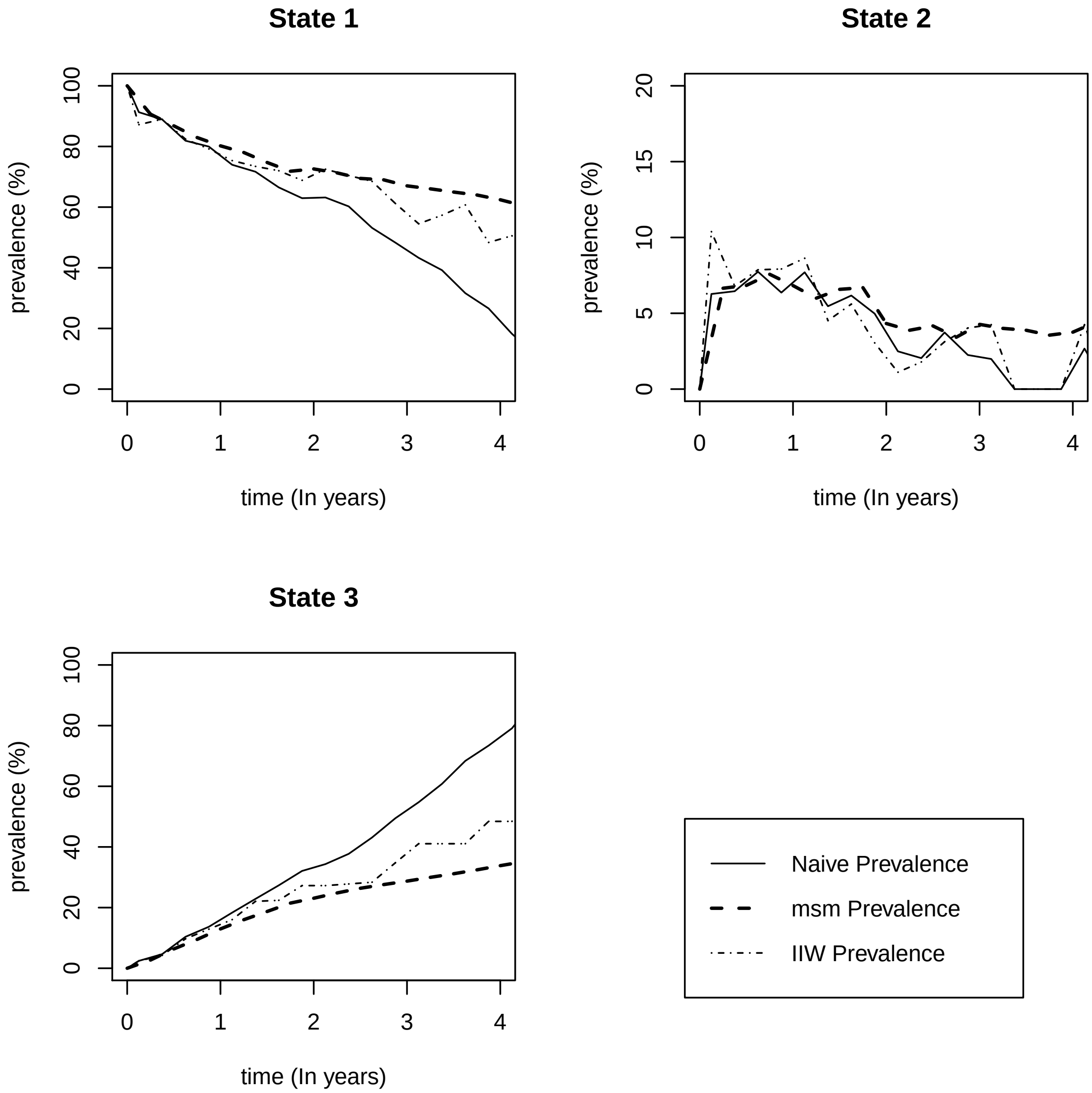

Figure 5.5: Comparison of Naive, msm, and IIW prevalences for injection drug users in the FARV2 group, considering the BLIP covariate

