The Application of Multistate Markov Models to HIV Disease Progression

Tarylee Reddy

March, 2011

The Application of Multistate Markov Models to HIV Disease Progression

by

Tarylee Reddy

Thesis submitted to the University of KwaZulu-Natal in fulfilment of the requirements for the Masters degree in Statistics.

UNIVERSITY OF KWAZULU-NATAL

WESTVILLE CAMPUS, DURBAN, SOUTH AFRICA

Disclaimer

This document describes work undertaken as a masters programme of study at the University of KwaZulu-Natal (UKZN). All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the institute.

Declaration

I, Tarylee Reddy, declare that

1. The research reported in this thesis, except where otherwise indicated, is my original research.
2. This thesis has not been submitted for any degree or examination at any other university.
3. This thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.
4. This thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:
   (a) Their words have been re-written but the general information attributed to them has been referenced
   (b) Where their exact words have been used, then their writing has been placed in italics and inside quotation marks, and referenced.
5. This thesis does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the thesis and in the references sections.

Tarylee Reddy

Student number 205505650 Signature

Prof Henry Mwambi

Supervisor Signature

March 14, 2011

Abstract

Survival analysis is a well developed area which explores time to single event analysis. In some cases, however, such methods may not adequately capture the disease process as the disease progression may involve intermediate events of interest. Multistate models incorporate multiple events or states. This thesis proposes to demystify the theory of multistate models through an application based approach. We present the key components of multistate models, relevant derivations, model diagnostics and techniques for modeling the effect of covariates on transition intensities.

The methods that are developed in the thesis are applied to HIV and TB data partly sourced from CAPRISA and the HPP programmes in the University of KwaZulu-Natal. HIV progression is investigated through the application of a five state Markov model with reversible transitions such that state 1: CD4 count ≥ 500, state 2: $350 \leq$ CD4 count < 500, state 3: $200 \leq$ CD4 count < 350, state 4: CD4 count < 200 and state 5: ARV initiation. The mean sojourn time in each state and transition probabilities are presented as well as the effect of covariates namely age, gender and baseline CD4 count on transition rates.

A key finding, consistent with previous research, is that the rate of decline in CD4 count tends to decrease at lower levels of the marker. Further, patients enrolling with a CD4 count less than 350 had a far lower chance of immune recovery and a substantially higher chance of immune deterioration compared to patients with a higher CD4 count. We noted that older patients tend to progress more rapidly through the disease than younger patients.

Contents

List of Figures

List of Tables

Acknowledgements

Professor Henry Mwambi has been the ideal supervisor. His advice, insightful criticisms and encouragement aided the writing of this thesis in several ways. I would like to thank SACEMA (South African Center for excellence in Epidemiological Modeling and Analysis) for funding my studies over the past two years. This thesis would not have been possible without the data provided by the HPP (HIV Pathogenesis Program) and CAPRISA (Centre for AIDS Program Research In South Africa). I would like to thank my employers, the MRC (Medical Research Council), for generously allowing me time to work on my thesis.

I would like to thank God for giving me the courage to pursue this daunting topic. It was not an easy road but I have enjoyed every moment. I would like to thank my parents, sister, brother, grandfather and aunt for always believing in me. Thank you to my fiancé Raneel for all his love, support and encouragement.

This thesis is dedicated to my late grandmother, Govindoo Chunglerian. The desire to make you proud spurred me on to complete this thesis.

Chapter 1

Introduction

Survival analysis is concerned with modeling time to event (e.g. death or disease onset) data where hazard rates or intensities or survival functions between groups are compared. This topic has been extensively researched. However, in practice the disease evolution or progression can be broken up into a finite number of intermediate states, which may offer a greater understanding and clarity of the evolution of disease than a pure survival analysis model would provide. A multistate model considers several discernable states of a process, a modeling process that allows one to compute transition probabilities, transition rates (hazards), mean sojourn time (mean time spent in each state) and to model the effect of covariates on transition rates. Treating a multistate model as several separate survival models would not suffice as this would not accommodate the process stage dependence. In spite of the many advantages of applying multistate models, these models are not often applied compared to classical survival analysis. Meira-Machado et al. (2009) state the reasons for this as daunting mathematical theory and lack of available software.

As a multistate process evolves over time, a history is naturally generated. This history contains information on previous states visited, times of entry into previous states and length of stay in states. The simplest, and most applied multistate model assumes that the transition to a future state is only dependent on the present state (Markov property) and time independent transition

rates. There are other multistate models which are less restrictive but more difficult to implement.

This thesis proposes to demystify the theory of multistate models through an application based approach. We present the key components of multistate models, relevant derivations, model diagnostics and techniques for modeling the effect of covariates on transition intensities. These concepts have been successfully applied to data on HIV (Alioum et al., 1998), depression (Marshall and Goldhamer, 1955), cancer (Schaubel et al., 1998) and hepatitis (Geng et al., 1998). Most of these studies, however, were performed in Europe and the United States. To my knowledge, this thesis is one of the first to study HIV disease progression in resource limited settings and the first in South Africa.

In 2008, it was estimated that there were an estimated 22.4 million adults and children living with HIV in Sub-Saharan Africa (UNAIDS, 2009). HIV/AIDS is therefore the greatest public health challenge facing South Africa and in particular the KwaZulu-Natal province . An understanding of HIV progression and factors that influence disease progression can have great value when understanding HIV pathogenesis and on the development of new treatment strategies (Sabin et al., 1998). With regards to HIV progression there are different ways in which one can define “progression”. One may examine the occurrence of clinical events which occur early in HIV infection or the values of biological markers as markers of disease progression. The most widely used marker of HIV progression is the CD4 lymphocyte count, which plays a crucial role in the immune system. When an individual loses CD4 cells he or she is more vulnerable to opportunistic infections and lymphomas.

Although substantial research has been done in the area of CD4 count modeling using linear mixed models, no attempt has been made to estimate the length of stay in different CD4 count intervals or to investigate probabilities of transitions to lower CD4 counts over time in South Africa. This thesis goes

even further, in that it seeks to estimate the effect of age, gender and baseline CD4 count on individual transition rates. The multistate approach is particularly powerful as it enables one to make predictions on the trajectory of an individual's disease progression through time.

The first endeavor to model the stages of HIV infection was by Longini et al. (1989) who applied a Markov model with five progressive stages: (1) infected but antibody-negative; (2) antibody-positive but asymptomatic; (3) pre-AIDS symptoms and/or abnormal haematologic indicator; (4) clinical AIDS and (5) death. This model was applied to data on 548 homosexual and bisexual men who were enrolled at the San Francisco City Clinic between 1978 and 1980. Longini et al. (1989) examined the mean sojourn times in each of the states and the length of the AIDS incubation period which is defined as the time from infection to clinical AIDS. This model offered valuable information regarding progression and was the start of Longini et al. (1989) contribution to the modeling of the natural history of HIV/AIDS but was lacking in that it did not take into account the effect of cofactors on rates of HIV progression. In 1991, Longini et al. (1991) modeled HIV progression with a different state structure that examined immune deterioration rather than clinical deterioration. Longini et al. (1991) applied an eight state progressive model which included cofactors on individual transition rates. In the 1796 HIV infected individuals from the US Army it was noted that rate of CD4 decline is faster in older age groups. Thus transition rates are age dependent. A limitation of this study was the inability to model long term behavior of CD4 count. This limitation, due to the fact that CD4 counts only started being taken in 1990, was explained by Bwayo et al. (1995) who addressed this limitation with their article in 1995. Bwayo et al. (1995) examined the decline in CD4 count in HIV positive female sex workers residing in Kenya.

In this thesis we propose a model with state structure based on four intervals of CD4 count and ARV initiation as an absorbing fifth state. The decision

to consider states based on CD4 counts is that apart from CD4 counts being the leading indicator of HIV progression, it is this marker that is used to determine when antiretroviral therapy should commence. This model is fitted to data from the Sinikithemba cohort, consisting of 451 HIV positive individuals residing in Durban, South Africa. Upon careful examination of CD4 count it became clear that reverse transitions need to be incorporated into the model. Bwayo et al. (1995) also noted erratic decline in CD4 counts in their data and were the first to include reverse transitions in CD4 count modeling. The decision to designate ARV status as the end point was based on the primary objective of this study, which is to determine rates of immune deterioration in *ARV naive* patients. The rate of CD4 count decline is examined and predictions of the trajectory of patients is made through the use of transition probability matrices. The effect of age, gender and baseline CD4 count on individual state transition rates is examined.

Chapter 2

An Introduction to Multistate Models

The key components associated with a multistate model according to Mullins (1996) are the states, observation times, actions, rewards, transitions and constraints of the model. These are briefly described below:

States

Each state in a Markov model represents a particular distinct situation and cannot overlap with another state. In statistical terms the above statement can be interpreted as *mutually exclusive* states. In biomedical applications the states might be based on clinical symptoms, biological markers (Alioum et al., 1998), some scale of the disease (Schaubel et al., 1998) or a non-fatal complication in the course of the illness (Mullins, 1996). Although this thesis is dedicated to biomedical applications, it must be stated that there are numerous applications of multistate models to other fields such as engineering and sociology. It is commonly assumed, for mathematical simplicity, that the number of states are finite. Each state is classified as either transient, absorbing or instantaneous.

Transient: A transient state is temporary. Hence, once it is visited it will be exited with certainty.

Absorbing: An absorbing state is a state from which there is no escape.

Instantaneous: An instantaneous state is one which will be exited immediately.

Observation times

These are the points in time at which the system is observed and data is collected. This system, in our case, refers to the disease process. Markov models can incorporate equally or unequally spaced stages. The four possible observation schemes discussed below are with reference to biomedical applications.

Fixed: Here patients are observed at fixed time intervals specified in advance by the investigator and the patient.

Random: The sampling times vary randomly, and are independent of the current state of disease. Such observation times arise in observational or open studies involving a cohort of individuals where no prior design to observe patients was put in place.

Doctor's care: It is important that severely ill patients are monitored more closely. Here the choice of the next sampling time is dependent on the state of the patient at the previous sampling time. The sampling times in this case are specified by the doctor.

Patient self selection: A patient may himself/herself decide to visit a doctor when in poor condition.

It is important to consider the reasons why observations are made at the given times. This could give information about the value of that observation. If a modeler were to fit a model ignoring the information in the observation times then the modeler's results could be biased. Jackson (2007) has shown that of the four observation schemes discussed above, the only informative observation scheme is patient self selection.

Actions and decisions

At each stage or observation time an action may be taken. Possible actions include surgery, a drug regimen or no action at all. When we are purely analyzing the natural history of a disease, as is done in this dissertation, the decision process and the rewards process discussed below are ignored.

Rewards

The actions described above are aimed at achieving some sort of reward such as a speedy recovery, higher quality of life or longer lifespan. A different state of disease may require a different sequence of actions in order to maximize the rewards.

State transitions

A change of state is called a *transition*. The transitions of a Markov model describe the likelihood of being in a particular state at some future time, given the present state. The Markov property (Definition 4.1) implies that a patient's progression rate to the next state is independent of their progression rate into the previous state. The parameters of interest in such a model are the inter-state *transition probabilities*, *transition rates* and distribution of *duration in state*.

States and state transitions can be effectively represented in a transition rate diagram. In these diagrams, states are represented by squares and arrows denote the allowed transitions to or from these states. Figure 2.1 is the transition rate diagram for the commonly used illness-death model.

The three states depicted in Figure 2.1 are disease free, disease and death. The bidirectional arrows connecting the disease and disease free state indicate

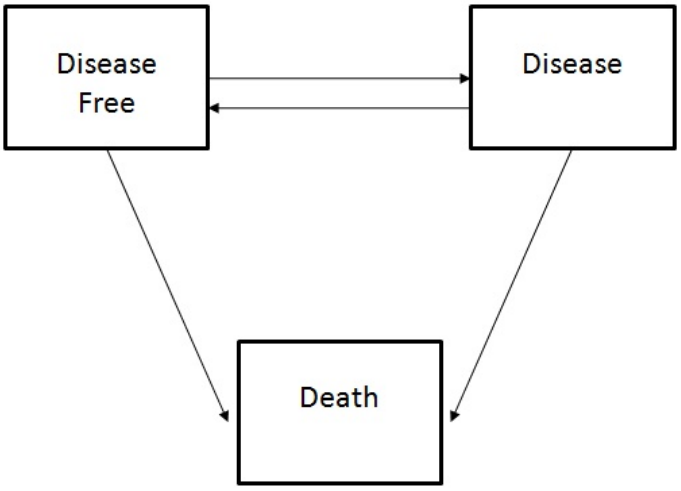

Figure 2.1: Illness-Death model

that a patient may progress to or recover from the disease state. Clearly death is an absorbing state.

In the homogeneous *continuous* time Markov context we define the transition probability, $p_{ij}(t)$, to be the conditional probability of entering state j at time t, given the system was in state i at time 0. This definition of $p_{ij}(t)$ implies that after leaving state i the process could have migrated to other states, say $j_1, j_2, \ldots, j_{k-1}$ before finally entering state j at time t. A detailed description of the properties of these transition probabilities and the transition probability matrix follows in Chapter 4.

Constraints

Constraints are introduced into the model to simplify it or to isolate typical behavior. In terminal disease modeling, a constraint is that state transitions are irreversible and sequential. That is, once a patient has progressed to a certain state they cannot go back to an earlier state. There are certain irreversible

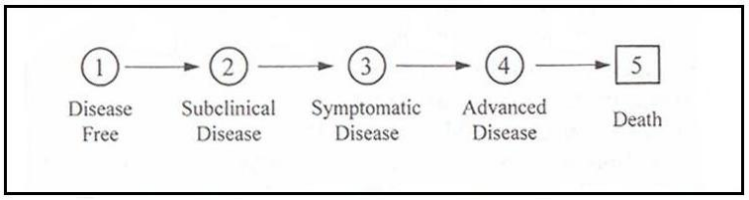

Figure 2.2: Irreversible multi-state terminal disease Markov model

and sequential states which have to be passed through in order to reach a more advanced state. In other words the disease states exhibit some natural hierarchy with order determined by the natural history of the disease. These states are called tunnel states because an individual passes through them in a set sequence, analogous to passing through a tunnel. The terminal disease model applied by Mullins (1996), which appears in Figure 2.2, illustrates the use of tunnel states.

It may be the case that, because of the timing of the observations, a patient appears to have skipped a state from one visit to the next. In such instances, it is assumed that the patient moved through the "skipped" state during the interval.

We consider two other components central to the understanding of multi-state models. These are:

1. The history generated by the multistate process
2. Survival analysis

The history generated by the multistate process

If one were to examine the state of an individual at a particular time t, it will become apparent that a particular path was traversed to get to this state. This "path" is referred to as the history of the process. The history contains information on previous states visited, length of stay in the current and previous states etc.

Survival analysis

Survival analysis deals with the analysis of time to event data. Here the event of interest may be death or disease onset (e.g. HIV infection). The important point to note is that survival analysis models a single event. Multistate models therefore are an extension of survival models. The beauty of dealing with only transitions to a single event is that many of the standard survival analysis techniques for prediction, model fitting and model analysis can be directly applied to multistate models. Chapter 3 will review key functions in survival analysis, nonparametric and parametric estimation of survival distribution and the modeling of survival data.

2.1 Preliminary Concepts

A multistate process is a stochastic process $(X(t), t \in T)$ with finite state space $S = \{1, 2, 3 \dots N\}$ where $T = [0, \tau]$ is the period of observation (Meira-Machado et al., 2009). As the process evolves over time a history $\mathcal{F}_{t-}$ consisting of the observation of the process over the interval $[0, t)$ is generated. $\mathcal{F}_{t-}$ is a $\sigma - algebra$ which can be explained as the history of the process just before time t. As mentioned previously this history contains information on previous states visited, length of stay in the current and previous states, time of transition into previous states and other related information. The two quantities which completely characterize a multistate process are transition probabilities and transition intensities.

Transition probabilities

Consider,

$$p_{ij}(s,t) = P(X(t) = j | X(s) = i, \mathcal{F}_{s-}) \tag{2.1}$$

for $i, j \in S$ and $s, t \in T$ where $p_{ij}(s,t)$ can be interpreted as the probability

of entering state j at time t, conditional on being in state i at time s and the history of the process prior to time s. The transition probabilities in Equation (2.1) form a transition probability matrix $\mathbf{P}(s,t)$.

$$\mathbf{P}(s,t) = \begin{pmatrix} p_{11}(s,t) & p_{12}(s,t) & \cdots & p_{1n}(s,t) \\ p_{21}(s,t) & p_{22}(s,t) & \cdots & p_{2n}(s,t) \\ \vdots & \vdots & \cdots & \vdots \\ p_{n1}(s,t) & p_{n2}(s,t) & \cdots & p_{nn}(s,t) \end{pmatrix}$$

Transition intensities

Let,

$$q_{ij}(t, \mathcal{F}_{t-}) = \lim_{\Delta t \to 0} \frac{p_{ij}(t, t+\Delta t)}{\Delta t} \tag{2.2}$$

The transition intensity can be interpreted as the instantaneous rate or hazard of making a transition from state i to state j at time t. The transition intensities in Equation (2.2) form a transition rate matrix $\mathbf{Q}$(t).

$$\mathbf{Q}(t) = \begin{pmatrix} q_{11}(t) & q_{12}(t) & \cdots & q_{1n}(t) \\ q_{21}(t) & q_{22}(t) & \cdots & q_{2n}(t) \\ \vdots & \vdots & \cdots & \vdots \\ q_{n1}(t) & q_{n2}(t) & \cdots & q_{nn}(t) \end{pmatrix}$$

Note that the transition probabilities and transition intensities both depend on the history $\mathcal{F}_{t-}$.

2.2 Types of Multistate Models

The more history on which the transition intensities and transition probabilities depend, the more complicated the model is to implement.
We will briefly introduce the various types of multistate models in this section, and explore them in more detail in subsequent chapters.

2.2.1 Markov Model

In a Markov model a transition to a future state is only dependent on the present state of the process as the history before this is irrelevant. Hence,

$$p_{ij}(s,t) = P(X(t) = j | X(s) = i) \tag{2.3}$$

This is an important departure from the model specified in Equation (2.1). The three types of Markov models are outlined below.

Time-homogeneous Markov model (THMM):
In a time-homogeneous Markov model transition intensities are assumed to be constant. Hence

$$q_{ij}(t, \mathcal{F}_{t-}) = q_{ij}$$

The THMM is the most widely applied multistate model, possibly owing to the well developed software for these models and less daunting theoretical framework. Chapter 4 explores the Markov assumption, time homogeneity assumption, properties of the transition probability matrix **P**(t), the transition rate matrix **Q**(t) and the computation of **P**(t).

Time dependent Markov model:
This model is used when there is an underlying reason for transition rates to change with time. For example, suppose we were examining the simple "illness-death" model. It is a known fact that rates of falling sick are higher in winter and recovery rates are lower. Hence, the transition rates vary with time of the season. It is also often the case that transition rates vary with age (time). This is easy to see in the following example.

Example of time dependence:
As patients get older, the risk of death due to illness increases and their chances of recovery decrease. The opposite is true for younger patients but this cannot be generalized to all diseases. In fact, for some childhood diseases the opposite occurs. That is the younger the patient the higher the risk of disease and death

and lesser the chance of recovery. Clearly, in this case transition rates would be a function of age.

Nonparametric Markov model: This approach can be thought of as a generalization of the Kaplan-Meier estimate outlined in Chapter 3. The non-parametric estimates of the transition probabilities are called the Aalen-Johansen Estimates and are presented in detail in Chapter 5.

2.2.2 Semi-Markov Model

In a Semi-Markov model the future evolution not only depends on the current state, but also on t_i, the entry time into the current state i. The form of the transition intensities which we model then changes to $q_{ij}(t, t - t_i)$ where $t - t_i$ denotes the duration in state i.

Although the semi-Markov model is more flexible than the time homogeneous Markov model there are two drawbacks to it use. Firstly, semi-Markov models contain many parameters which make them more difficult to fit. Secondly, we can only use these models if we know what distribution the sojourn or residence time in each state follows.

2.2.3 Non-Markov Model

Non-Markov models depend arbitrarily on the history of the process. These models were relatively unused and difficult to implement until recently when Meira-Machado et al. (2009) developed “Markov free” estimators of transition probabilities for an illness-death model.

Chapter 3

Time to Single Event Analysis

3.1 Introduction

Time to single event analysis or survival analysis can actually be considered as the simplest multistate model. This necessitates a thorough understanding of the topic before approaching more complicated models. As depicted in Figure 3.1, survival analysis can be represented as a multistate model with two states and one transition associated with a transition rate q_{12}. Questions which arise with respect to survival analysis are:

- Why do we require special techniques and models for time to event data?
- Why are the more standard statistical methods such as t-tests, linear regression and others not appropriate?

A response to the above questions is that survival data has many special features which require the assumptions of distributions other than the normal distribution. In addition, survival data frequently suffers from the prob-

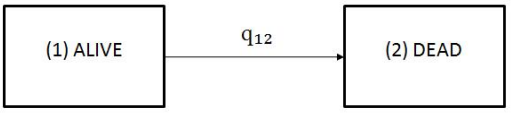

Figure 3.1: Survival as a multistate model

lem of censored observations hence the need for non-standard methods to deal with partially observed outcomes other than those appropriate to fully observed data.

3.2 The Special Features of Survival Data

Incomplete observation

Unlike blood pressure or weight, time to event is not measured instantaneously. In other words, one has to "wait" for the event to occur. Due to time and monetary constraints, studies cannot continue until every person has experienced the event. Hence some people may never experience the event during the study. People may also choose to withdraw from the study or may be lost to followup. The above are all examples of censoring. A censored observation is one whose value is incomplete due to factors observed or unobserved specific to each subject or group (Hougaard, 2002). An observation is left censored if the event occurred before observation began and right censored if the actual value lies to the right of the censored value. Interval censoring, which is of particular importance in the next chapter, occurs when the event of interest is known to have occurred between two time points. An important assumption made in the analysis of censored survival data is the assumption of *independent* censoring. This assumes that the actual survival time of an individual is independent of the mechanism that causes the survival time to be censored at c where $c < t$.

Distribution of survival times

Survival data are not symmetrically distributed and are strictly positive. In fact if we were to construct a histogram of survival times of similar individuals the histogram will exhibit a long tail to the right, meaning that most time to event data is right skewed.

Conditioning

Hougaard (2002) introduced this concept through an example which we have adapted below:
If we have information that the median lifetime of a South African female is 65 years and 70 days, can a woman celebrating her 65th birthday use this information to infer that she has only 70 days left to live? The answer to this question is obviously "no" since the death pattern of woman aged 0 to 65 years old is irrelevant to the 65 year old as she has already reached this age. Hence, a more accurate picture would be the conditional median where we condition on survival to 65 years. The above illustration warrants the form of the hazard function which will be discussed in the next section.

3.3 Important Functions in Survival Analysis

Let T be the time to event random variable and assume that each individual i has an event time t_i and a censoring time c_i, two new variables may now be introduced. $x_i = min(t_i, c_i)$ is the time which we actually observe and δ_i is an indicator of death which takes on a value of 1 if t_i is observed and 0 if c_i is observed. Putter et al. (2007) state that these events and censoring times of the individuals should be seen as a random sample $(X_1, C_1) \dots (X_n, C_n)$ from a survival distribution $X_i \sim F$ with $S(t) = P(T > t)$ and censoring distribution $C_i \sim G$. The basic assumption of independent censoring assumes that the actual survival time is independent of the censoring mechanism for right censored data. The survival function $S(t)$ and related functions are briefly discussed below.

Survival function

Let T denote the survival time, then the probability that an individual survives longer than t (before experiencing the event) is given by

$$S(t) = P(T > t)$$

The graph of $S(t)$ is called the survival curve and allows us to find the quartiles and compare the distribution of two or more groups. Section 3.4 covers methods for the estimation of $S(t)$ in the presence of right censored observations.

The hazard function

Lee and Wang (2003) define the hazard function as the probability of failure during a very small time interval assuming that the individual has survived to the beginning of that interval. Thus the hazard function is mathematically expressed or defined as

$$\alpha(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{3.1}$$

By the independence assumption the hazard of censored individuals is equal to the hazard of individuals who are still under observation. The hazard function can increase, decrease, remain the same or exhibit a combination of these features. Lee and Wang (2003) explore different shapes of hazards. The hazard function also plays a central role in regression modeling which will be discussed in Section 3.5.

The cumulative hazard function

The cumulative hazard function can take on any value between 0 and ∞ and represents the integrated hazard up to time t.

$$A(t) = \int_0^t \alpha(s) ds \tag{3.2}$$

The function $A(t)$ is also frequently referred to as the integrated hazard function.

The three functions described are mathematically related. First, from Equation (3.1) it is easy to state that

$$\alpha(t) = \frac{f(t)}{S(t)}$$

But $f(t) = \frac{d}{dt}[1 - S(t)]$
Hence

$$\alpha(t) = -\frac{S'(t)}{S(t)}$$

It also follows that

$$-\int_0^t \alpha(s)ds = \log S(t)$$

Therefore $A(t) = -\log S(t)$ or equivalently

$$S(t) = \exp[-A(t)] = \exp[-\int_0^t \alpha(s)ds]$$

3.4 Nonparametric Methods for the Estimation of Survival Functions

First we consider the estimation of $S(t)$. There exist two nonparametric methods for the estimation of $S(t)$. These are

1. The Kaplan-Meier or product limit estimator
2. The life table estimate

The Kaplan-Meier estimate is based on individual survival times whereas the lifetable estimate groups survival times into intervals. We will only discuss the Kaplan-Meier estimator for the purpose of this thesis.

Suppose that there are n individuals with observed survival times $t_1, t_2, \ldots, t_n$. If there are $m \leq n$ recorded event times then the ranked survival times are denoted as $t_{(1)}, t_{(2)} \ldots t_{(m)}$. Let n_j be the number of individuals at risk (still under followup and have not yet experienced the event) at time $t_{(j)}$ and d_j be the number of observed events at time $t_{(j)}$ for $j = 1, 2, \ldots, m$. The Kaplan-Meier estimator is defined as follows

$$\hat{S}(t) = \prod_{t:t_{(j)} \leq t} \left(1 - \frac{d_j}{n_j}\right) \tag{3.3}$$

Several approaches have been suggested to estimate the cumulative hazard function $A(t)$ given by (3.2).

The Nelson-Aalen estimator of $A(t)$ is given by

$$\hat{A}(t) = \sum_{t_{(j)} \leq t} \frac{d_j}{n_j} \tag{3.4}$$

The estimator of the survival function based on Equation (3.4) is

$$\hat{S}(t) = \exp[-\hat{A}(t)]$$

The slope of the cumulative hazard function gives us insight into the shape of the hazard function, which is useful for model identification.

3.5 Modeling Survival Data

Another distinguishing feature of survival analysis is that the hazard function is modeled directly, whereas in most other regression models the mean response or some function of the mean response is modeled as in generalized linear models (McCullagh and Nelder, 1989). Collett (2003) describes the two main objectives of modeling survival data as :

(1) To determine the effect that explanatory variables have on the hazard.

(2) To estimate the hazard function for a particular individual. Using the relationship between the survival function and the hazard function outlined in Section 3.3, this in turn will enable us to estimate survival probability or median survival time as a function of explanatory variables.

The most common regression type model for time to event data is the proportional hazards model (Cox, 1972).

3.5.1 The Cox Proportional Hazards Model

Definition 3.1. The proportional hazards property
This property states that the ratio of hazard rates for two different individuals with covariates $\mathbf{x}_1 = (x_{11}, x_{21}, \ldots, x_{p1})'$ and $\mathbf{x}_2 = (x_{12}, x_{22}, \ldots, x_{p2})'$ is a constant (i.e. independent of time).

In simple terms, the above definition means that the ratio of the risk of an event for two individuals is the same, regardless of how long they survive (Lee and Wang, 2003). The above property gives rise to the following form of the hazard function. Given a set of covariates in $\mathbf{x} = (x_1, x_2, \ldots, x_p)'$, the hazard function at time t is

$$\alpha(t \mid \mathbf{x}) = \alpha_0(t) g(\mathbf{x}) \tag{3.5}$$

In Equation (3.5) $\alpha_0(t)$ is referred to as the baseline hazard function and can be interpreted as the hazard function when all covariates are set to zero and $g(\mathbf{x})$ is a function of covariates only. The Cox proportional hazards model, due to Cox (1972), assumes that g(**x**) in Equation (3.5) is an exponential function of covariates. That is,

$$g(\mathbf{x}) = \exp\left(\sum_{j=1}^{p} b_j x_j\right) = \exp(\mathbf{b}'\mathbf{x}) \tag{3.6}$$

Therefore the hazard function in Equation (3.5) becomes :

$$\alpha(t \mid \mathbf{x}) = \alpha_0(t) \exp\left(\sum_{j=1}^{p} b_j x_j\right) = \alpha_0(t) \exp(\mathbf{b}'\mathbf{x}) \tag{3.7}$$

where $\mathbf{b} = (b_1, b_2, \ldots, b_p)'$ are the regression coefficients associated with the p covariates.

The hazard ratio for a subject with a set of predictors say $\mathbf{x}_1$, compared to a subject with a set of predictors $\mathbf{x}_2$ is

$$HR(\mathbf{x}_1, \mathbf{x}_2) = \frac{\alpha_0(t)\exp(\mathbf{b}'\mathbf{x}_1)}{\alpha_0(t)\exp(\mathbf{b}'\mathbf{x}_2)} \tag{3.8}$$

$$= \exp(\mathbf{b}'(\mathbf{x}_1 - \mathbf{x}_2) \tag{3.9}$$

3.5.2 The Partial Likelihood Function for Survival Times

This section, written with reference to Lee and Wang (2003), covers the partial likelihood method for estimation of parameters in the absence of tied observations. Suppose that k of the n individual survival times are uncensored and that $n-k$ are right censored. We let $t_{(1)} < t_{(2)} < \ldots < t_{(k)}$ and $\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \ldots, \mathbf{x}_{(k)}$ denote the ordered, distinct event times and covariates respectively. As defined previously the risk set at time $t_{(i)}$ here denoted by $R(t_{(i)})$ consists of all individuals still under followup and who have not yet experienced the event. For a particular observed event at time $t_{(i)}$, given $R(t_{(i)})$, the probability that the event is on the individual observed is :

$$\frac{\exp(\sum_{j=1}^{p} b_j x_{j(i)})}{\sum_{l \epsilon R(t_{(i)})} \exp(\sum_{j=1}^{p} b_j x_{jl})} = \frac{\exp(\mathbf{b}'\mathbf{x}_{(i)})}{\sum_{l \epsilon R(t_{(i)})} \exp(\mathbf{b}'\mathbf{x}_l)} \tag{3.10}$$

Since each observed event time contributes a factor as given above, the overall partial likelihood function is

$$\mathcal{L}(\mathbf{b}) = \prod_{i=1}^{k} \frac{\exp(\sum_{j=1}^{p} b_j x_{j(i)})}{\sum_{l \epsilon R(t_{(i)})} \exp(\sum_{j=1}^{p} b_j x_{jl})} = \prod_{i=1}^{k} \frac{\exp(\mathbf{b}'\mathbf{x}_{(i)})}{\sum_{l \epsilon R(t_{(i)})} \exp(\mathbf{b}'\mathbf{x}_l)} \tag{3.11}$$

Expression (3.11) is called a partial likelihood because, as is evident, the baseline hazard is left unspecified (Cox, 1972). The maximum partial likelihood estimate $\hat{\mathbf{b}} = (\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_p)'$ can be found by setting the derivative of the log of the expression in Equation (3.11) with respect to $\mathbf{b}$ equal to zero and solving the resulting simultaneous equations iteratively using numerical methods such as the Newton-Raphson procedure.

3.5.3 Non-proportional Hazards Model

The model described above is only relevant for time independent covariates. In practice, however, we may find that both time dependent and baseline covari-

ates contribute to the hazard at time t. If this happens the time independent hazard ratio assumption clearly fails to hold. The two types of time dependent variables are

(i) Time varying covariates resulting from repeated observations at different time points prior to the event or censoring.

(ii) Covariates whose values change according to a mathematical function of time.

The partial likelihood function accommodating time dependent covariates is

$$\mathcal{L}(\mathbf{b}) = \prod_{i=1}^{k} \frac{\exp(\sum_{j=1}^{p} b_j x_{j(i)}(t_{(i)}))}{\sum_{l \epsilon R(t_{(i)})} \exp(\sum_{j=1}^{p} b_j x_{jl}(t_{(i)}))} \tag{3.12}$$

$$= \prod_{i=1}^{k} \frac{\exp(\mathbf{b}'\mathbf{x}_{(i)}(t_{(i)}))}{\sum_{\ell \epsilon R(t_{(i)})} \exp(\mathbf{b}'\mathbf{x}_{\ell}(t_{(i)}))} \tag{3.13}$$

where k is the number of uncensored event times and
$\mathbf{x}_{\ell}(t_{(i)}) = (x_{1l}(t_{(i)}), x_{2l}(t_{(i)}), \ldots, x_{pl}(t_{(i)}))'$ denotes the covariates observed from person ℓ at time $t_{(i)}$. Clearly the likelihood $\mathbf{x}_{\ell}(t_{(i)})$ in Equation (3.13) is time dependent despite the proportional hazard structure at each observation time.

3.6 Application: Survival of HIV-TB co-infected patients under different forms of Treatment

This section examines the survival of HIV patients co-infected with TB who were enrolled in the CAPRISA SAPIT trial. The SAPIT trial was a three arm open label randomized controlled trial, the objective of which was to determine the optimal time to start ARVs in HIV-TB co-infected patients.
The three treatment arms were

Arm 1: ART to be initiated within 4 weeks of starting tuberculosis treatment.

Arm 2: ART to be initiated within 4 weeks of completing the intensive phase of tuberculosis treatment.

Arm 3: ART to be initiated within 4 weeks after completing tuberculosis treatment.

In the analysis which follows arm 1 and arm 2 combined are referred to as the "Integrated arm", and arm 3 as the "Sequential arm".

Between 28 June 2005 and 11 July 2008 HIV-infected, smear positive patients, 18 years or older were recruited. A total of 642 HIV-tuberculosis co-infected patients were enrolled; 429 in the integrated treatment arms and 213 in the sequential treatment arm (Abdool Karim et al., 2010). At baseline, patients in the integrated and sequential treatment arms were comparable in terms of age, CD4 count, viral load (VL), WHO stage and gender. These baseline characteristics are presented in Table 3.1. Difference between arms was assessed through the use of t-tests for continuous variables and Fishers exact test for categorical variables (gender and WHO Stage).

3.6.1 Description of the Population at Baseline

Table 3.1: Baseline characteristics

Characteristic	**Integrated Arm**	**Sequential Arm**	**P-value**
Mean Age (SD)	34.4 (8.38)	33.9 (8.18)	0.48
Gender (Proportion Males)	48.7%	52.1%	0.45
Mean CD4 (SD)	181(136.2)	167 (124.1)	0.22
Mean Log VL (SD)	5 (0.91)	5.12(0.74)	0.12
WHO Stage 4	4.9%	4.7%	1

3.6.2 Survival Analysis

By September 2008 there were a total of 52 deaths in the study. A statistical test was performed to compare the two arms to examine the effect of treatment arm on death. Table 3.2 shows that there were significantly higher deaths in the sequential treatment arm than in the integrated treatment arm (p-value $= 0.001$). It is of interest to examine the impact of treatment arm on time to death.

Table 3.2: Mortality between treatment arms

Treatment Arm	n	Number of Deaths	Proportion	P-value
Integrated	429	29	6.76	0.001
Sequential	213	23	10.80	

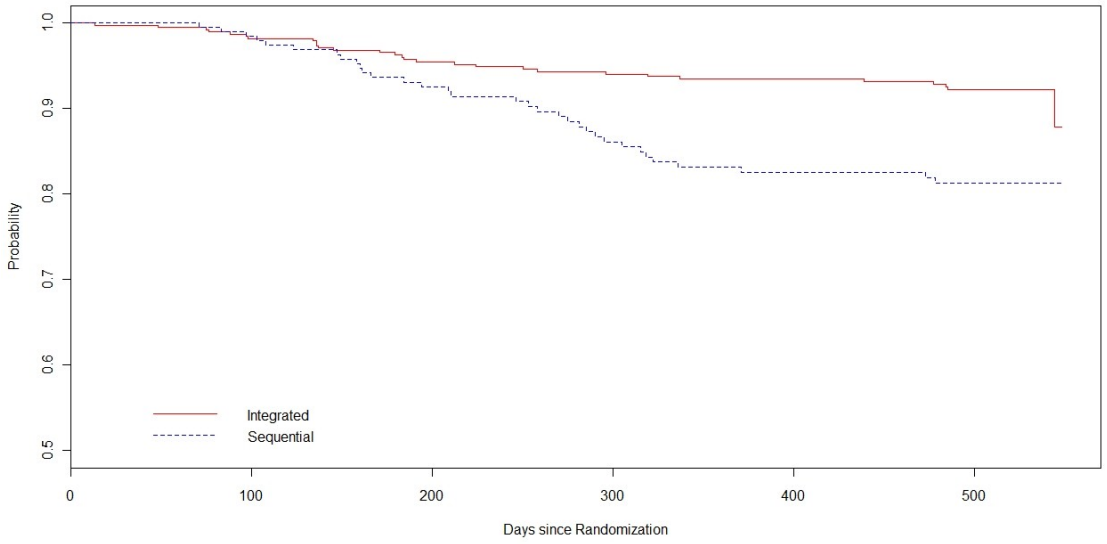

Figure 3.2: Kaplan-Meier survival curves

The Kaplan-Meier survival curves associated with the two treatment arms appears in Figure 3.2. Patients who did not experience a death were censored at the time of their last physical examination. From this curve we see a large number of the deaths in the sequential arm occurred between 60 and 240 days. This time corresponds to the continuation phase of TB therapy. Thus we see that treatment administered between the time of completion of TB therapy and ARV initiation has a significant impact on survival.