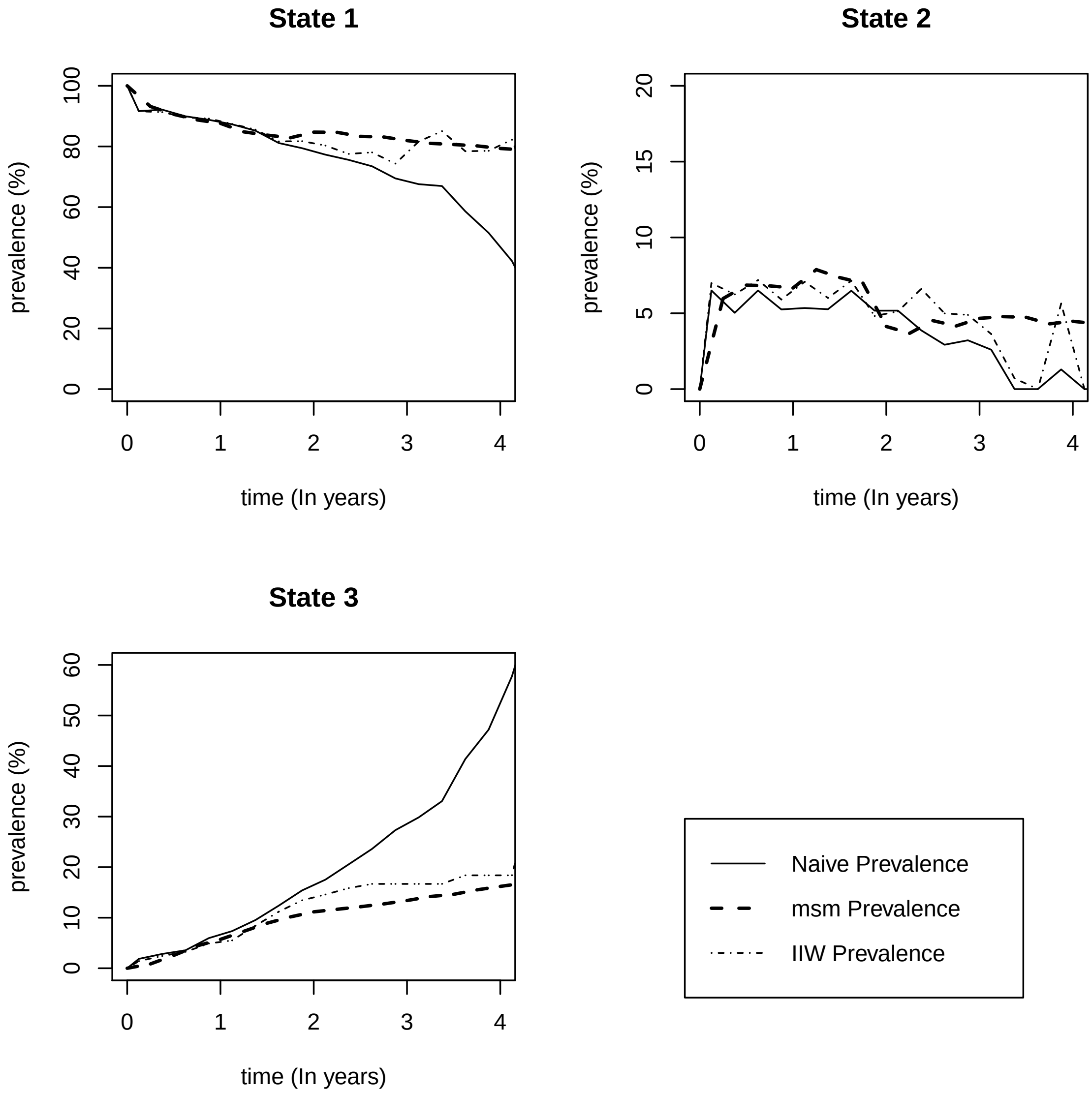

Figure 5.6: Comparison of Naive, msm, and IIW prevalences for non-injection drug users in the FARV2 group, considering the BLIP covariate

Table 5.14: The msm prevalence estimates for Figures 5.4 and 4.16.