The Cox proportional hazards model was fitted using the R software which has a built in function “coxph” to fit the model. We were interested in examining the effect of treatment arm, baseline weight and CD4 count on time to death.

Based on previous literature we took the decision to include age and gender as they are well known to be confounders.

The specification of the model to account for covariate effects on the hazard of death is given below.

$$\begin{aligned} \alpha(t \mid \mathbf{x}) &= h_0(t)\exp(\mathbf{b}'\mathbf{x}) \\ &= h_0(t)\exp(b_1\text{gender} + b_2\text{age} + b_3\text{arm} \\ &\quad + b_4\text{weight} + b_5\text{CD4}) \end{aligned} \tag{3.14}$$

Weight and baseline CD4 count were transformed to categorical variables as follows

Weight

1 Weight < 60
2 Weight ≥ 60

Baseline CD4 count

1 CD4 Count $0 - 100$
2 CD4 Count $101 - 200$
3 CD4 Count $201 - 350$
4 CD4 Count $351 - 500$

The parameter estimates, hazard ratios, standard error of the coefficient, the z or Wald statistic and the p-value after applying the Cox proportional hazards model in R appear in Table 3.3. We see that the hazard of death for males is almost 1.7 times that of females. It is also clear that lower levels of CD4 count is associated with higher risk of death. This effect is particularly pronounced in patients with a baseline CD4 count less than 100, who have a 3.63 times greater hazard of death than patients with a CD4 count greater than 350.

Table 3.3: Results of the fitted Cox proportional hazards model

Variable	Coefficient	Hazard Ratio	Standard Error	z	p-value
Gender (Male)	0.5299	1.699	0.272	1.947	0.0510
Age	-0.0129	0.987	0.017	-0.761	0.4500
Arm (Sequential)	0.8179	2.266	0.256	3.192	0.0014
Baseline Weight < 60	0.2488	1.282	0.278	0.897	0.3700
Baseline CD4 (0-100)	1.2885	3.627	0.602	2.140	0.0320
Baseline CD4 (101-200)	0.6295	1.877	0.631	0.997	0.3200
Baseline CD4 (201-350)	-0.3379	0.713	0.732	-0.462	0.6400

As can be viewed from Table 3.3, the effect of Treatment Arm on time to death is significant (p-value < 0.05). The hazard in the sequential treatment arm is 2.27 times the hazard in the integrated treatment arm.

Using the stepwise regression procedure for which output appears in Appendix C, based on the AIC criterion, weight and age were excluded from the model as they were not significant predictors of time to death. Thus, the final model was refitted with both baseline weight and age excluded from the model. The results appear in Table 3.4.

Table 3.4: Results of the refitted Cox proportional hazards model

Variable	Coefficient	Hazard Ratio	Standard Error	z	p-value
Gender (Male)	0.478	1.613	0.261	1.829	0.06700
Baseline CD4 (0-100)	1.280	3.597	0.600	2.133	0.03300
Baseline CD4 (101-200)	0.630	1.878	0.630	1.001	0.32000
Baseline CD4 (201-350)	-0.361	0.697	0.731	-0.494	0.62000
Arm (Sequential)	0.850	2.340	0.255	3.335	0.00085

To determine whether the assumption of proportional hazards holds, we used two methods : One graphical and the other quantitative.

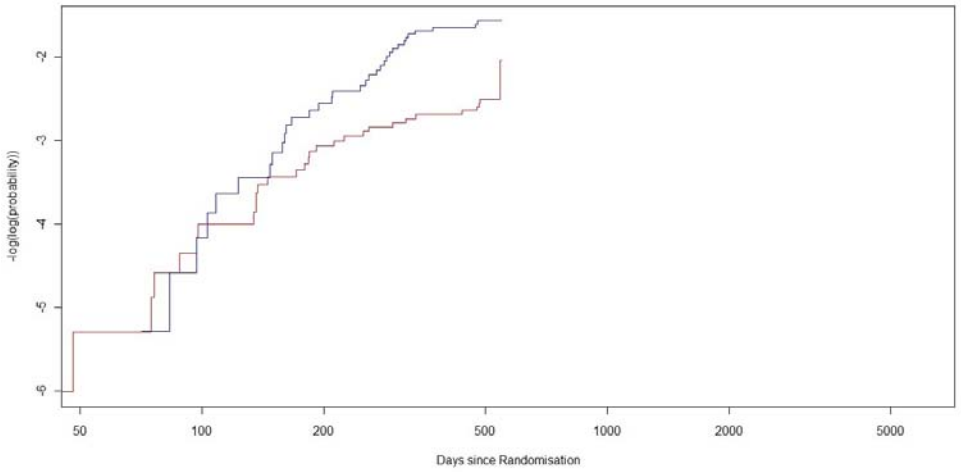

Figure 3.3: Testing the proportional hazards assumption

In light of Definition 3.1, one method to check proportional hazards is to fit a Cox proportional hazards model stratified on levels of some covariate, and plot

$$\log(-\log(\hat{S}_j(t;\tilde{\mathbf{x}}_j)))$$

where j represents the j^{th} stratum and $\tilde{\mathbf{x}}_j$ represents the vector of average values of other covariates for the j^{th} stratum (Lee and Wang, 2003). If the assumption of proportional hazards holds, the curves will be parallel.

Since the two curves in Figure 3.3 cross more than once it is difficult to conclude whether the assumption has been violated.

The formal test "coxzph" in R gives an estimate of the time-dependent coefficient $b(t)$ and tests its significance. If the proportional hazards assumption is true, $b(t)$ will be a constant and graphically a horizontal line. The test was applied to the current problem to test for time dependence of the regression coefficients, the result of which appears in Table 3.5.

As can be seen from Table 3.5 and p-values, there seems to be no evidence to suggest that the proportional hazards assumption is violated. Plots of scaled

Table 3.5: An investigation of the proportional hazards assumption

Variable	rho	chisq	p-value
Gender(Male)	-0.0684	0.294	0.587
Baseline CD4 (0-100)	-0.1473	1.342	0.247
Baseline CD4 (101-200)	-0.1434	1.290	0.256
Baseline CD4 (201-350)	-0.1924	2.303	0.129
Arm (Sequential)	0.1480	1.365	0.243
Global	NA	3.841	0.572

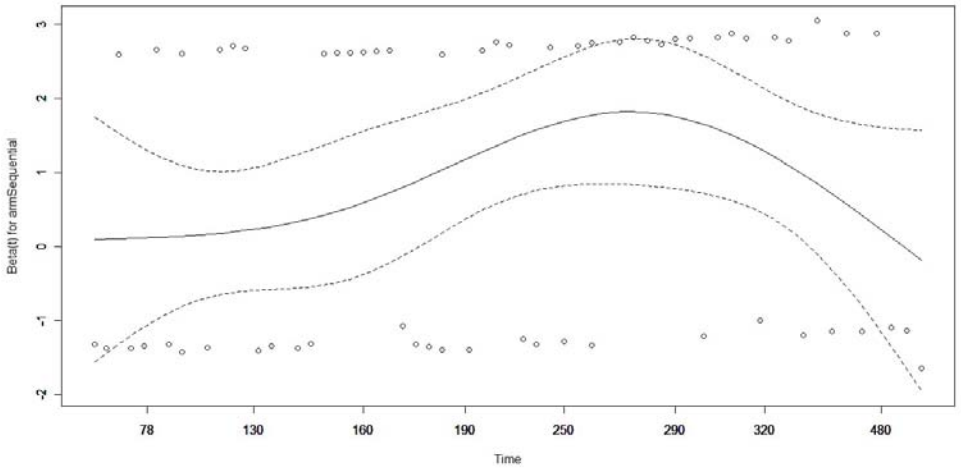

Figure 3.4: Time trend of the hazard ratio

Schoenfeld residuals against time for treatment arm appears in Figure 3.4. The solid line is a smoothing-spline fit to the plot, with the broken lines representing standard error bands around the fit. This plot reaffirms the finding in Table 3.5, that the assumption of proportional hazards appears to be supported for the covariate treatment arm. Thus, we are confident that the proportional hazards model assumed in our analysis is justified.

The SAPIT data provided a platform to apply the theoretical concepts of survival analysis which were introduced in previous sections. This was necessary

since survival analysis forms a foundation for multistate models, which is the main topic of this thesis. The ideas introduced in this chapter are now extended to multistate models in subsequent chapters with an application in Chapter 6.

Chapter 4

Time Homogeneous Markov Model

4.1 Introduction

In this chapter we briefly discuss the important and key properties of continuous time homogeneous Markov models. These basic concepts are key to understanding multistate models. We start with some important definitions and notations.

4.2 Definitions

Definition 4.1. The Markov property for processes in continuous time For a continuous time stochastic process $\{X(t), t \geq 0\}$ whose state space is S, we say it has the Markov property if

$$P(X(t) = j|X(s) = i, X(t_{n-1}) = i_{n-1}, ..., X(t_1) = i_1) = P(X(t) = j|X(s) = i)$$

where $0 \leq t_1 \leq ... \leq t_{n-1} \leq s \leq t$ is any nondecreasing sequence of $n + 1$ state occupation times and $i_1 \ldots, i_{n-1}, i, j \in S$.

In other words, Definition 4.1 states that the state of the process at time t depends only on the most recent state occupied prior to time t.

Definition 4.2. A continuous time stochastic process $\{X(t), t \geq 0\}$ is called a continuous time Markov process (CTMC) if it has the Markov property

Definition 4.3. Time homogeneity in a continuous time Markov chain A CTMC is said to be time homogeneous if for any $s \leq t$ and any states $i, j \in S$

$$P(X(t) = j|X(s) = i) = P(X(t-s) = j|X(0) = i)$$

That is, dependence on time is only through the length of time elapsed between events.

The time homogeneous property means that whenever state i is entered, the way the process evolves is equivalent to having started in state i at time 0.

4.3 The Sojourn Time

When the process enters state i, the time it spends there before moving to another state is called the holding time in state i or the *sojourn time*. The sojourn time is of great interest in disease modeling as it gives us an indication of how rapidly the disease is progressing. Longer sojourn times in a disease state mean a slow progressing disease and shorter sojourn times mean a rapidly progressing disease. With the assumption of time homogeneity, it should be evident that the sojourn time in state i would be the same every time state i is entered. Hence we can speak of a holding time or sojourn time *distribution*. An immediate question that arises is : What distribution does the holding time follow? The answer to this question is presented in the following theorem which has been adapted from Random Processes, Statistics (2007) notes.

Theorem 4.1. *For a time homogeneous continuous time Markov chain, T_i (the sojourn time in state i) is exponentially distributed*

Proof:

The proof is based on the memoryless property which is unique for the exponential distribution. By time homogeneity we assume that the process starts in state i. Then,

$$P(T_i > s+t|T_i > s) = P(X(u) = i \text{ for } 0 \leq u \leq s+t|X(u) \text{ for } 0 \leq u \leq s) \quad (4.1)$$

$$= P(X(u) = i \text{ for } s < u \leq s+t|X(u) = i \text{ for } 0 \leq u \leq s) \quad (4.2)$$

$$= P(X(u) = i \text{ for } s < u \leq s+t|X(s) = i) \quad (4.3)$$

$$= P(X(u) = i \text{ for } 0 < u \leq t|X(0) = i) \quad (4.4)$$

$$= P(T_i > t)$$

This is exactly the memoryless property which is unique for an exponentially distributed random variable, therefore T_i must be exponentially distributed.

Step (4.1) follows from the simple observation that, for $s \geq 0$ the event $\{T_i > s\}$ is equivalent to the event $\{X(u) = i \text{ for } 0 \leq u \leq s\}$. Similarly, for $s, t \geq 0$ the event $\{T_i > s+t\}$ is equivalent to the event $\{X(u) = i \text{ for } 0 \leq u \leq s+t\}$. Step (4.2) follows from the fact that $P(A \cap B|A) = P(B|A)$ where we let $A = \{X(u) = i \text{ for } 0 \leq u \leq s\}$ and $B = \{X(u) = i \text{ for } s < u \leq s+t\}$. Step (4.3) follows from the Markov property in Definition 4.1 and step (4.4) follows from time homogeneity defined in Definition 4.3.

Thus T_i is exponentially distributed with the corresponding state i mean sojourn time given by μ_i. This also means that $var(T_i) = \mu_i^2$ which implies that the variance is not independent of the mean. This is important because modeling such variables requires special approaches such as the generalized linear model (GLM) methodology (McCullagh and Nelder, 1989).

4.4 The Transition Rate Matrix

We have just established that in a homogeneous CTMC, for each state i, the amount of time spent in that state in a given visit is an exponentially distributed random variable. We define the parameter of this exponentially distributed random variable to be q_i. Here q_i is referred to as the transition intensity, and by the time homogeneous assumption of the Markov process, this parameter can be regarded as a constant. We let q_{ij} be the rate of going from state i to j. Then we define $\mathbf{Q}$ to be the transition rate matrix or infinitesimal transition generator, with elements

$$\mathbf{Q} = \begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1n} \\ q_{21} & q_{22} & \cdots & q_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ q_{n1} & q_{n2} & \cdots & q_{nn} \end{pmatrix}$$

for a continuous time Markov process with n states.

In the case where the state i is absorbing, that is once state i is entered it will never be exited, $q_i = 0$. A state i is called instantaneous if $q_i = \infty$, since once it is entered it is exited immediately. Throughout this thesis, and in the examples to follow, we assume that the CTMC has no instantaneous states, that is $0 < q_i < \infty$ for all $i \in S$.

Now we define important properties of the transition rate matrix, before proceeding to a description of the transition probability matrix.

(i) $\sum_{j \in S} q_{ij} = 0$ for all $i \in S$

(ii) $q_i = \sum_{j \neq i} q_{ij}$

(iii) $q_{ii} = -\sum_{j \neq i} q_{ij} = -q_i$ for all $i \in S$.

Using (ii) and (iii) above, we are now able to modify the $\mathbf{Q}$ matrix defined pre-

viously as

$$\mathbf{Q} = \begin{pmatrix} -q_1 & q_{12} & \cdots & q_{1n} \\ q_{21} & -q_2 & \cdots & q_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ q_{n1} & q_{n2} & \cdots & -q_n \end{pmatrix}$$

To fit a multistate model we need to estimate this transition intensity matrix $\mathbf{Q}$. Fortunately, this is relatively simple to do in R, as will be demonstrated in Chapter 6 provided that the right data for the observed process is available.

As we shall see in the following section, we can use the estimated $\mathbf{Q}$ matrix ($\hat{\mathbf{Q}}$) to find the transition probability matrix described in the next section.

4.5 The Transition Probability Matrix

As defined in Chapter 2, $p_{ij}(t)$ is the conditional probability of entering state j at time t, given the system was in state i at time 0. So,

$$p_{ij}(t) = P(X(t) = j | X(0) = i) \text{ for all } i, j \in S \text{ and } t \geq 0$$

The quantities $p_{ij}(t)$ are called the transition probabilities, and form a matrix $\mathbf{P}(t)$ called the transition probability matrix. Hence for a CTMC with n possible states we can define

$$\mathbf{P}(t) = \begin{pmatrix} p_{11}(t) & p_{12}(t) & \cdots & p_{1n}(t) \\ p_{21}(t) & p_{22}(t) & \cdots & p_{2n}(t) \\ \vdots & \vdots & \cdots & \vdots \\ p_{n1}(t) & p_{n2}(t) & \cdots & p_{nn}(t) \end{pmatrix}$$

A detailed description of the properties of this matrix and its individual elements follows in the next section. We are now able to describe the behavior of a CTMC. Suppose that once a system enters state i at time t it remains there for an exponentially distributed period of time with parameter q_i, where $q_i \geq 0$, and then jumps to state j ($j \neq i$). It will be noted that despite the assumption of time homogeneity, the state specific sojourn times are assumed to have separate exponential distribution rates thereby allowing state heterogeneity.

Theorem 4.2. *A continuous time Markov chain is completely characterized by its initial distribution and its transition probability matrix* $\boldsymbol{P}(t)$.

Based on the importance placed on $\mathbf{P}(t)$ in the above theorem, it is imperative that we have a thorough knowledge of the transition probability matrix $\mathbf{P}(t)$ and its properties.

4.6 The Properties of $\mathrm{P}(t)$

The theory of this section, unless otherwise stated, has been adapted from Kulkarni (1995).

Theorem 4.3. *The transition probability matrix* $\boldsymbol{P}(t)$ *with elements* $\{p_{ij}(t)\}$ *has the following properties.*

(i) $p_{ij}(t) \geq 0$ *for all* $i, j \in S$ *and* $t \geq 0$.

(ii) $\sum_{j \in S} p_{ij}(t) = 1$ *for all* $i \in S$ *and* $t \geq 0$.

(iii) $p_{ij}(t+s) = \sum_{k \in S} p_{ik}(t) p_{kj}(s)$ *for all* $t \geq 0, s \geq 0$ *and* $i, j \in S$.

Proof

(i) $p_{ij}(t) = P(X(t) = j | X(0) = i)$. These are conditional probabilities, so clearly $p_{ij}(t) \geq 0$.

(ii) To prove this part we let S_n be the time at which the n^{th} jump takes place. Now, define $X_n = X(S_n^+)$ to be the state of the system immediately after the n^{th} jump takes place. We can also define $N(t)$ to be the number of transitions up to

and including time t. Then

$$\sum_{j\in S} p_{ij}(t) = \sum_{j\in S} P(X(t) = j|X(0) = i)$$

$$= \sum_{j\in S} \sum_{n=0}^{\infty} P(X(t) = j|X(0) = i,\ N(t) = n)P(N(t) = n|X(0) = i)$$

$$= \sum_{j\in S} \sum_{n=0}^{\infty} P(X_n = j|X(0) = i,\ N(t) = n)P(N(t) = n|X(0) = i)$$

$$= \sum_{n=0}^{\infty} \sum_{j\in S} P(X_n = j|X(0) = i,\ N(t) = n)P(N(t) = n|X(0) = i)$$

$$= \sum_{n=0}^{\infty} P(N(t) = n|X(0) = i) \tag{4.5}$$

$$= P(N(t) < \infty|X(0) = i)$$

$$= 1 \tag{4.6}$$

Step (4.5) follows because we have $P(X_n \in S) = 1$ and hence

$$\sum_{j\in S} P(X_n = j|X_0 = i,\ N(t) = n) = 1$$

Step (4.6) follows because the CTMC is assumed to undergo a finite number of transitions in a finite amount of time.

(iii) To prove this equation, commonly known as the Chapman-Kolmogorov equation for CTMCs, we condition on $X(t)$ and also use the time homogeneity

property to get

$$p_{ij}(t+s) = P(X(t+s) = j|X(0) = i)$$

$$= \sum_{k\in S} P(X(t+s) = j|X(t) = k,\ X(0) = i)P(X(t) = k|X(0) = i)$$

$$= \sum_{k\in S} P(X(t+s) = j|X(t) = k)P(X(t) = k|X(0) = i) \tag{4.7}$$

$$= \sum_{k\in S} P(X(s) = j|X(0) = k)P(X(t) = k|X(0) = i) \tag{4.8}$$

$$= \sum_{k\in S} p_{ik}(t)p_{kj}(s) \tag{4.9}$$

Step (4.7) follows from the Markov property

Step (4.8) follows from time homogeneity

In matrix form (iii) can be written as $\mathbf{P}(t+s) = \mathbf{P}(t)\mathbf{P}(s)$.

Another property of the transition probabilities $p_{ij}(t)$ that is of interest in the study of CTMCs, is its behavior near $t = 0$. In order to achieve this goal we need to derive the differential equation for $p_{ij}(t)$. First we consider a number of preliminary concepts.

Theorem 4.4. *Let $\boldsymbol{P}(t)$ be a transition matrix of a CTMC. Let $\boldsymbol{Q} = [q_{ij}]$ be its generator matrix. Then,*

(i) $\lim\limits_{t\to 0} \dfrac{p_{ii}(t) - p_{ii}(0)}{t} = \lim\limits_{t\to 0} \dfrac{p_{ii}(t) - 1}{t} = q_{ii} = -q_i$*, where* $i \in S$

(ii) $\lim\limits_{t\to 0} \dfrac{p_{ij}(t) - p_{ij}(0)}{t} = \lim\limits_{t\to 0} \dfrac{p_{ij}(t) - 0}{t} = q_{ij}$*, where* $i \neq j$ *and* $i, j \in S$

Proof: We define $X_n = X(S_n^+)$ to be the state of the system immediately after the n^{th} jump, and $Y_n = S_n - S_{n-1}$ where S_n denotes the time of the n^{th} jump. Note that X_n is not a time variable but a state variable while S_n is a time

variable. Recall that T_i is the i^{th} holding or sojourn time. Let $\{X(t),\ t \geq 0\}$ be a CTMC and $N(t)$ be the number of transitions up to t. We have

$$P(N(t) = 0|X(0) = i) = P(Y_i > t|X(0) = i)$$

$$= e^{-q_i t}$$

$$= 1 - q_i t + o(t)$$

$$= 1 + q_{ii}(t) + o(t) \tag{4.10}$$

Also,

$$P(N(t) \geq 2|X(0) = i) = P(Y_1 + Y_2 \leq t|X(0) = i)$$

$$= \sum_{j \in S} P(Y_1 + Y_2 \leq t|X(0) = i,\ X_i = j)p_{ij} \tag{4.11}$$

Now, $X_0 = i$ and $X_1 = j$ imply that $Y_1 \sim \exp(q_i)$ and $Y_2 \sim \exp(q_j)$ are independent random variables. Hence,

$$P(Y_1 + Y_2 \leq t|X_0 = i,\ X_1 = j) = 1 - \frac{q_j}{q_j - q_i}\ e^{-q_i t} + \frac{q_i}{q_j - q_i}\ e^{-q_j t}$$

$$= 1 - \frac{q_j}{q_j - q_i}(1 - q_i t + o(t)) + \frac{q_i}{q_j - q_i}(1 - q_j t + o(t))$$

$$= o(h) \tag{4.12}$$

Now, substituting (4.12) into Equation (4.11) we get,

$$P(N(t) \geq 2|X(0) = i) = \sum_{j \in S} p_{ij} o(t) \tag{4.13}$$

Since the right hand side is a probability, it is bounded. Hence we get

$$P(N(t) \geq 2|X(0) = i) = o(t) \tag{4.14}$$

Now

$$
\begin{aligned}
p_{ii}(t) &= P(X(t) = i | X(0) = i) \\
&= P(X(t) = i | X(0) = i,\ N(t) = 0) P(N(t) = 0 | X(0) = i) \\
&+ P(X(t) = i | X(0) = i,\ N(t) = 1) P(N(t) = 1 | X(0) = i) \\
&+ P(X(t) = i | X(0) = i,\ N(t) \geq 2) P(N(t) \geq 2 | X(0) = i) \\
&= 1(1 + q_{ii}t + o(t)) + 0 \cdot P(N(t) = 1 | X(0) = i) \\
&+ P(X(t) = i | X(0) = i,\ N(t) \geq 2) o(t) \\
&= 1 + q_{ii}t + o(t) \qquad (4.15)
\end{aligned}
$$

Hence we have

$$
\begin{aligned}
\frac{p_{ii}(t) - 1}{t} &= q_{ii} + \frac{o(t)}{t} \\
\lim_{t \to 0} \frac{p_{ii}(t) - 1}{t} &= q_{ii} \quad ,
\end{aligned}
$$

which completes the proof.

(ii) This part is similarly proved, by showing that

$$
\begin{aligned}
p_{ij}(t) &= q_{ij}t + o(t) \qquad (4.16) \\
\frac{p_{ij}(t)}{t} &= q_{ij} + \frac{o(t)}{t} \\
\lim_{t \to 0} \frac{p_{ij}(t)}{t} &= q_{ij}
\end{aligned}
$$

Corollary 4.1. *$\boldsymbol{P}(t)$ is continuous at $t = 0$ and $\boldsymbol{P}(0) = \boldsymbol{I}$.*

Proof: The results from Theorem 4.4, along with the fact that $q_i < \infty$, imply that

$$\lim_{t\to 0} p_{ii}(t) = 1 = p_{ii}(0)$$

$$\lim_{t\to 0} p_{ij}(t) = 0 = p_{ij}(0) \quad \text{for } i \neq j$$

Hence $\lim_{t\to 0} \mathbf{P}(t) = \mathbf{P}(0) = \mathbf{I}$.

Corollary 4.2. *$\boldsymbol{P}(t)$ is continuous at all values of $t \geq 0$.*

Proof: From Theorem 4.3 part (iii), we know that $\mathbf{P}(t+h) = \mathbf{P}(t)\mathbf{P}(h)$. Letting $h \to 0$ we get

$$\begin{aligned}
\lim_{h\to 0} \mathbf{P}(t+h) &= \lim_{h\to 0} \mathbf{P}(t)\mathbf{P}(h) \\
&= \mathbf{P}(t) \lim_{h\to 0} \mathbf{P}(h) \\
&= \mathbf{P}(t)\mathbf{I} \\
&= \mathbf{P}(t)
\end{aligned}$$

Now that we have established continuity of $\mathbf{P}(t)$, we can establish differentiability of $\mathbf{P}(t)$ and its relationship with the rate matrix $\mathbf{Q}$.

Theorem 4.5. *Let $\boldsymbol{P}(t)$ be a transition matrix of a CTMC with rate matrix $\boldsymbol{Q}$. Then $\boldsymbol{P}(t)$ is differentiable and satisfies,*

$$\frac{d}{dt}\boldsymbol{P}(t) = \left[\frac{d}{dt} p_{ij}(t)\right] = \boldsymbol{P}(t)\boldsymbol{Q}$$

This is commonly known as Kolmogorov's forward equation (in matrix form).

Proof: We have

$$p_{ij}(t+h) = P(X(t+h) = j|X(0) = i)$$

$$= \sum_{k\in S} P(X(t+h) = j|X(t) = k,\ X(0) = i)P(X(t) = k|X(0) = i)$$

$$= \sum_{k\in S} P(X(t+h) = j|X(t) = k)P(X(t) = k|X(0) = i) \tag{4.17}$$

$$= \sum_{k\in S} P(X(h) = j|X(0) = k)p_{ik}(t) \tag{4.18}$$

$$= \sum_{\substack{k\in S\\ k\neq j}} p_{ik}(t)(q_{kj}h + o(h)) + p_{ij}(t)(1 + q_{jj}h + o(h)) \tag{4.19}$$

$$= \sum_{k\in S} p_{ik}(t)q_{kj}h + \sum_{k\in S} p_{ik}(t)o(h) + p_{ij}(t)$$

Hence

$$\begin{aligned} p_{ij}(t+h) - p_{ij}(t) &= \sum_{k\in S} p_{ik}(t)q_{kj}h + o(h) \\ \frac{p_{ij}(t+h) - p_{ij}(t)}{h} &= \sum_{k\in S} p_{ik}(t)q_{kj} + \frac{o(h)}{h} \\ \lim_{h\to 0} \frac{p_{ij}(t+h) - p_{ij}(t)}{h} &= \sum_{k\in S} p_{ik}(t)q_{kj} \\ \frac{d}{dt}\, p_{ij}(t) &= \sum_{k\in S} p_{ik}(t)q_{kj} \end{aligned}$$

In matrix form, the above result leads to

$$\frac{d}{dt}\mathbf{P}(t) = \mathbf{P}(t)\mathbf{Q} \tag{4.20}$$

Step (4.17) follows from the Markov property (Definition 4.1) and step (4.18)

follows from time homogeneity (Definition 4.2). The expression (4.19) follows from (4.15) and (4.16).

To obtain Kolmogorov's backward equation, we condition on $X(h)$ instead of $X(t)$ to arrive at

$$\frac{d}{dt}\mathbf{P}(t) = \mathbf{Q}\mathbf{P}(t)$$

A question which often arises is which equation should we use. The forward or backward equation? If we want to obtain the conditional distribution of $X(t)$ we should use the forward equation. However, if we want to find the probability that $X(t)$ is in a given state j for various initial states, we should solve backward equations.

4.7 Computation of $\mathrm{P}(t)$

For both the forward and backward equations in Theorem 4.3, we have the same initial boundary condition.

$$\mathbf{P}(0) = \mathbf{I}$$

For a continuous time Markov process with n finite states, $\mathbf{I}$ is the $n \times n$ identity matrix. This boundary condition follows since

$$\begin{aligned} p_{ii}(0) &= P(X(0) = i | X(0) = i) = 1 \\ p_{ij}(0) &= P(X(0) = j | X(0) = i) = 0 \end{aligned}$$

Although the backward and forward equations are different, they do have the same solution. The solution to the matrix differential Equation (4.20) is given by

$$\mathbf{P}(t) = e^{t\mathbf{Q}} = \sum_{n=0}^{\infty} \frac{(t\mathbf{Q})^n}{n!}$$

$$= \mathbf{I} + t\mathbf{Q} + \frac{(t\mathbf{Q})^2}{2!} + \frac{(t\mathbf{Q})^3}{3!} + \cdots \tag{4.21}$$

To see whether the above solution satisfies the backward equation defined in Theorem 4.3, we consider

$$\mathbf{P}'(t) = \frac{d}{dt}\left[\mathbf{I} + t\mathbf{Q} + \frac{(t\mathbf{Q})^2}{2!} + \frac{(t\mathbf{Q})^3}{3!} + \cdots\right]$$

$$= \mathbf{Q} + t\mathbf{Q}^2 + \frac{t^2}{2!} + \frac{t^3}{3!}\mathbf{Q}^4 + \cdots \tag{4.22}$$

$$= \mathbf{Q}\left[\mathbf{I} + t\mathbf{Q} + \frac{t^2\mathbf{Q}^2}{2!} + \frac{t^3\mathbf{Q}^3}{3!} + \cdots\right]$$

$$= \mathbf{Q}\mathbf{P}(t)$$

Similarly, to show that $\mathbf{P}(t) = e^{t\mathbf{Q}}$ satisfies the forward equation we can write (4.22) as

$$\begin{aligned}\mathbf{P}'(t) &= \left[\mathbf{I} + t\mathbf{Q} + \frac{t^2\mathbf{Q}^2}{2!} + \frac{t^3\mathbf{Q}^3}{3!} + \cdots\right]\mathbf{Q} \\ &= \mathbf{P}(t)\mathbf{Q}\end{aligned}$$

Now that we have established that both the backward and forward equations satisfy $\mathbf{P}(t) = e^{t\mathbf{Q}}$, the next step is the computation of $e^{t\mathbf{Q}}$. One possible method is to use the Taylor expansion of $e^{t\mathbf{Q}}$ in (4.21). However, Kulkarni (1995) has shown that this method is numerically unstable. He presents an alternative method which is simpler to implement on a computer. This method works when all eigenvalues of $\mathbf{Q}$ are distinct. An appendix on eigenvalue and eigenvector expansion appears in Appendix A.

Let $\mu_1, \mu_2, \ldots, \mu_N$ be the N distinct eigenvalues of $\mathbf{Q}$. Then it is possible to write the matrix decomposition

$$\mathbf{Q} = \mathbf{A}\mathbf{D}\mathbf{A}^{-1} \tag{4.23}$$

where $\mathbf{D} = \mathrm{diag}(\mu_1, \mu_2, \ldots, \mu_N)$ and $\mathbf{A}$ is a matrix whose i^{th} column is equal to the right eigenvector corresponding to μ_i. Then the positive powers of $\mathbf{Q}$ can easily be generated as

$$\mathbf{Q}^n = \mathbf{A}\mathbf{D}^n\mathbf{A}^{-1}$$

for $n = 1, 2, 3, \ldots$ hence,

$$\exp(\mathbf{Q}t) = \mathbf{A}\mathbf{A}^{-1} + \sum_{n=1}^{\infty} \mathbf{A}\, \frac{\mathbf{D}^n t^n}{n!}\, \mathbf{A}^{-1}$$

$$= \mathbf{A}\left[\mathbf{I} + \sum \frac{\mathbf{D}^n t^n}{n!}\right] \mathbf{A}^{-1}$$

$$= \mathbf{A}\exp(\mathbf{D}t)\mathbf{A}^{-1} \tag{4.24}$$

But,

$$\mathbf{D} = \mathrm{diag}(\mu_1, \mu_2, \ldots, \mu_N)$$

Hence,

$$\exp(\mathbf{D}t) = \mathrm{diag}(e^{\mu_1 t}, e^{\mu_2 t}, \ldots, e^{\mu_N t})$$

Substituting the above expression into (4.24), we are now able to compute $\exp(\mathbf{Q}t)$ easily. The following example, taken from Elston et al. (2002), uses the above theory to compute the transition probability matrix $\mathbf{P}(t)$ of a three state model as an illustration.

Example 4.1

This example tracks the flow of substances in the body. The states or compartments of the model refer to containers such as organs or the blood stream itself. The matrix $\mathbf{Q}$ corresponding to the three state model is given by

$$\mathbf{Q} = \begin{pmatrix} q_{11} & q_{12} & q_{13} \\ q_{21} & q_{22} & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Recall from Section 4.4 that the structure of the matrix $\mathbf{Q}$ means $q_{11} = -(q_{12} + q_{13})$ and $q_{22} = -q_{21}$. State 1 refers to the bloodstream, state 2 refers to the liver while state 3 refers to the bladder, from which the pollutant or toxin will be expelled, as shown in Figure 4.1. The drug will reside in the bloodstream for an exponentially distributed period of time, then move either to the bladder or to the liver. The drug will similarly stay in the liver for an exponentially

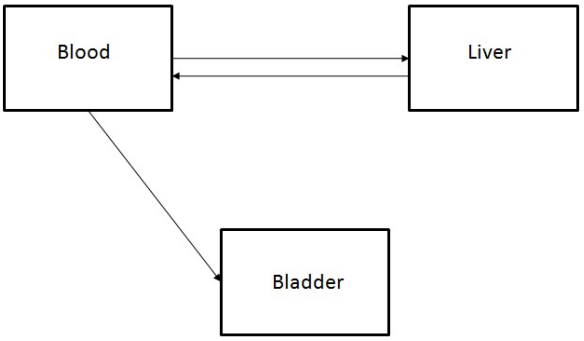

Figure 4.1: The transition rate diagram for toxins

distributed period, then move back to the bloodstream. Finally any drug that reaches the bladder will be removed from the system, so this represents an absorbing state. Clearly, the bloodstream state and the liver state are transient. Note that by virtue of having different exponential distribution rates for each compartment in the model, we account for state to state heterogeneity, that is not treating the process as one homogeneous compartment. The model allows for structure in the system.

Consequently, it is of interest to calculate the total amount of time that the toxin will spend in the liver where damage can occur. As an illustration suppose $\mathbf{Q}$ was found to be

$$\mathbf{Q} = \begin{pmatrix} -2 & 1 & 1 \\ 1 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Before we can compute the transition probability matrix, we need to find the eigenvalues and eigenvectors of the matrix $\mathbf{Q}$. For such a small matrix these can be found by hand (Appendix B) as well as via the R command "eigen". The eigenvalues and eigenvectors hence the matrix $\mathbf{A}$ corresponding to $\mathbf{Q}$ are given below as

$$(\mu_1, \mu_2, \mu_3) = (-2.618034, -0.381966, 0.0000),$$

$$\mathbf{A} = \begin{pmatrix} -0.8506508 & -0.5257311 & 0.5773503 \\ 0.5257311 & -0.8506508 & 0.5773503 \\ 0.0000000 & 0.0000000 & 0.5773503 \end{pmatrix}$$

If we assume that there is a bolus injection of pollutant at time 0 into the blood stream, then the transition probabilities for a single particle can be found from the Kolmogorov forward equations. Given the eigenvalue of $\mathbf{Q}$ we can write expressions for $p_{ij}(t)$ for $j = 1, 2, 3$ as

$$\begin{aligned} p_{11}(t) &= 0.276\exp(-0.382t) + 0.724\exp(-2.618t) \\ p_{12}(t) &= 0.448\exp(-0.382t) - 0.448\exp(-2.618t) \\ p_{13}(t) &= -0.724\exp(-0.382t) - 0.276\exp(-2.618t) + 1 \end{aligned}$$

Now, more specifically, we can calculate the probability of the substance remaining in or leaving the bloodstream after five units of time as

$$\begin{aligned} p_{11}(5) &= 0.04087 \\ p_{12}(5) &= 0.06633 \\ p_{13}(5) &= 0.89279 \end{aligned}$$

Through examination of the form of the transition probabilities we see that as time increases the pollution concentration decreases in the bloodstream, increases and then decreases in the liver and eventually it all resides in the bladder.