Time		1	2	3	4
$P_{11}(t)$	without Blip	0.864	0.828	0.798	0.760
	with Blip	0.860	0.829	0.792	0.763
$P_{12}(t)$	without Blip	0.070	0.038	0.031	0.033
	with Blip	0.069	0.040	0.042	0.039
$P_{13}(t)$	without Blip	0.066	0.133	0.171	0.208
	with Blip	0.071	0.131	0.165	0.198

5.7 Concluding remarks

In this chapter, we proposed a method for estimating prevalence probabilities in modulated Markov models for which transition intensities depend on internal covariates related to previous process history. This method, which is based on an algorithm that simulates sample paths of processes, can be used in model checking and prediction of the modulated Markov models. Modeling the covariate process, enables us to estimate the distribution of time to viral rebound in the presence of a time dependent covariate such as blip. This is an advantage of the multistate models over survival methods.

We also discussed difficulties in model assessment with irregular intermittent observations when comparing empirical and model-based estimates. We addressed these problems using likelihood ratio tests within the Markov process family. In addition, methods of estimating the power of these tests were proposed. Finally, besides assessing models in terms of how well they fit the observed data, we may be interested in evaluating their performance in predicting new outcomes. In this chapter, we proposed a method which enables us to compare the prediction performance of different models, even if they have different outcome spaces.

Chapter 6

Topics for Research

In this last chapter of the thesis, we give a brief overview of the proposed methods in the previous chapters, and briefly discuss further research topics.

6.1 Overview

As a brief summary, in Chapter 3 of the thesis, we discussed the use of asymptotic covariance matrices for investigation of the loss of efficiency in estimation of parameters for intermittently observed data. We concluded that in progressive models, the precision of transition intensity and probability estimators based on panel data, relative to the precision of estimates obtained from continuous observations, did not change much with the increase of the gap times between visits. In bi-directional models, the relative precision of transition intensity estimates dropped off rapidly, but the decrease was smaller for transition probabilities. In Chapter 4, we proposed the IIW method for estimating marginal process features such as failure time distributions and prevalence probabilities, when the observation times were process-dependent. We demonstrated via simulations that the IIW prevalence estimates have little bias but the naive estimates that ignore the dependency between inter-visit times and previous process history can have large bias. In Chapter 5, we developed a method that facilitates estimation of the prevalence probabilities in the presence of internal covariates. In addition, we addressed various problems that may arise in the assessment of multistate models with irregular intermittent observations by the use of likelihood ratio tests within the Markov process family. Finally, we proposed a method that compares the prediction performance of models based on different outcome spaces.

More detailed discussion and practical recommendations on the usage of the proposed methods are presented at the end of each individual chapter, and the illustrations involving CANOC data demonstrated practical aspects of their application.

6.2 Heterogeneity in Markov Models

Sometimes individuals may follow a Markov process, but even after conditioning on covariates that are significantly associated with the transition intensities, there may remain substantial heterogeneity with respect to the transition probabilities. In this case, there are basically two approaches to consider, each of which involves incorporating unobserved random effects.

The first is when the unobservable heterogeneity can be generally incorporated in a model through continuous random effects. In this approach, each individual's transition intensity matrix can include random effects. Introducing random effects to Markov models has two advantages. Firstly, it gives processes that depend on the whole event history, but still keeps the model parsimonious, unlike, e.g. models that increase the order of the Markov chain. Secondly, it accounts for and provides measures of the heterogeneity. Mandel and Betensky (2008) proposed random effects Markov models to account for heterogeneity in the population, and studied prediction based on them. Another examination of random effects in multi-state models is given by O'Keeffe et al. (2011).