4.8 Maximum Likelihood Estimation

In the example of the flow of substances through the body, the transition rate matrix $\mathbf{Q}$ was given and hence spared us the task of estimating the parameters. The current section explores methods for parameter estimation in a time

homogenous Markov model. Maximum likelihood estimation is simple for the time homogeneous Markov model. This simplicity can be attributed to a relatively small, finite number of parameters arising from constant transition intensities, as well as the simple relationship $\mathbf{P}(t) = e^{t\mathbf{Q}}$ outlined in Section 4.7.

If i indexes a particular individual out of M individuals, then the data for individual i consists of the times $(t_{i,1} < \ldots < t_{i,n_i})$ that individual i was observed and the states $(X(t_{i,1}), \ldots X(t_{i,n_i}))$ the individual was in at each of these time points. Then the likelihood function is the product over all individuals and all time points.

$$L = \prod_{i=1}^{M} \prod_{j=1}^{n_i - 1} l_{i,j} \tag{4.25}$$

Jackson (2007) presents the form of the individual likelihood function under different observation schemes. The forms of $l_{i,j}$ under an example of different assumptions are presented below.

Intermittently observed process: Considering a general multistate process, if $(X(t_j), X(t_{j+1}))$ denotes a pair of successive observed states at times t_j and t_{j+1} respectively, then the contribution to the likelihood is

$$l_{i,j} = p_{X(t_j),X(t_{j+1})}(t_{j+1} - t_j)$$

Death states: Usually when the death state is reached, the time of entry into the state is known but the state occupied immediately before death is unknown. In other words we have a scenario similar to left censored information. Letting $X(t_{j+1}) = D$ denote the death state it is logical that the likelihood should be summed over all possible states that could have been occupied the day before death. Hence,

$$l_{i,j} = \sum_{m \neq D} p_{X(t_j)m}(t_{j+1} - t_j) q_{mD}$$

Censored states: It is often the case that at the end of a study or observation, it is known that a patient is alive but in an unknown state. In this case

the likelihood takes the following form

$$l_{i,j} = \sum_{m \in C} p_{X(t_j)m}(t_{j+1} - t_j)$$

where C denotes the set of all possible states.

Exactly observed transition times: If the times $(t_{i,1}, \ldots, t_{i,n_i})$ were exact transition times and no transitions took place in between these transition times then the contribution to the likelihood will take the following form

$$l_{i,j} = p_{X(t_j)X(t_j)}(t_{j+1} - t_j) q_{X(t_j)X(t_{j+1})}$$

The rationale behind this form is that the individual is known to be in state $X(t_j)$ in the interval $[t_j, t_{j+1})$ until moving to state $X(t_{j+1})$ precisely at time t_{j+1}. The use of exactly observed transition times is advantageous and will be demonstrated in this section as a means of computing initial parameter estimates.

There are iterative computer procedures available to handle the maximization process. Kay (1986) outlines the calculation of the starting values of parameters for this procedure. The principal assumption here is that the times t_j represent exact transition times between states. Letting T_i denote the total time spent in state i by all individuals and m_{ij} the total number of transitions from state i to state j, the maximum likelihood estimates of the parameters are then

$$q_{ij}^0 = \frac{m_{ij}}{T_i} \tag{4.26}$$

where values of q_{ij}^0 are zero or undefined. Kay (1986) suggests the use of a global average value over all possible transitions. That is:

$$q_{ij}^0 = \frac{\sum_i \sum_j m_{ij}}{\sum_i T_i} \tag{4.27}$$

Note that Equations (4.26) and (4.27) above assume time independent transition rates.

4.9 Diagnostics for Model Assessment

The assumptions of time homogeneity and the Markov assumption are assumptions which need to be assessed as an incorrect application can lead to bias (Meira-Machado et al., 2009). It is also important to get an overall idea of the goodness of fit and to recognize time intervals in which the model may greatly under or overestimate parameters of interest.

4.9.1 Assessing the Markov Assumption

The Markov assumption that the future state of a process depends only on the present state, independent of the past (Definition 4.1), is a key simplifying assumption in multistate models. Kay (1986) stated that the difficulty of assessing the Markov assumption lies when faced with the absence of exact transition times. He proposed a method of interpolation to estimate exact transition times after which one may apply the formal test explained below. Assuming an illness-death model allowing recovery with states: (1) healthy, (2) illness and (3) death, let x denote the time spent in state 2 from the most recent transition from state 1. Kay (1986) proposed fitting a model where the transition intensity q_{23} is given by

$$q_{23} = \lambda_0 \exp(\beta x)$$

and thereafter testing the hypothesis $H_0 : \beta = 0$. A limitation argued by Titman (2007) is that this test would only be valid if the interpolation to estimate exact transition times is accurate.

4.9.2 Assessing the Time Homogeneity Assumption

The assumption of time homogeneity in Definition 4.3 implies that transition rates are constant over time. This assumption can be tested by fitting a piecewise constant intensities model and thereafter using a likelihood ratio test as a test for time independence. The piecewise intensities model is discussed in detail in Chapter 5. Likelihood ratio tests are often used to compare two nested

models, and take the form

$$LRT = -2\ln\left(\frac{L_s(\hat{\theta})}{L_g(\hat{\theta})}\right) \tag{4.28}$$

where $L_s(\hat{\theta})$ denotes the likelihood of the simplified model (model with fewer parameters) and $L_g(\hat{\theta})$ denotes the likelihood for the general model. The Likelihood Ratio Test statistic asymptotically follows a χ^2 distribution with degrees of freedom equal to the difference in the number of parameters between the general and the simplified model. Although Markov models had been implemented since 1985 (Kalbfleisch and Lawless, 1985), prior to 2002 very little attention was placed on formally measuring the goodness of fit of these models (Aguirre-Hernández and Farewell, 2002).

4.9.3 Prevalence Counts

Gentleman et al. (1994) introduced a method which compares the observed and expected (fitted) data. This method requires two assumptions. The first assumption is that an individual's state at a time t was the same as the state at their previous observation time. Gentleman et al. (1994) state that provided subjects are observed sufficiently frequently bias should be minimal. The second assumption is that the process begins at a common time for all individuals. Suppose at this time, each individual is in state 1. Then, given $n(t_i)$ individuals are under observation at time t_i, the expected number of individuals in state r at time t_i is $n(t_i)P_{1r}(t_i)$. Thereafter one may plot observed and expected prevalence against time.

Letting O_{ri} and E_{ri} denote the observed and expected counts respectively, for a particular state r and time t_i we may gain an indication of the times at which the data deviate from the model by calculating (Titman and Sharples, 2008)

$$\frac{(O_{ri} - E_{ri})^2}{E_{ri}} \tag{4.29}$$

However, one cannot apply formal tests to assess whether these deviances are statistically significant. The reasons for this limitation, given by Titman (2007),

are that the interpolation of observed states and the dependence between the rows of the tables render formal tests nonapplicable.

4.9.4 Residual Plots

Residuals are defined as the differences between the observed values of the outcome variable and the fitted values and are a powerful tool to detect outlying observations. Titman (2007) proposed a method for calculating *score residuals*. Assuming that we have n subjects and a parameter vector $\theta \in \Theta$ with maximum likelihood estimate $\hat{\theta}$ we define the score function

$$U(\theta) = \frac{\partial \ln L(\theta)}{\partial \theta} \tag{4.30}$$

where $L(\theta)$ denotes the likelihood function. The score residual for a single subject is given by

$$U_i(\hat{\theta})' I_i(\hat{\theta})^{-1} U_i(\hat{\theta}) \tag{4.31}$$

where $I(\hat{\theta})$ is the observed Fisher information matrix, that is, minus the matrix of second order derivatives of the log-likelihood for that subject evaluated at $\hat{\theta}$. The greatest challenge in computing score residuals lies in the computation of the derivative of the exponential matrix. This is achieved through the use of the eigenvalue decomposition derived in Section 4.7. That is, $\mathbf{P}(t) = \mathbf{A}\exp(\mathbf{D}t)\mathbf{A}^{-1}$ where $\mathbf{A}$ is the matrix of eigenvectors and $\mathbf{D}$ is the diagonal matrix of eigenvalues of the transition intensity matrix $\mathbf{Q}$. The first derivative with respect to the u^{th} component of θ is (Titman, 2007)

$$\frac{\partial \mathbf{P}(t)}{\partial \theta_u} = \mathbf{A}\mathbf{V}_u\mathbf{A}^{-1} \tag{4.32}$$

where $\mathbf{V}_u$ is a matrix with ij^{th} element given by

$$\begin{cases} \dfrac{g_{ij}^{(u)}(\exp(d_i t) - \exp(d_j t))}{(d_i - d_j)}, & i \neq j \\ g_{ii}^{(u)} t \exp(d_i t), & i = j \end{cases}$$

and $g_{ij}^{(u)}$ is the $(i,j)^{th}$ entry of the matrix $\mathbf{G}_u = \mathbf{A}\dfrac{\partial \mathbf{Q}}{\partial \theta_u}\mathbf{A}^{-1}$. $d_1, \ldots, d_R$ are the eigenvalues of the intensity matrix $\mathbf{Q}(\theta)$, which are assumed to be distinct.

Jackson (2007) explains that individuals who have larger score residuals, have a greater influence on the maximum likelihood estimates.

4.9.5 Quantitative Tests

Aguirre-Hernández and Farewell (2002) proposed a Pearson-type goodness of fit test that involved partitioning the data for each transition based on observation number into observation categories denoted by h and by time interval category l_h. Observations are further grouped by covariate classes c, according to quantiles of the transition intensity q_{rr}. For a particular transition $r \rightarrow s$, for a patient with observations at times $t_j, j = 1, \ldots n$ the observed and expected cell frequencies can be calculated as

$$o_{hl_hrsc} = \sum I[(X(t_{j+1}) = s, X(t_j) = r)] \tag{4.33}$$

and

$$e_{hl_hrsc} = \sum P[(X(t_{j+1}) = s, X(t_j) = r)] \tag{4.34}$$

where $I(A)$ is an indicator function for event A and the summation is over the set of transitions in the category defined by h, l_h, c over all individuals i. The Pearson-type test statistics is then

$$T = \sum_h \sum_{l_h} \sum_r \sum_s \sum_c \frac{(o_{hl_hrsc} - e_{hl_hrsc})^2}{e_{hl_hrsc}} \tag{4.35}$$

Although the Pearson test statistic has an assumed χ^2 distribution, the null distribution of T is not exactly χ^2 owing to non-identical time intervals and consequently observed transitions are realizations of independent non-identical multinomial distributions. Aguirre-Hernández and Farewell (2002) showed that in the absence of covariates the χ^2 distribution is a good approximation. This test, however, is not applicable when censoring or an absorbing state is present.

Titman and Sharples (2008) modified this test to correct for biases introduced by absorbing states and censoring. In the case of exact death times these

are handled by times of death classified transitions to the absorbing state according to the next scheduled observation time after the death, which is estimated by multiple imputation from a Kaplan-Meier estimate of the distribution of time intervals between observations (Jackson, 2007).

4.10 Modeling Multistate Data

The models we have considered so far assume that transition intensities and transition probabilities are the same for all individuals. Individual differences can be due to observed covariates and or unobserved factors. In the current thesis we focus on the case of observed covariates. Multistate models can also assist in understanding the effect of risk factors associated with the onset or progression of illness. The four types of risk factors defined by Mullins (1996) follow:

Predisposing factors e.g. age and previous illness

Precipitating factors e.g. exposure to a viral or carcinogenic agent

Enabling factors e.g. poor nutrition, restricted access to medical care and poverty

Reinforcing factors e.g. repeated exposure and stress

Alioum et al. (1998) used multistate models to study the effect of gender, age, transmission category and antiretroviral therapy on the clinical progression of HIV. This section serves to introduce methods for inferring the effect of covariates on transitions which will be applied in Chapter 6.

In multistate models it is the transition intensities that are modeled, since one can directly allow for these covariate dependent intensities in the model. The modeling of transition intensities was first introduced by Kay (1982). This paper extended the model proposed by Cox (1972), which was introduced in Section 3.5 to multiple events. In recent years alternative strategies have been introduced (Fiocco et al., 2008).

4.10.1 Proportional Hazards Regression Models

The transition intensities are now allowed to depend on a vector of covariates $\mathbf{z}$. That is,

$$q_{ij}(t \mid \mathbf{z}(t)) = q_{ij0}(t) \exp(\boldsymbol{\beta}'_{ij}\mathbf{z}(t)) \tag{4.36}$$

where $\boldsymbol{\beta}_{ij} = (\beta_{ij1}, \beta_{ij2}, \ldots, \beta_{ijr})'$ is the vector of regression coefficients associated with vector $\mathbf{z}(t)$ for the transition from state i to state j. The quantity $q_{ij0}(t)$ represents the baseline transition intensity function. If the ordered transition times are $t_{(1)} < t_{(2)} < \ldots$ then the partial likelihood contribution of a transition from state i to state j at time $t_{(k)}$ with independent variables $\mathbf{z}_{(k)}$ due to Kay (1982) is

$$\frac{q_{ij0}(t_{(k)}) \exp(\boldsymbol{\beta}'_{ij}\mathbf{z}_{(k)}(t_{(k)}))}{\sum_{l \epsilon R(t_{(k)},i)} q_{ij0}(t_{(k)}) \exp(\boldsymbol{\beta}'_{ij}\mathbf{z}_l(t_{(k)}))} \tag{4.37}$$

The risk set $R(t_{(k)}, i)$ in Equation (4.37) is the risk set for making transitions from state i to state j. The way that the risk set is defined is one of the features in multistate models which differs from the risk set defined in Section 3.5 under the Cox (1972) partial likelihood approach. If the ordered transition times from state i to state j are

$$t_{ij1} < t_{ij2} < \ldots < t_{ijn}$$

then the partial likelihood for that transition is given by

$$L(\boldsymbol{\beta}_{ij}) = \prod_{m=1}^{n} \frac{\exp(\boldsymbol{\beta}'_{ij}\mathbf{z}_{ijm}(t_{ijm}))}{\sum_{l \epsilon R(t_{ijm},i)} \exp(\boldsymbol{\beta}'_{ij}\mathbf{z}_l(t_l))} \tag{4.38}$$

The above formulation is therefore a generalization of the Cox (1972) partial likelihood function to multistate models. The notation in Equation (4.36) indicates that separate baseline intensities are fitted for each transition. In the literature, however, it is often assumed that all transitions going into the same state have proportional baseline hazards (Fiocco et al., 2008). Meira-Machado et al. (2007) present the illness death model as an example. Here baseline intensities for the transitions $healthy \rightarrow dead$ and $sick \rightarrow dead$ are assumed proportional. That is, $q_{13}(t \mid z) = q_{130}(t) \exp(\boldsymbol{\beta}'_{13}z)$ and $q_{23}(t \mid z) = q_{130}(t) \exp(\boldsymbol{\beta}'_{23}z + \delta)$ where $\exp \delta$ is the constant of proportionality.

4.10.2 Additive Hazards Regression Models

Aalen (1980) proposed an alternative method for modeling survival data, which allowed time varying regression coefficients. The model proposed was

$$\alpha(t \mid \mathbf{z}) = \beta_0(t) + \beta_1(t)z_1 + \beta_2(t)z_2 + \ldots + \beta_p(t)z_p \tag{4.39}$$

As stated in Hosmer and Royston (2002), the coefficients in the above model can be interpreted as the change in hazard at time t, from the baseline hazard function $\beta_0(t)$ for a one-unit change in the covariate, holding all other covariates constant. The model extends to multistate models and is commonly referred to as Aalen's non-parametric additive model. The extension, due to Andersen and Keiding (2002) can be specifically stated as

$$q_{ij}(t \mid \mathbf{z}) = q_{ij0}(t) + \boldsymbol{\beta}'_{ij}(t)\mathbf{z} \tag{4.40}$$

Further detail regarding the model and methods employed to estimate parameters can be found in Hosmer and Royston (2002).

Chapter 5

Time Non-homogeneous Markov Model

The non-homogeneous model is used when there is an underlying reason for transition rates to change with time or age. The following two examples illustrate this concept.

1. Suppose we were examining the simple "illness-death" model. It is a known fact that rates of falling sick are higher in winter and recovery rates are lower. Hence, the transition rates vary with time.
2. As patients get older, the risk of death due to certain disease increases and their chances of recovery decrease. The opposite is true for younger patients. Clearly, in this case transition rates would be a function of age.

As introduced in Section 2.2, the time non-homogeneous Markov model considers time dependent transition intensities of the following form

$$q_{ij}(t, \mathcal{F}_{t-}) = q_{ij}(t) \tag{5.1}$$

There are two approaches for handling the aforementioned model that have been documented in the literature. These are

- Parametric methods
- Non-Parametric methods

5.1 Parametric Methods

As stated by Meira-Machado et al. (2009) the simplest and most widely used parametric procedure is conducted by partitioning the observation time into intervals, and assuming that within each interval the transition rate or intensity remains constant. This is often referred to as the piecewise constant intensities model. In other words, the transition intensities q_{ij} are step functions and the transition rate matrix is of the following form

$$\mathbf{Q}(t) = \begin{cases} \mathbf{Q}_0 & t < \tau_1 \\ \mathbf{Q}_i & \tau_i \leq t < \tau_{i+1}, i = 1, 2, \ldots, m-1 \\ \mathbf{Q}_m & t \geq \tau_m \end{cases}$$

where $\tau_1, \tau_2 \ldots \tau_m$ are the designated cutpoints and t denotes time since the initiation of the process.
Now, using results from the previous chapter and times t_a and t_b such that $\tau_k < t_a < t_b < \tau_{k+1}$

$$\mathbf{P}(t_a, t_b) = \mathbf{P}(t_a, \tau_{i+1})\mathbf{P}(\tau_{i+1}, \tau_{i+2}) \ldots \mathbf{P}(\tau_{j-1}, \tau_j)\mathbf{P}(\tau_j, t_b)$$

Each of the terms of this product give a transition probability matrix within a time interval of constant transition rate and therefore can be computed using the Kolmogorov forward equation.

The above holds when t_a and t_b both fall within the same interval. If this is not the case such that $\tau_i \leq t_a < \tau_{i+1}$ and $\tau_j \leq t_b < \tau_{j+1}$ where $j > i$ then by the Chapman Kolmogorov equation proved in Theorem 4.3

$$\mathbf{P}(t_a, t_b) = \exp(\mathbf{Q}_k(t_b - t_a)$$

The likelihood function below is the same as that for the time homogeneous process where i indexes an individual out of M individuals and j indexes the times at which the individual is observed.

$$L = \prod_{i=1}^{M} \prod_{j=1}^{n_i - 1} l_{i,j} \tag{5.2}$$

The terms of the likelihood however, are computed differently as will be described below. According to Pérez-Ocón et al. (2001), if the observed transition interval for a patient is between two cutpoints, the contribution to the likelihood is the transition probability with the corresponding $\mathbf{Q}$ matrix. If the observed transition interval has one cutpoint, the contribution to the likelihood is the product of the transition probability in the interval between the instant of the jump and the cutpoint, and from this point to the next jump or censoring, with the corresponding $\mathbf{Q}$ matrices in each period. If the observed transition interval is between k cutpoints, the contribution is the product of $k+1$ transition probabilities with the corresponding $\mathbf{Q}$ matrices. If the last time observed is a death then the last product in the likelihood is a transition probability to the absorbing state. However, if the last observed time is a censoring, then the last product term is a probability of surviving in the last state.

The piecewise process is a very powerful tool, not only owing to its ability to handle time dependence, but also in its ability to test whether the assumption of time homogeneity is valid in other models. This concept was discussed in Section 4.9. Pérez-Ocón et al. (2001) studied relapse and survival times for 300 breast cancer patients under different forms of treatment. A piecewise constant intensities model was shown to be more appropriate than the homogeneous model. The two intervals chosen were $0 \leq t < 48$ and $t \geq 48$. Meira-Machado et al. (2007) outlines the limitation of the piecewise model as the lack of clarity on how to choose the number of cutpoints and the values of the cutpoints.

5.2 Non-parametric Methods

5.2.1 Introducing the Product Integral

The product integral plays a central role in survival analysis and stochastic process theory. In fact, the Kaplan-Meier estimator of the survival function is the product integral of the Nelson-Aalen estimator of the cumulative intensity function. This section introduces the product integral for a simple survival

analysis. Consider breaking the period $(0, t]$ into k subintervals such that $0 = t_0 < t_1 < \ldots < t_K = t$. We have :

$$\begin{aligned} S(t) &= P(T_1 > t_1)P(T_2 > t_2|T_1 > t_1)\ldots P(T_K > t_K|T_{K-1} > t_{K-1}) \\ &= \prod_{k=1}^{K} P(T > t_k|T > t_{k-1}) \\ &= \prod_{k=1}^{K} S(t_k|t_{k-1}) \end{aligned} \tag{5.3}$$

where $S(t_k|t_{k-1}) = P(T > t_k|T > t_{k-1})$.

This follows by noting that

$$\begin{aligned} S(t_k|t_{k-1}) &= P(T > t_k|T > t_{k-1}) \\ &= \frac{P(T > t_k \cap T > t_{k-1})}{P(T > t_{k-1})} \\ &= \frac{P(T > t_k)}{P(T > t_{k-1})} \end{aligned}$$

and

$$\begin{aligned} S(t_1|t_0) &= P(T > t_1|T > t_0) \\ &= \frac{P(T > t_1)}{P(T > t_0)} \\ &= P(T > t_1) \end{aligned}$$

since $P(T > t_0) = 1$.

Using results introduced in Section 3.3 we know that

$$dS(t) = -S(t-)dA(t) \tag{5.4}$$

From Equation (5.4) we have the approximation

$$S(t_k) - S(t_{k-1}) \approx -S(t_{k-1})(A(t_k) - A(t_{k-1})) \tag{5.5}$$

or

$$S(t_k \mid t_{k-1}) \approx 1 - (A(t_k) - A(t_{k-1})) \tag{5.6}$$

Substituting the above equation into Equation (5.3) we have

$$S(t) \approx \prod_{k=1}^{K}(1 - (A(t_k) - A(t_{k-1}))) \tag{5.7}$$

As stated by Aalen (2008) if the number of subintervals increase while their lengths tend to zero in a uniform way, the product on the right hand side will approach a limit called the *product integral* and the approximation will improve. Hence $S(t)$ may be expressed in terms of the product integral as

$$S(t) = \prod_{u \leq t}(1 - dA(u)) \tag{5.8}$$

When $A(u)$ is absolutely continuous we have

$$dA(u) = \alpha(u)du$$

Recall that the Nelson-Aalen estimator, which was introduced in Section 3.4 is given by

$$\hat{A}(t) = \sum_{t_{(j)} \leq t} \frac{d_j}{n_j} \tag{5.9}$$

The product integral of $\hat{A}(t)$ is then

$$\begin{aligned} \hat{S}(t) &= \prod_{0}^{t}(1 - d\hat{A}) \\ &= \prod_{t_j \leq t}\left(1 - \frac{d_j}{n_j}\right) \end{aligned}$$

which is the Kaplan-Meier estimator.

5.2.2 Expressing P(s,t) as a function of Q(t)

Let $p_{ij}(s,t)$ denote the time dependent transition probability for a Markov process. Then

$$p_{ij}(s,t) = P(X(t) = j | X(s) = i) \tag{5.10}$$

for $i, j \in S$ and $s, t \in T$. The transition probabilities in Equation (5.10) yield the transition probability matrix **P**(s,t) given by

$$\mathbf{P}(s,t) = \begin{pmatrix} p_{11}(s,t) & p_{12}(s,t) & \cdots & p_{1n}(s,t) \\ p_{12}(s,t) & p_{22}(s,t) & \cdots & p_{2n}(s,t) \\ \vdots & \vdots & \cdots & \vdots \\ p_{n1}(s,t) & p_{n2}(s,t) & \cdots & p_{nn}(s,t) \end{pmatrix}$$

This section illustrates the application of product integration to Markov processes. Returning to the Kolmogorov equation which was introduced in Theorem 4.5, and extending the equation to a non-homogeneous Markov process we have

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

which holds for **Q**(t) due to Aalen (2008) defined as

$$\mathbf{Q}(t) = \lim_{h \longrightarrow 0} \frac{1}{h}(\mathbf{P}(t, t+h) - \mathbf{I})$$

Using the general solution case, the Kolmogorov equation can be written as

$$\mathbf{P}(s,t) = \mathbf{I} + \int_s^t \mathbf{P}(s, u-) d\mathbf{A}(u) \tag{5.11}$$

which is the extension of the expression in (5.4) to multistate Markov processes. We define $\mathbf{A}(t)$ to be a square matrix function with ij^{th} element

$$A_{ij}(t) = \int_0^\infty q_{ij}(s) ds \tag{5.12}$$

and ii^{th} element

$$A_{ii}(t) = -\sum_{i \neq j} A_{ij}(t) \tag{5.13}$$

Given interval specific transition matrix $P(t_i, t_j)$, $t_i < t_j$ we can write

$$P(s,t) = P(t_0, t_1)P(t_1, t_2)P(t_2, t_3) \ldots P(t_{k-1}, t_k)$$

where $s = t_0 < t_1 < \ldots < t_K = t$.

By the same argument as in the previous section we can approximate Equation (5.11) as

$$\mathbf{P}(s,t) \approx \prod_1^K (\mathbf{I} + (\mathbf{A}(t_k) - \mathbf{A}(t_{k-1}))) \tag{5.14}$$

Now, letting the lengths of intervals go to zero,

$$\mathbf{P}(s,t) = \prod_{u\epsilon(s,t]} (\mathbf{I} + d\mathbf{A}(u)) \tag{5.15}$$

The product integral representation of the transition probability matrix only assumes existences of cumulative transition intensities which need not necessarily be continuous (Borgan, 1997).

5.2.3 The Aalen-Johansen Estimator

Assuming that exact transition times are recorded and denoting

d_{ijk} = Number of individuals who experience a transition from state i to j at time t_k (the risk set)

r_{ik} = Number of individuals in state i just prior to time t_k

d_{ik} = $\sum_{i \neq j} d_{ijk}$ denotes the number of transitions out of state i at time t_k

The Aalen-Johansen estimator takes the following form

$$\widehat{\mathbf{P}(s,t)} = \prod_{s<t_j \leq t} (\mathbf{I} + \widehat{\mathbf{q}}_j) \tag{5.16}$$

where $\mathbf{I}$ is an $n \times n$ identity matrix, and $\widehat{\mathbf{q}}_j$ is a $n \times n$ matrix with ij^{th} element

$$\widehat{q}_{ijk} = \frac{d_{ijk}}{r_{ik}}$$

for $i \neq j$ while the ii^{th} element is given by

$$\widehat{q}_{iik} = -\frac{d_{ik}}{r_{ik}}$$

Returning to Equation (5.8) in the previous section, we will now illustrate the link between the Aalen-Johansen estimator and the Nelson-Aalen Estimator:

$$\mathbf{P}(s,t) = \prod_{u\epsilon(s,t]} (\mathbf{I} + d\mathbf{A}(u)) \tag{5.17}$$

For $i \neq j$ we estimate the cumulative transition intensity $A_{ij}(t)$ by the Nelson-Aalen estimator introduced in survival analysis. That is

$$\widehat{A}_{ij}(t) = \sum_{t_k \leq t} \widehat{q}_{ijk} \tag{5.18}$$

$$= \sum_{t_k \leq t} \frac{d_{ijk}}{r_{ik}} \tag{5.19}$$

and,

$$\widehat{A}_{ii}(t) = \sum_{t_k \leq t} \widehat{q}_{iik} \tag{5.20}$$

$$= \sum_{t_k \leq t} -\frac{d_{ik}}{r_{ik}} \tag{5.21}$$

These elements form the square matrix $\widehat{\mathbf{A}}(t) = \sum_{t_j \leq t} \widehat{\mathbf{q}}_j$ Returning to

$$\mathbf{P}(s,t) = \prod_{u \epsilon (s,t]} (\mathbf{I} + d\mathbf{A}(u)) \tag{5.22}$$

we see that

$$\widehat{\mathbf{P}}(s,t) = \prod_{u \epsilon (s,t]} (\mathbf{I} + d\widehat{\mathbf{A}}(u)) \tag{5.23}$$

Since $\widehat{\mathbf{A}}(t)$ is matrix of step functions we see that this is the same as the Aalen-Johansen estimator in Equation (5.16).

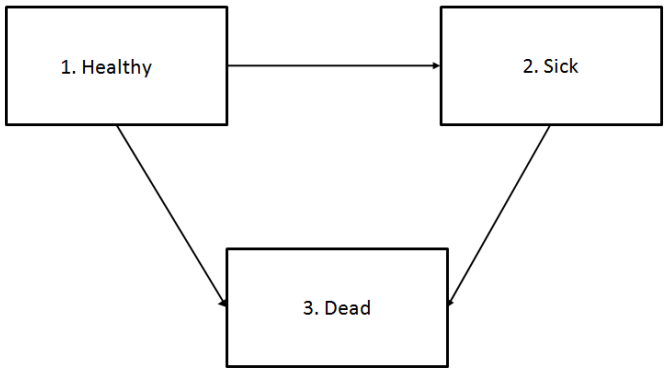

Figure 5.1: Illness-Death model

Example 5.1

For simple models it is possible to give explicit expressions for transition probabilities, as will be demonstrated below for the Illness-Death model which is illustrated in Figure 5.1.

The transition rate matrix takes the following form

$$\mathbf{Q}(t) = \begin{pmatrix} q_{11}(t) & q_{12}(t) & q_{13}(t) \\ 0 & q_{22}(t) & q_{23}(t) \\ 0 & 0 & 0 \end{pmatrix}$$

The corresponding transition probability matrix is as follows

$$\mathbf{P}(s,t) = \begin{pmatrix} p_{11}(s,t) & p_{12}(s,t) & p_{13}(s,t) \\ 0 & p_{22}(s,t) & p_{23}(s,t) \\ 0 & 0 & 1 \end{pmatrix}$$

Note that $q_{31}(t) = 0$, $q_{32}(t) = 0$ and $q_{33}(t) = 0$ because "dead" is an absorbing state. Letting $t_{(1)} < t_{(2)} < \ldots < t_{(d)}$ denote the ordered event times for illness onset or death and

d_{12k} = Number of individuals who become diseased at time $t_{(k)}$

d_{13k} = Number of healthy individuals who die at time $t_{(k)}$

d_{23k} = Number of sick individuals who die time $t_{(k)}$

r_{1k} = Number of healthy individuals just prior to time $t_{(k)}$

r_{2k} = Number of sick individuals just prior to time $t_{(k)}$

Using the Aalen-Johansen estimator we have

$$\widehat{p_{11}}(s,t) = \prod_{s<t_{(k)}\leq t}\left(1 - \frac{d_{12k} + d_{13k}}{r_{1k}}\right) \tag{5.24}$$

and

$$\widehat{p_{22}}(s,t) = \prod_{s<t_{(k)}\leq t}\left(1 - \frac{d_{23k}}{r_{2k}}\right) \tag{5.25}$$

The estimate of $p_{12}(s,t)$ is slightly more complicated and requires the following reasoning due to Meira-Machado et al. (2009):

$$p_{12}(s,t) = \int_s^t p_{11}(s,u)q_{12}(u)p_{22}(u,t)du \tag{5.26}$$

The estimator for $p_{12}(s,t)$ is given by

$$\widehat{p_{12}}(s,t) = \sum_{s<t_{(k)}\leq t}\widehat{p_{11}}(s,t_{(k-1)})\frac{d_{12k}}{r_{1k}}\widehat{p_{22}}(t_{(k)},t) \tag{5.27}$$

which is found by substituting estimates for each term in Equation (5.26). $q_{12}(u)$ is substituted by $d\widehat{A}_{12}(u)$ which is the increment of the estimator $A_{12}(u) = \sum_{t_{(k)}\leq t}\frac{d_{12k}}{n_{1k}}$.

5.3 Other Markov Models

All the previous theory has been dedicated to *Markov* processes. In many cases, however, the Markov assumption may not be appropriate. There exist other types of Markov models which possess different properties. These processes are briefly discussed below.

5.3.1 The Hidden Markov Model

Hidden Markov models were originally introduced and studied in the late 1960s and early 1970s, and since the 1980s have become very popular models for speech recognition and protein modeling. In a hidden Markov model (HMM) the states of the Markov chain are not observed. The observed data are governed by some probability distribution (the emission distribution) conditionally on the unobserved state. This definition is difficult to grasp and hence is better introduced through an example.

Karlsson (2000) describes the following urn and coloured ball example.

Example 5.2

Consider a room with N urns. Within each urn there are a large number of coloured balls. We assume that there are M different colours in total. Furthermore, assume that an urn is initially chosen according to some probability distribution. From this urn, a ball is chosen at random, and its colour is recorded as the observation. The ball is then replaced in the urn from which it was selected. A new urn is selected according to a random selection process associated with the current urn.

The ball selection process is repeated for the new urn, after which the next urn is selected according to a selection process associated with the second urn, and so forth. The entire process generates a finite observation sequence of colours which we would like to model as the observable output of a HMM. We can now see that we have an underlying Markov chain, where each state corresponds to the selection of a particular urn. This chain is, however, not observable but can be observed through the sequence of colours which obviously is a probabilistic function of the embedded Markov chain, since a colour is chosen randomly depending on the state which we are currently in, i.e. the urn, which we are currently choosing the ball from.

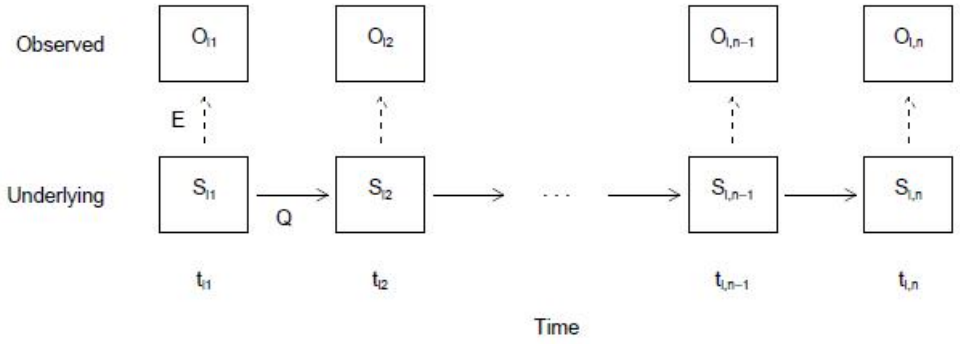

Figure 5.2: A hidden Markov model in continuous time

Jackson (2007) describes HMMs as mixture models, where observations are generated from unknown distributions. This distribution, however, changes through time according to states of a hidden Markov chains. The true state of the model S_{ij} evolves as a Markov process which is underlying and not observed. Observed data O_{ij} are generated conditionally on the true states S_{ij} according to a set of distributions $f_1(y \mid \theta_1, \gamma_1), f_2(y \mid \theta_2, \gamma_2), \ldots f_n(y \mid \theta_n, \gamma_n)$ where n denotes the number of states and θ_i the parameter for the state i distribution which can accommodate dependence on covariates through a link transformed linear model with coefficients γ_i (Jackson, 2007). Figure 5.2, taken from Jackson (2007), illustrates the evolution of a hidden Markov model.

General HMMs have proven to be useful when modeling a chronic disease whose stages can only be identified by an error prone continuous biological marker such as CD4 count in HIV/AIDS and FEV in bronchial obliterans syndrome or cancer stages through cancer screening tests.

5.3.2 The Misclassification Model

Misclassification of states occur often, particularly between adjacent states. In the misclassification model, which is a type of hidden Markov model (Karlsson, 2000), the observed data are states which are assumed to be misclassifications

of the true, underlying states.

Consider a disease progression model with at least a disease-free and a disease state. When screening for the presence of the disease, the screening process can sometimes be subject to error. In such cases the Markov disease process $S_i(t)$ for individual i is not observed directly, but through realizations $O_i(t)$.

The quality of a diagnostic test is often measured by the probabilities that the true and observed states are equal, that is $P(O_i(t) = r|S_i(t) = r)$. When r represents a "positive" or disease state, this is the sensitivity, or the probability that a true positive is detected by the test. When r represents a "negative" or disease-free state this represents the specificity, or the probability that, given the condition of interest is absent, the test produces a negative result.

Observed states O_{ij} for patient i at time t_{ij} are generated conditionally on true states S_{ij} according to a misclassification matrix $\mathbf{E}$. For a finite number of states n, $\mathbf{E}$ is an $n \times n$ matrix whose (r, s) entry is

$$e_{rs} = P(O_{ij}(t) = r|S_{ij}(t) = r)$$

so that

$$\mathbf{E} = \begin{pmatrix} e_{11} & e_{12} & \cdots & e_{1n} \\ e_{21} & e_{22} & \cdots & \\ \vdots & & & \\ e_{n1} & & \cdots & e_{nn} \end{pmatrix}$$

5.3.3 The Semi-Markov Model

Semi-Markov multistate models are Markov extension models which allow for dependence on the time spent in the current state. This dependence is referred to as duration dependence. In a semi-Markov model the future evolution not only depends on the current state, but also on t_i , the entry time into the current state i. The form of the transition intensities which we model then changes to $q_{ij}(t, t - t_i)$. In a semi-Markov model two assumptions of Chapter 4 are relaxed.

- The durations in each state need not be exponential.
- The transitions to the next state may depend on the time spent in the current state.