The second approach for dealing with heterogeneity involves discrete random effects, and arises when there are unobservable and different types of individuals, and all individuals of the same type follow a Markov process with specific transition intensity matrix. We can either consider these models as latent class models or discrete mixture models. One such model is the so called Mover-Stayer model. This model extends the Markov process by allowing for the existence of a subgroup of individuals who, with probability one, will remain in their initial state throughout the observation period. These individuals are called "Stayers". The remaining individuals, called "Movers", follow a common Markov process with specific transition probabilities. This approach has been used by some authors (e.g. Goodman (1961), Frydman (1984), Cook and Lawless (2014)). This model is very restrictive and in general, more than two classes of individuals may be called for.

In the CANOC data, the viral load paths are highly variable between and within individuals, and continuous random effects may not represent such heterogeneity well. However, we might consider latent classes based on baseline covariates, and then use transition intensity matrices which are independent of covariates, conditioning on the classes. Another potential research topic is to compare random effects models with models that incorporate more dependence on past history, which will be discussed in the next section.

6.3 Models with dependence on history

One of the important challenges in modeling is to specify intensity functions that adequately describe life history processes or events. In the analysis of the CANOC viral load data, the occurrence of Blips and the impact of Blips on virologic rebounds are of interest, for example. However, the exact nature of a Blip is uncertain, and different definitions can be used to define this event practically. In the thesis, we used the definition of Blip in Grennan et al. (2012). However, this definition of Blip is affected by the gap times between visits. In the CANOC study, the viral load values were designed to be measured roughly every three months, but some people delayed or missed their visits. Thus, persons who missed their appointments while they were in state 2, could have moved back to state 1 by the next visit. Since these transitions were not observed, the Blip occurrence would not be noted. Ways to define internal covariates that can deal with irregular observation times is an important topic for research. For the CANOC data, we can investigate the effects of other covariates on the occurrence of viral rebounds. For instance, we may consider covariate "recent-Blip" which can stand for having experienced Blip within 3 months before time t; that is $Z(t) = 1\{\text{Blip in the last three months before } t\}$. This would help us to distinguish between the occurrence of Blip at any time in the past and the recent ones. In addition, it might give us a better understanding of the impact of viral load values on the occurrence of viral rebounds. van Houwelingen and Putter (2012) in Chapter 6 of their book considered the effects of different types of time dependent covariates on prediction of specific events.

6.4 Recurrent Events

We discussed the case where viral load processes produce a single event, viral rebound. However, we may be interested in considering viral rebounds as recurrent events. In this case, the occurrence of an earlier viral rebound may affect the probabilities of a new viral rebound. These type of effects can be modeled by defining covariates associated with the occurrence of the previous events, and then including these covariates in the transition intensities of the multistate models associated with the future.

In the CANOC study, we can use the multistate model in Figure 3.1(b), and the transition intensities in Sections 3.4 or 5.6 for modeling the viral load processes after the first viral rebound. However, we may set the time to zero after the occurrence of a viral rebound and subsequent viral suppression, and include covariates which indicate the occurrence of the previous viral rebounds in the transition intensities.

6.5 IIW Prevalence Estimates

In the second simulation study of subsection 4.3.4, we found that extreme weight values cause influential observations, and consequently high variability of the IIW-SM prevalence estimates. We recommended the use of truncation for the adjustment of large weight values. This approach was applied in the second simulation, and the results demonstrated that the variability in prevalence estimates was reduced to some extent. Another approach that can be considered as future work is to use parametric models instead of nonparametric semi-Markov models for the gap times. For instance, we can fit flexible parametric models such as piecewise constant. This may avoid extreme weight values and give estimates of θ

with less variability.