Although the semi-Markov model is more flexible than the time homogeneous or piecewise time homogeneous Markov model, there are two drawbacks to its use. Firstly, these models contain many parameters, which make them more difficult to fit. Secondly, we can only use these models if we know what distribution the sojourn time (time spent in each state) follows.

Chapter 6

Application of Multistate Models to HIV Progression

6.1 Data

The Sinikithemba Study was started in 2005 (and is still in progress). Thus the data used in the current thesis is part of the continuing study. This study recruited 451 HIV positive individuals. Initial information such as demographics, HLA typing, cellular immunology, CD4 count and viral load were collected. CD4 counts were then taken every three months and viral loads were taken every six months. The study takes place at McCords hospital in Durban, South Africa.

Of the 451 participants enrolled, there were 115 participants with less than two CD4 count readings or missing baseline information who were excluded from the current analysis. This analysis examines the immune deterioration of 336 ARV naive HIV positive patients enrolled in the Sinikithemba study between January 2005 and December 2009. The decision to exclude observations and enrollments post December 2009 was based on limiting the period bias that could be introduced due to the change in ARV treatment guidelines as at 1 December 2009. Patients were followed for a median of 3.54 years (IQR 1.91 – 4.52 years) and had a median of 12 visits (IQR 4 – 17 visits). The median time between visits was 0.26 years. There were 257 (76.49%) participants who had

a baseline CD4 count less than or equal to 500. Compared to the CAPRISA SAPIT study used in Chapter 3 to demonstrate time to single event analysis, the Sinikithemba study data is suitable for multistate models because patients were followed over a longer period of time with smaller intervals between visits.

Table 6.1 summarizes baseline information on study participants. From Table 6.1 we see that the number of females exceeded the number of males and mean baseline CD4 count was well over the 200 cell threshold.

Table 6.1: Baseline characteristics of study participants

Characteristic	Statistic
Age mean (95% CI)	32.60 (31.73, 33.48)
Females n(%)	270 (80.36%)
Baseline CD4 count Mean (95% CI)	381.22 (359.21, 403.22)

In Figure 6.1 the CD4 counts are plotted against time for two randomly selected participants.

Figure 6.1(b) shows that participant SK-063 had a relatively steady decline in CD4 count over six years post enrollment. Participant SK-010 in Figure 6.1(a) however, displays more "erratic" behavior of CD4 count exhibiting an undulating pattern. There are many reasons that have been cited in the literature explaining CD4 count fluctuation. One of these reasons, reported by Malone et al. (1990), is diurnal fluctuation of CD4 count and other subject specific characteristics including genetic factors. In this study a significant CD4 cell count diurnal increase of 59 cells/mm3 (p-value = 0.018), was detected between 8am and 10pm for male HIV positive patients. Other reasons for fluctuation include seasonal change, psychological and physical stress, diet and the menstrual cycle (Crowe et al., 1996).

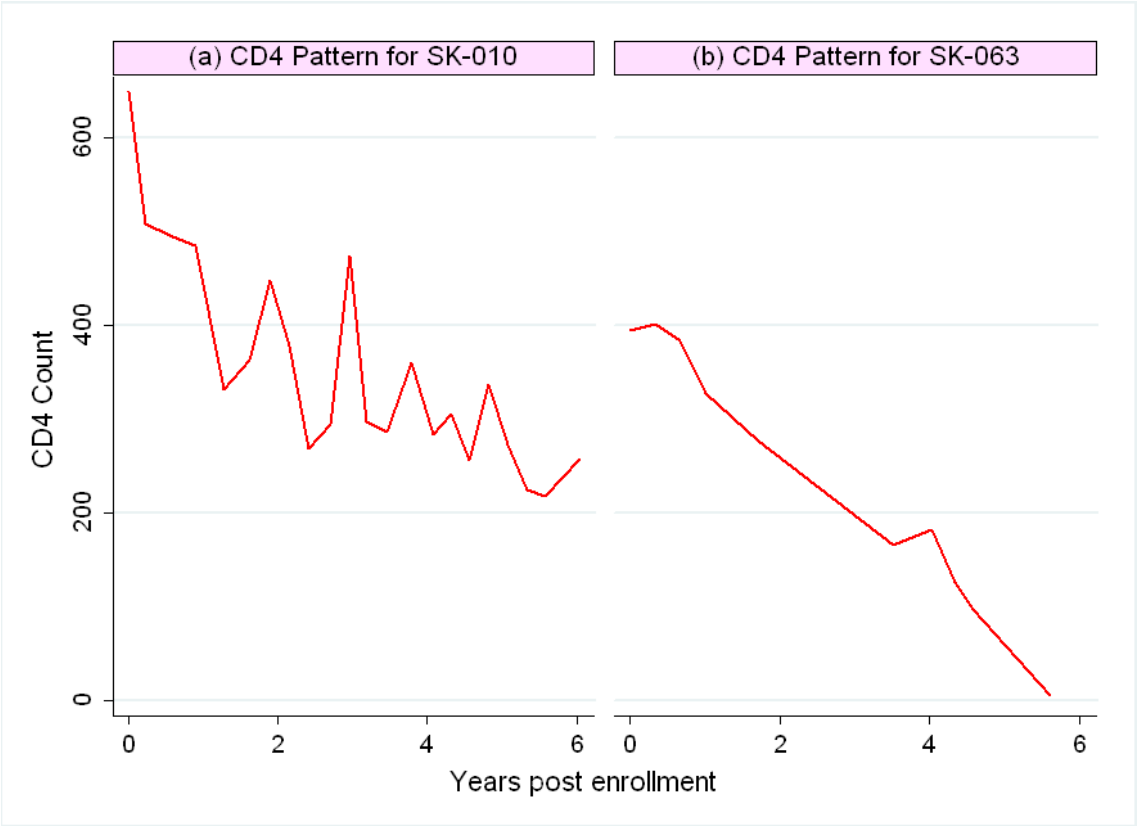

Figure 6.1: Pattern of decline for two randomly selected participants

6.2 Motivation for Multistate Model

In 2008 UNAIDS (UNAIDS, 2008) estimated that there were 5.7 million people living with HIV in South Africa. More than 3 million of these are women aged 15 and above and 280,000 are children aged between 0 and 14 years making South Africa the world's largest population of people living with HIV. As stated in the introduction, an understanding of HIV progression and factors that influence disease progression can have great value in understanding HIV pathogenesis and in the development of new treatment strategies (Sabin et al., 1998). Although substantial research has been done in the area of CD4 count modeling using linear mixed models, no attempt has been made to estimate the length of stay in different CD4 count intervals or to investigate probabilities of transitions to lower CD4 counts over time in South Africa. This thesis goes even further in that it seeks to estimate the effect of age, gender and baseline CD4 count on individual transition rates.

The general pattern of CD4 count decline in HIV positive individuals throughout HIV infection is illustrated in Figure 6.2 as reported in Sabin et al. (1998). We see a sudden drop in CD4 cells at seroconversion, which soon after returns to normal, after which a gradual decline is observed. Our main objective is to describe this gradual decline of cells in the Sinikithemba (SK) cohort, and to assess the effect of cofactors or covariates on the disease progression. The multistate approach is particularly useful for this type of problem because the data is characterized by a high degree of interval and right censoring. Furthermore, patients are at different stages in disease progression as this was not an acute infection cohort.

We have chosen a state structure based on four intervals of CD4 count and ARV initiation as an absorbing fifth state. Upon careful examination of CD4 count it became clear that reverse transitions need to be incorporated into the model. Bwayo et al. (1995) also noted erratic decline in CD4 counts in their data and were the first to include reverse transitions in CD4 count modeling. The

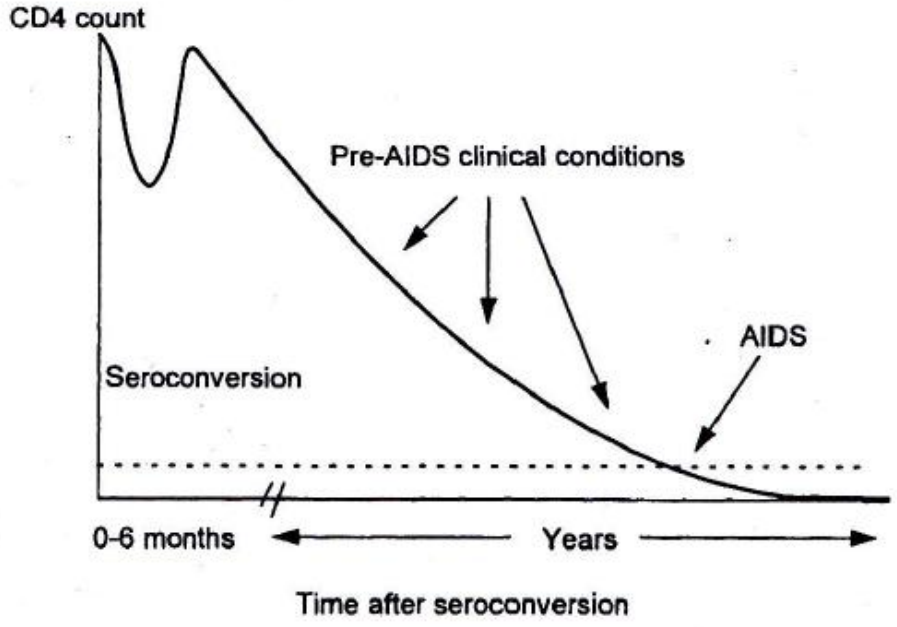

Figure 6.2: Pattern of CD4 count change throughout HIV infection

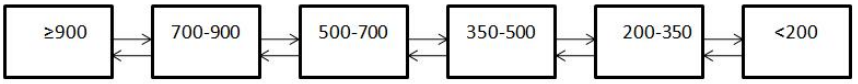

Figure 6.3: State structure applied to HIV progression in Kenyan sex workers and mothers

state structure used in Bwayo et al. (1995) is shown in Figure 6.3. The decision to designate ARV status as the end point was based on the primary objective of this study, which is to determine rates of immune deterioration in *ARV naive* patients. The transition state diagram is illustrated in Figure 6.4.

The labeling of states associated with HIV disease progression in terms of CD4 count are as follows:

1: CD4 Count ≥ 500

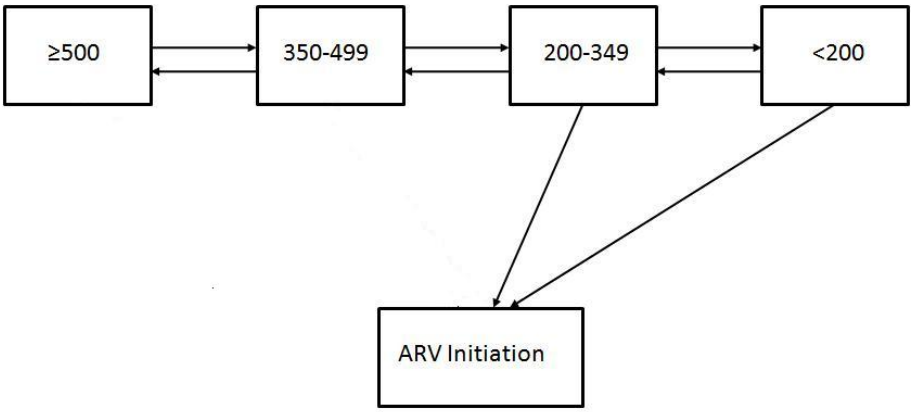

Figure 6.4: Graphic representation of the Sinikithemba data model

2: $350 \leqq$ CD4 Count < 500

3: $200 \leqq$ CD4 Count < 350

4: CD4 Count < 200

5: ARV Initiation

The number of state transitions observed in the SK data are summarized in Table 6.2

Table 6.2: Observed transitions between states

	To State				
From State	1	2	3	4	5
1	376	178	19	0	0
2	148	473	269	7	2
3	16	205	909	159	25
4	0	2	80	247	81

Through examination of the state transitions we see that there were 2 ARV initiations after a previous CD4 count of 350-499 and 25 ARV initiations after a previous CD4 count of 200-349. Treatment guidelines state that all patients

presenting with WHO stage 4 illnesses be initiated on ARVs regardless of CD4 count. Such conditions include Karposi Sarcoma. Upon reaching a CD4 count less than 250 patients are referred to an ARV clinic for treatment. As is visible in the observed transitions, prior to 1st December 2009 patients were being initiated on therapy quite late (81 Transitions to ARV initiation from CD4 count less than 200).

As mentioned in the previous section there is quite strong evidence of a CD4 increase in patients. The state transitions confirm this, since there were a total of 433 (148+205+80) transitions of immune recovery that is, from a higher to a lower state. The model in Figure 6.4 is specified by the transition intensity matrix $\mathbf{Q}$.

$$\mathbf{Q}(t) = \begin{pmatrix} q_{11}(t) & q_{12}(t) & 0 & 0 & 0 \\ q_{21}(t) & q_{22}(t) & q_{23}(t) & 0 & 0 \\ 0 & q_{32}(t) & q_{33}(t) & q_{34}(t) & q_{35}(t) \\ 0 & 0 & q_{43}(t) & q_{44}(t) & q_{45}(t) \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

The $\mathbf{Q}$ matrix defines which instantaneous transitions can occur in the Markov process, and that the data are "snapshots" of the process. Although there were 19 occasions on which a patient was observed in state 1 followed by state 3, the underlying model specifies that the patient must have passed through state 2 in between. The same rationale applies to other observed "jumped" states.

We begin by fitting the simplest model to the data which is the time homogenous Markov model. These two simplifying assumptions allow us to modify the transition intensity matrix such that :

$$\mathbf{Q} = \begin{pmatrix} -q_{12} & q_{12} & 0 & 0 & 0 \\ q_{21} & -(q_{21}+q_{23}) & q_{23} & 0 & 0 \\ 0 & q_{32} & -(q_{32}+q_{34}+q_{35}) & q_{34} & q_{35} \\ 0 & 0 & q_{43} & -(q_{43}+q_{45}) & q_{45} \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

The state transition diagram is depicted in Figure 6.5 below. Here the between state transition rates are denoted by q_{ij} where $i \neq j$ and $i, j = 1, 2, 3, 4, 5$

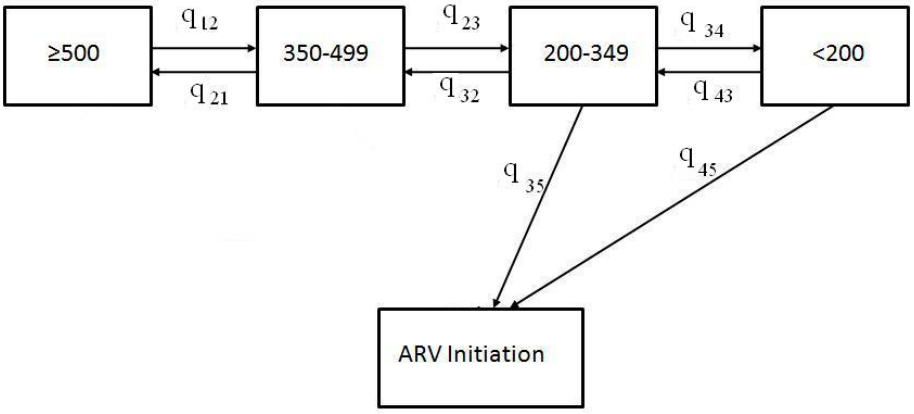

Figure 6.5: Graphic representation of the model with parameters

Initial estimates of the transition intensity matrix were calculated by assuming exact transition times. The likelihood calculation and iterative procedure were explored in the previous section (Section 4.8). The initial estimate of the transition rate matrix is given by

$$\mathbf{Q}_0 = \begin{pmatrix} -0.925 & 0.925 & 0.000 & 0.000 & 0.0000 \\ 0.561 & -1.580 & 1.019 & 0.000 & 0.0000 \\ 0.000 & 0.511 & -0.969 & 0.396 & 0.0623 \\ 0.000 & 0.000 & 0.606 & -1.219 & 0.6131 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.0000 \end{pmatrix}$$

This estimate is used to compute maximum likelihood estimates of q_{ij} where $i \neq j$ and $i, j = 1, 2, 3, 4, 5$, which are given in Table 6.3.

Table 6.3: Time homogeneous Markov model parameter estimates

Parameter	Estimate (95% CI)
q_{12}	1.528 (1.310,1.783)
q_{21}	0.922 (0.779,1.091)
q_{23}	1.576 (1.394,1.783)
q_{32}	0.8721 (0.759,1.003)
q_{34}	0.5612 (0.479,0.657)
q_{35}	0.02387 (0.009,0.066)
q_{43}	0.8412 (0.670,1.056)
q_{45}	0.6682 (0.540,0.827)
$-2\ln L$	5891.98

Before extracting important functions such as mean sojourn times and transition probability matrices it is important to find out whether there are certain individuals having a profound effect on the parameter estimates. We computed score residuals for each patient to achieve this purpose. Score residuals are expressed as a function of the score function and observed Fisher information matrix which were outlined in Section 4.9. The score residuals for all patients appear in Appendix D and are displayed graphically in Figure 6.6.

As Figure 6.6 shows, there are three subjects with particularly large influence on the model likelihood. The state trajectory of patients with score residuals greater than 0.40 appear in Figure 6.7.

The pattern of influence generally exhibits a close association with the amount of follow-up time. Patients SK-086 and SK-163 have maintained a CD4 Count greater than 500 over a six year period. These patients could be “elite controllers”, which are people infected with HIV who do not lose CD4 cells or have the virus in their blood. Subjects for which an ARV initiation was observed also tend to have higher influence.