Another source of variability in nonparametric IIW prevalence estimates is insufficient number of observations at each time t, or time interval. This becomes more challenging when we deal with irregular observation times. In this situation, for each time t, there might be only one observation, leading to binary estimates of $P_{1r}(t)$ as discussed in subsection 4.3.3. In order to tackle this problem, we applied smoothing approaches such as loess for the simulation study, and a grouping method for the CANOC data. Since different smoothing methods may influence the estimation of IIW prevalences, the impact of these approaches requires further investigation.

In this thesis, we proposed the IIW prevalence estimates to reduce the bias in nonparametric estimates when the observation times are process dependent. However, in order to make a more precise comparison between estimates obtained from Markov models and corresponding IIW estimates, and more generally, to use the IIW estimates more broadly, we would like to construct confidence intervals based on the IIW estimates. Research is needed to develop variance estimates and confidence intervals.

Bibliography

[1] Aalen, O.O., Borgan, O., Fekjaer, H. (2001). Covariate adjustment of event histories estimated from Markov chains: the additive approach. *Biometrics*, **57**, 993-1001.

[2] Aalen, O.O., Borgan, O., Gjessing, H.K. (2008). *Survival and Event History Analysis: A Process Point of View*, New York, Springer.

[3] Aguirre-Hernandez, R., Farewel, V.T. (2002). A Pearson-type goodness-of-fit test for stationary and time-continuous Markov regression models. *Statistics in Medcine*, **21**, 1899-1911.

[4] Andersen, P.K., Borgan, O., Gill, R.D., Keiding, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer-Verlag

[5] Borgan, O., Fiaccone, R.L., Henderson, R., Barreto, M.L. (2007). Dynamic analysis of recurrent event data with missing observations, with application to infant Diarrhoea in Brazil. *Scandinavian Journal of Statistics*, **34**, 53-69.

[6] Buzkova, P., and Lumley, T. (2007). Longitudinal data analysis for generalized linear models with follow-up dependent on outcome-related variables. *Canadian Journal of Statistics*, **35**, 485-500.

[7] Buzkova, P., and Lumley, T. (2009). Semiparametric modeling of repeated measurements under outcome-dependent followup. *Statistics in Medcine*, **28**, 987-1003.

[8] Chi, Y.Y., and Ibrahim, J.G. (2006). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics*, **62**, 432-445.

[9] Cook, R.J. (2000). Information and efficiency considerations in planing studies based on two-state Markov process. *Journal of Statistical Research*, **34**, 161-178.

[10] Cook, R.J., Lawless, J.F. (2007). *Statistical Analysis of Recurrent Events*, New York: Springer Science+ Business Media, LLC.

[11] Cook, R.J., Lawless, J.F. (2014). Statistical issues in modeling chronic disease in cohort studies. *Statistics in Bioscience*, to appear.

[12] Copas, A.J, Farewell, V.T. (2001). Incorporating retrospective data into an analysis of time to illness. *Biostatistics*, **1**, 1-12.

[13] Cox, D.R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B*, **34**, 187-220.

[14] Datta, S., and Satten, G.A. (2001). Validity of the Aalen-Johansen estimators of stage occupation probabilities and Nelson-Aalen estimators of integrated transition hazards for non-Markov models. *Statistics and Probability Letters*, **55**, 403-411.

[15] de Stavola, B.L. (1988). Testing departures from time homogeneity in multistate Markov processes. *Applied Statistics*, **37**, 242-250.

[16] Diggle, P.J., Sousa, I., and Chetwynd, A.G. (2008). Joint modelling of repeated measurements and time-to-event outcomes: The fourth Armitage lecture. *Statistics in Medcine*, **27**, 2981-2998.

[17] Faucett, C.L., Thomas, D.C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: A GIBBS sampling approach. *Statistics in Medcine*, **15**, 16631685.

[18] Frydman, H. (1984). Maximum likelihood estimation in the mover-stayer model. *Journal of the American Statistical Association*, **79**, 632-638.

[19] Gentleman, R.C., Lawless, J.F., Lindsey, J.C., Yan, P. (1994). Multi-state Markov models for analysing incomplete disease history data with illustrations for HIV disease. *Statistics in Medcine*, **13**, 805-821.

[20] Goodman, L.A. (1961). Statistical methods for the mover-stayer model. *Journal of the American Statistical Association*. **56**, 841-868.

[21] Grennan, J.T. et al. (2012). Magnitude of virologic blips is associated with a higher risk for virologic rebound in HIV-infected individuals: a recurrent events analysis. *Journal of Infectious Diseases*, **205**, 1230-1238.

[22] Henderson, R., Diggle, P., AND Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, **4**, 465-480.

[23] Hogan, J.W., and Laird, N.M. (1997a). Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine*, **16**, 239-257.

[24] Hogan, J.W., and Laird, N.M. (1997b). Model-based approaches to analyzing incomplete longitudinal and failure time data. *Statistics in Medicine*, **16**, 259-272.

[25] Hogan, J.W., and Laird, N.M. (1998). Increasing efficiency from censored survival data by using random effects to model longitudinal covariates. *Statistical methods in medical research*, **7**, 28-48.

[26] Hwang, W., and Brookmeyer, R. (2003). Design of panel studies for disease progression with multiple stages. *Lifetime Data Analysis*, **9**, 261-274.

[27] Jackson, C.H. (2011). Multi-state models for panel data: the msm package for R. *Journal of Statistical Software*, **38** (8), 1-28.

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

[29] Kalbfleisch, J.D, Lawless, J.F. (1989). Some statistical methods for panel life history data. *Proceedings of the Statistics Canada Symposium on the Analysis of Data in Time*, Ottawa, Statistics Canada, 185-192.

[30] Kvist, K., Andersen, P.K., Angst, J., Kessing, L.V. (2010). Event dependent sampling of recurrent events. *Lifetime Data Analysis*, **16**, 580-598.

[31] Lawless, J.F. (2003). *Statistical Models and Methods for Lifetime Data.* 2nd edition. Wiley, Hoboken, NJ.

[32] Liang, Y., Lu, W., Ying, Z. (2009). Joint modeling and analysis of longitudinal data with informative observation times. *Biometrics*, **65**, 377-384.

[33] Lin, H., Scharfstein, D.O., and Rosenheck, R.A. (2004). Analysis of longitudinal data with irregular, outcome-dependent follow-up. *Journal of the Royal Statistical Society, Series B*, **66**, 791-813.

[34] Liquet, B., and Commenge, D. (2011). Choice of estimators based on different observations: Modified AIC and LCV criteria. *Scandinavian Journal of Statistics*, **38**, 268-287.

[35] Little, R. J. A., and Rubin, D. B. (1987). Statistical analysis with missing data. *John Wiley and Sons*, Inc., 1st ed.

[36] Liu, L., Huang, X. (2009). Joint analysis of correlated repeated measures and recurrent events processes in the presence of death, with application to a study on acquired immune deficiency syndrome. *Applied Statistics*, **58**, 65-81.

[37] Mandel, M., and Betensky, R.A. (2008). Estimating time-to-event from longitudinal ordinal data using random effects Markov models: application to multiple sclerosis progression. *Biostatistics*, **9**, 750-764.

[38] Mathieu, E., Loup, P., Dellamonica, P., Daures, J.P. (2005). Markov modelling of immunological and virological states in HIV-1 infected patients. *Biometrical Journal*, **47**, 834-846.

[39] Mehtala, J., Auranen, K., Kulathinal, S. (2011). Optimal designs for epidemiologic longitudinal studies with binary outcomes. *Statistical methods in medical research*, DOI: 10.1177/0962280211430663.

[40] O'Keeffe, A.G., Tom, B.D.M., and Farewell, V.T. (2011). A case study in the clinical epidemiology of psoriatic arthritis: multi-state models and causal arguments. *Applied Statistics*, **60**, 675-699.

[41] Palmer, A.K., Klein, M.B., Raboud, J., et al. (2011). The CANOC Collaboration.Cohort profile: the Canadian Observational Cohort collaboration. *International Journal of Epidemiology*, **40**, 2532.

[42] Proust-Lima, C., Taylor. J.M.G. (2009). Development and validation of dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modelling approach. *Biostatistics*, **10**, 535-549.

[43] Pullenayegum, E.M., and Feldman, B.M. (2013). Doubly robust estimation, optimally truncated inverse-intensity weighting and increment-based methods for the analysis of irregularly observed longitudinal data. *Statistics in Medicine*, **32**, 1054-1072.

[44] Raboud, J., Loutfy, M.R., Su, D., Bayoumi, A.M., Klein, M.B., Cooper, C., Machouf, N., Rouke, S., Walmsley, S., Rachlis, A., Harrigan, R., Smieja, M., Tsoukas, C., Montaner, J.S., Hogg, R.S. (2010). Regional differences in rates of HIV-1 viral load monitoring in Canada: Insights and implications for antiretroviral care in high income countries. *BMC Infectious Diseases*, **10**, 40.

[45] Sun, J. (2006). *The Statistical Analysis of Interval-Censored Failure Time Data.* New York: Springer.

[46] Titman, A.C. (2009). Comparison of the asymptotic null distribution of goodness-of-fit tests for multi-state models. *Lifetime Data Analysis*,**15**, 519-533.

[47] Titman, A.C., Sharples, L.D. (2008). A general goodness-of-fit test for Markov and hidden Markov models. *Statistics in Medicine*, **27**, 2177-2195.

[48] Titman, A.C., Sharples, L.D. (2010a). Model diagnostics for multi-state models. *Statistical Methods in Medical Research*, **19**, 621-651.

[49] Titman, A.C., Sharples, L.D. (2010b). Semi-Markov models with phase-type sojourn distributions. *Biometrics*, **66**, 742-752.

[50] Tom, B.D.M., Farewell, V.T. (2011). Intermittent observation of time-dependent explanatory variables: a multistate modelling approach. *Statistics in Medicine*, **30**, 3520-3531.

[51] Tsiatis, A.A., and Davidian, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika*, **88**,

447-458.

[52] Tsiatis, A.A., and Davidian, M. (2004). Joint modelling of longitudinal and time to-event data: an overview. *Statistica Sinica*, **14**, 809-834.

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

[54] van Houwelingen, J.C., Putter, H. (2012). *Dynamic Prediction in Clinical Survival Analysis*, Boca Raton: CRC Press.

[55] White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, **50**,1-26.

[56] Wulfsohn, M.S., and Tsiatis, A.A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, **53**, 330-339.

[57] Xu, J., Zeger, S.L. (2001). Joint analysis of longitudinal data comprising repeated measures and times to events. *Appl. Statist.* **50**, 375-387.

[58] Yang, Y., Nair, V.N. (2011). Parametric inference for time to failure in multistate semi-Markov models: a comparison of marginal and process approaches. *Canadian Journal of Statistics*, **39**, 537-555.

[59] Yu, M., Law, N.J., Taylor, J.M.G., and Sandler, H.M. (2004). Joint longitudinal survival- cure models and their application to prostate cancer. *Statistica Sinica*, **14**, 835-862.

[60] Yu, L., Tyas, S.L., Snowdown, D.A., and Kryscio, R.J. (2008). Effects of ignoring baseline on modelling transitions from intact cognition to dementia. *Computational Statistics and Data Analysis*, **9**, 3334-3343.

[61] Zhu, L., Sun, J., Tong, X., Pounds, S. (2011). Regression analysis of longitudinal data with informative observation times and application to medical cost data. *Statistics in Medicine*, **30**, 1429-1440.