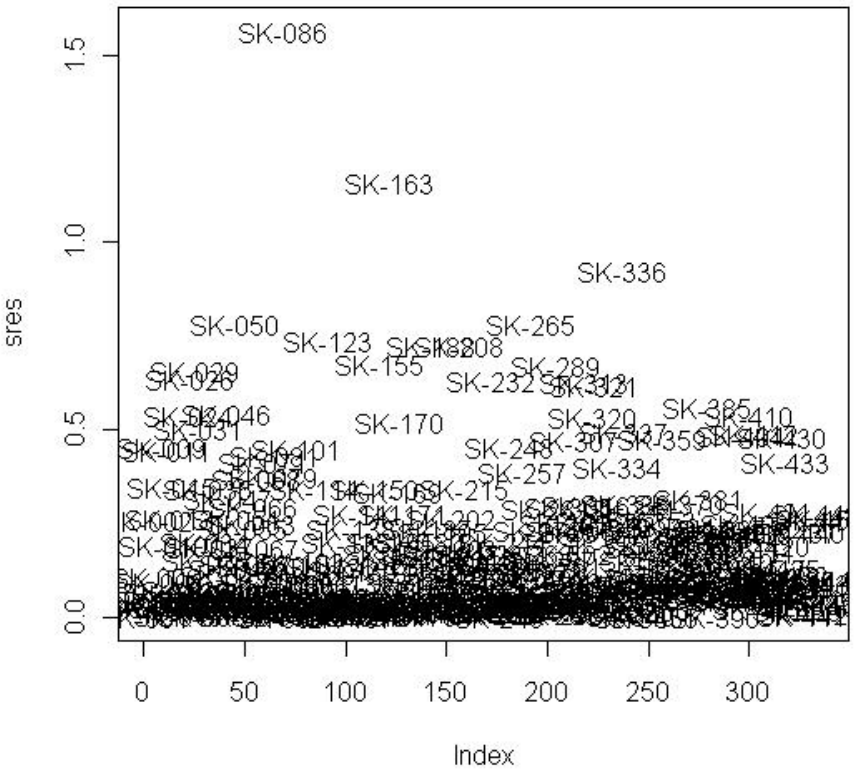

Figure 6.6: Plot of score residuals to examine individual patient influence on likelihood

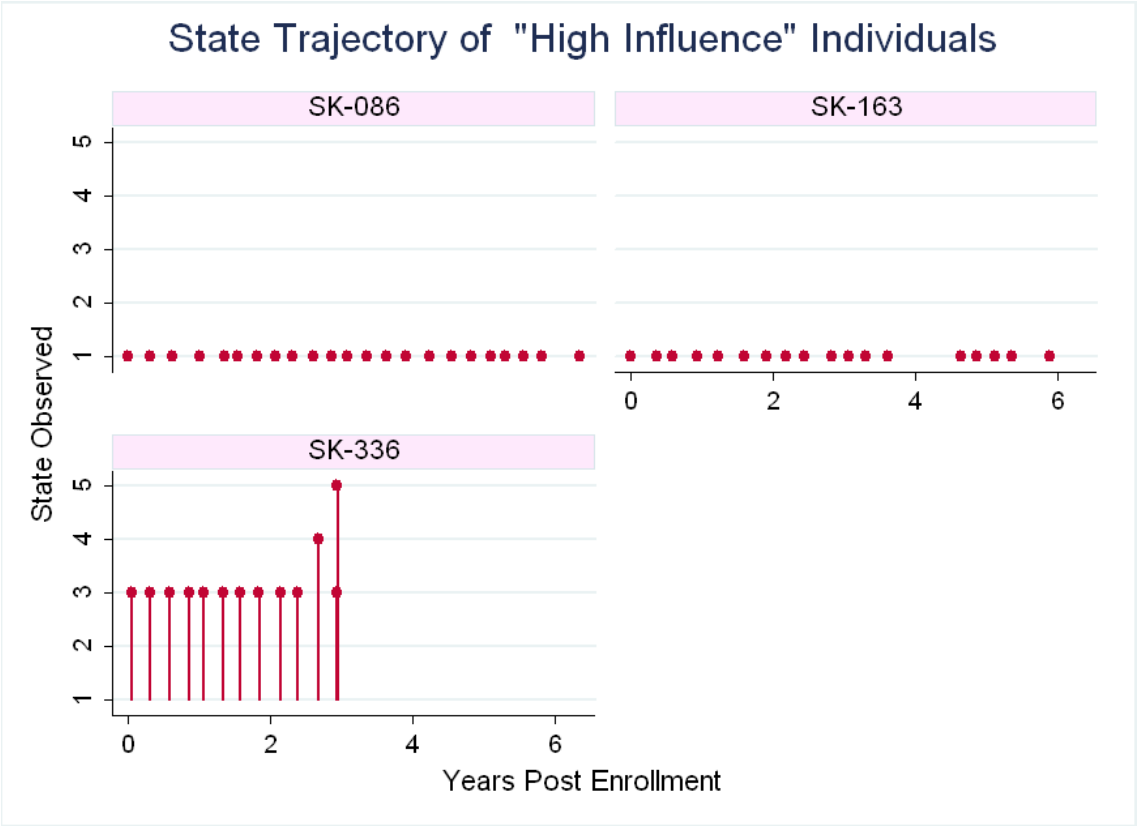

Figure 6.7: The state trajectory of "high influence individuals"

Upon removal of the three participants displaying a large influence on the likelihood, the model was refitted. The revised initial rate matrix $\mathbf{Q}_0$ is given by

$$\mathbf{Q}_0 = \begin{pmatrix} -0.988 & 0.988 & 0.000 & 0.000 & 0.000 \\ 0.561 & -1.580 & 1.019 & 0.000 & 0.000 \\ 0.000 & 0.514 & -0.971 & 0.396 & 0.060 \\ 0.000 & 0.000 & 0.599 & -1.213 & 0.614 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \end{pmatrix}$$

A comparison of the maximum likelihood estimates between the first model and the refitted model is presented in Table 6.4.

Table 6.4: Comparison of parameter estimates

Parameter	Model 1 estimate	Refitted model estimate
q_{12}	1.528	1.728
q_{21}	0.922	1.002
q_{23}	1.576	1.561
q_{32}	0.872	0.846
q_{34}	0.561	0.584
q_{35}	0.024	0.014
q_{43}	0.841	0.822
q_{45}	0.668	0.689
$-2\ln L$	5891.98	5836.05

Thus the removal of the three influential subjects had a substantial effect on model likelihood and transition intensity estimates. The transition intensities of the refitted model presented in Table 6.4 above, indicate that the rate of loss of CD4 cells declines as the CD4 count drops. This finding is consistent with that of Longini et al. (1991) who applied an eight state progressive model to data on HIV infected individuals from the US Army. The state structure used in Longini et al. (1991) is shown in Figure 6.8 where "OI" denotes the onset of an opportunistic infection.

Figure 6.8: The flow through eight states of infection as defined by Longini et al. (1991)

The Table 6.5 contains the mean sojourn time in each state expressed in years where mean sojourn times describe the average period in a single stay in a state. These are calculated as $\frac{-1}{q_{rr}}$ where q_{rr} is the r^{th} diagonal entry of the estimated transition intensity matrix $\mathbf{Q}$.

Table 6.5: Mean sojourn time in transient states

State	Estimate (Years)	Standard Error	95% CI
State 1	0.579	0.0462	(0.495, 0.677)
State 2	0.390	0.0201	(0.353, 0.431)
State 3	0.692	0.0367	(0.624, 0.768)
State 4	0.662	0.0548	(0.563, 0.779)

The interpretation of the mean sojourn times in states 1,2,3 and 4 are that an average stay in the CD4 count states are 0.58, 0.39, 0.69 and 0.66 years respectively. Thus we see that a typical individual just entering a CD4 count between 350 and 500, can expect to spend around five months (0.39 years) at that level before moving to either a higher or lower level of CD4 count. The mean sojourn time estimates the expected length of stay in a single stay in a state. It is also of interest, in the presence of reverse transitions, to estimate the *total* length of stay in each transient state. These estimates are presented in Table 6.6.

Table 6.6: Estimated total length of stay in transient states

State	Total length of stay (years)
State 1	2.07
State 2	2.57
State 3	3.56
State 4	1.38

The transition probability matrix $\mathbf{P}(t)$ is evaluated below at times 1, 2, 3, 4, 5 and 6 years. As expected, greater time is associated with higher probabilities for transitions to ARV initiation. It is of interest to note that as the years since enrollment increase, the probability of immune recovery and immune maintenance decrease. Immune recovery represents a transition from a state of lower CD4 count to a state of higher CD4 count. Immune maintenance refers to remaining in a particular state. Therefore the transitions representing immune recovery are

1. State $2 \longrightarrow 1$
2. State $3 \longrightarrow 2$
3. State $4 \longrightarrow 3$

Transitions representing immune deterioration are

1. State $1 \longrightarrow 2$
2. State $2 \longrightarrow 3$
3. State $3 \longrightarrow 4$

$$\mathbf{P}(1) = \begin{pmatrix} 0.329 & 0.336 & 0.27 & 0.052 & 0.014 \\ 0.195 & 0.298 & 0.37 & 0.098 & 0.037 \\ 0.085 & 0.202 & 0.44 & 0.172 & 0.098 \\ 0.023 & 0.074 & 0.24 & 0.283 & 0.378 \\ 0.000 & 0.000 & 0.00 & 0.000 & 1.000 \end{pmatrix}$$

$$\mathbf{P}(2) = \begin{pmatrix} 0.198 & 0.27 & 0.35 & 0.11 & 0.078 \\ 0.156 & 0.24 & 0.35 & 0.13 & 0.125 \\ 0.109 & 0.19 & 0.34 & 0.15 & 0.216 \\ 0.049 & 0.10 & 0.21 & 0.13 & 0.512 \\ 0.000 & 0.00 & 0.00 & 0.00 & 1.000 \end{pmatrix}$$

$$\mathbf{P}(3) = \begin{pmatrix} 0.149 & 0.224 & 0.33 & 0.127 & 0.17 \\ 0.130 & 0.204 & 0.32 & 0.129 & 0.22 \\ 0.105 & 0.172 & 0.29 & 0.124 & 0.31 \\ 0.056 & 0.098 & 0.17 & 0.085 & 0.59 \\ 0.000 & 0.000 & 0.00 & 0.000 & 1.00 \end{pmatrix}$$

$$\mathbf{P}(4) = \begin{pmatrix} 0.124 & 0.19 & 0.30 & 0.123 & 0.26 \\ 0.112 & 0.18 & 0.28 & 0.118 & 0.31 \\ 0.095 & 0.15 & 0.25 & 0.106 & 0.40 \\ 0.054 & 0.09 & 0.15 & 0.067 & 0.64 \\ 0.000 & 0.00 & 0.00 & 0.000 & 1.00 \end{pmatrix}$$

$$\mathbf{P}(5) = \begin{pmatrix} 0.107 & 0.17 & 0.27 & 0.112 & 0.34 \\ 0.098 & 0.16 & 0.25 & 0.105 & 0.39 \\ 0.085 & 0.14 & 0.22 & 0.093 & 0.47 \\ 0.050 & 0.08 & 0.13 & 0.056 & 0.68 \\ 0.000 & 0.00 & 0.00 & 0.000 & 1.00 \end{pmatrix}$$

$$\mathbf{P}(6) = \begin{pmatrix} 0.093 & 0.149 & 0.24 & 0.100 & 0.42 \\ 0.086 & 0.138 & 0.22 & 0.093 & 0.46 \\ 0.075 & 0.120 & 0.19 & 0.081 & 0.53 \\ 0.044 & 0.071 & 0.11 & 0.049 & 0.72 \\ 0.000 & 0.000 & 0.00 & 0.000 & 1.00 \end{pmatrix}$$

***Interpretation of the transition probability matrix*:**

- A typical individual presenting at the clinic with a CD4 count greater than 500, has a 35% chance of having a CD4 count between 200 and 350 and an 8% chance of being initiated on ARVs within two years. Six years later, however, this individual has a much higher probability of being initiated on ARVs (42%).
- An individual presenting at the clinic with a CD4 count between 350 and 500 has a 19.5% chance of experiencing immune recovery (an increase in CD4 Count) within a year. As time goes on, this chance of immune recovery decreases to 8.6% at six years.
- Prior to December 2009, the uptake of ARVs appears to be not as high as expected. This is apparent from the 50% probability of a patient currently presenting with a CD4 count less than 200, receiving ARV therapy within two years. There are many possible reasons for this. Although patients are referred to an ARV clinic when their CD4 counts are particularly low, they may not have adhered to this advice until they felt physically sick. The date of ARV initiation recorded could be much later than the actual initiation date. Clinicians may have been waiting for patients to present WHO Stage 3 or 4 symptoms before initiating them on ARVs, as the ARV rollout in South Africa was relatively poor in 2001.
- It is clear that transition probabilities from states 1, 2, 3, 4 into state 5 (ARV initiation) increase as t increases. For example $p_{45}(2) = 0.512$ while $p_{45}(6) = 0.72$.

Defining survival as "not initiating ARV therapy", we can plot the estimated survival functions from each of the four stages of infection.

$$S_{T_i}(t) = 1 - \hat{p}_{i5}(t)$$

where $i = 1, 2, 3, 4$

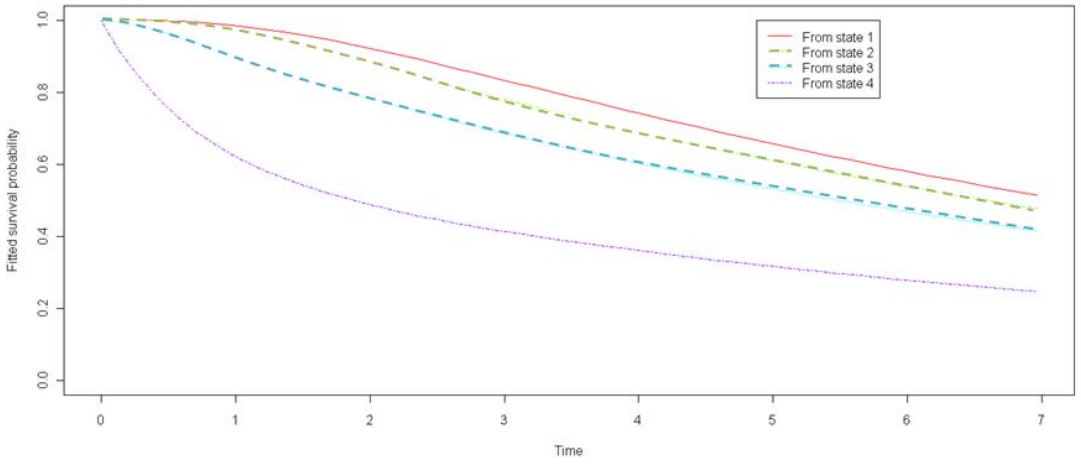

Figure 6.9: Estimated survival functions from each stage of infection

These functions are illustrated graphically in Figure 6.9.

As expected, the probability of ARV initiation within time t is highest for those with a previously observed CD4 count less than 200. We also see that as time increases, so too does the probability of initiating ARVs.

We now assess model fit by applying the methods introduced in Section 4.9. We assume that an individual's state at a time t was the same as the state at their previous observation time and that the process begins at a common time for all individuals. Then, given $n(t_i)$ individuals are under observation at time t_i, the expected number of individuals in state r at time t_i is $n(t_i)P_{1r}(t_i)$. The appropriateness of the former assumption was established by noting that observation times were relatively close together (Median 0.26 years). It is of interest to note that 16 patients were enrolled in the study without a CD4 count measurement who had three monthly measurements post enrollment. To prevent the loss of valuable followup data these patients were not excluded from the analysis. This is evident in the increase in observed total number of patients from time 0 to time 0.696 years.

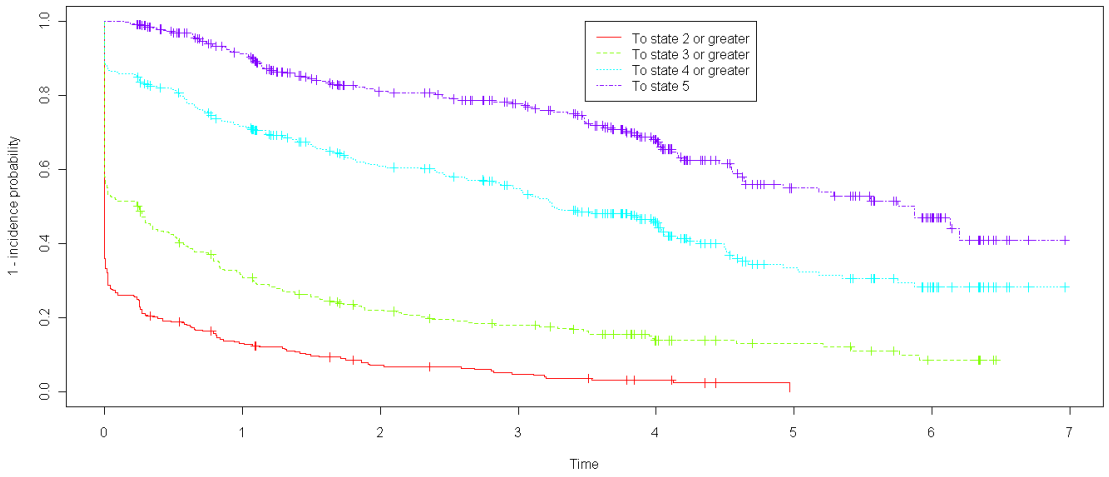

Figure 6.10: Kaplan-Meier incidence

The observed number of patients at each time point clearly demonstrate patterns in the data. As time increases the number of patients occupying the transient states 1, 2, 3 and 4 decreases and the number of patients initiating ARV therapy tends to increase. At approximately 7 years post enrollment there were 108 patients still under observation, 107 of whom had initiated ARV Therapy.

By examination of the expected frequencies in each state, areas of poor model fit can be identified. The observed counts are relatively close to the expected counts in the transient states, but tend to differ in the absorbing state (state 5). Using the observed and expected frequencies in each state, prevalences are calculated, which appear in Appendix D and are represented graphically in Figure 6.11.

Further investigation of the observed and expected prevalences in each state reveals that the predicted number of individuals who are initiated on ARVs (State 5) is underestimated by the model from about year 4 onwards. Similarly the number of individuals with CD4 count greater than 500 (State 1), State 2,

Table 6.7: Observed number of patients in each state evaluated at 10 equally spaced intervals

Time	State 1	State 2	State 3	State 4	State 5	Total
0	62	73	102	38	0	275
0.6965372907	53	81	99	41	17	291
1.3930745814	46	63	87	28	44	268
2.0896118721	38	62	78	24	56	258
2.7861491628	31	52	80	22	61	246
3.4826864535	20	46	71	20	74	231
4.1792237442	12	23	38	11	90	174
4.8757610349	8	12	23	10	98	151
5.5722983256	5	8	21	5	102	141
6.2688356163	2	8	3	0	107	120
6.965372907	0	1	0	0	107	108

Table 6.8: Expected number of patients in each state evaluated at 10 equally spaced intervals

Time	State 1	State 2	State 3	State 4	State 5	Total
0	62.0	73	102	38.0	1.7e-14	275
0.6965372907	50.4	73	106	41.0	2.1e+01	291
1.3930745814	39.6	61	93	37.1	3.8e+01	268
2.0896118721	33.8	53	83	34.0	5.4e+01	258
2.7861491628	29.1	46	73	30.3	6.7e+01	246
3.4826864535	24.8	40	63	26.4	7.7e+01	231
4.1792237442	17.1	27	44	18.3	6.8e+01	174
4.8757610349	13.6	22	35	14.6	6.6e+01	151
5.5722983256	11.6	19	30	12.5	6.9e+01	141
6.2688356163	9.0	14	23	9.8	6.4e+01	120
6.965372907	7.5	12	19	8.1	6.1e+01	108

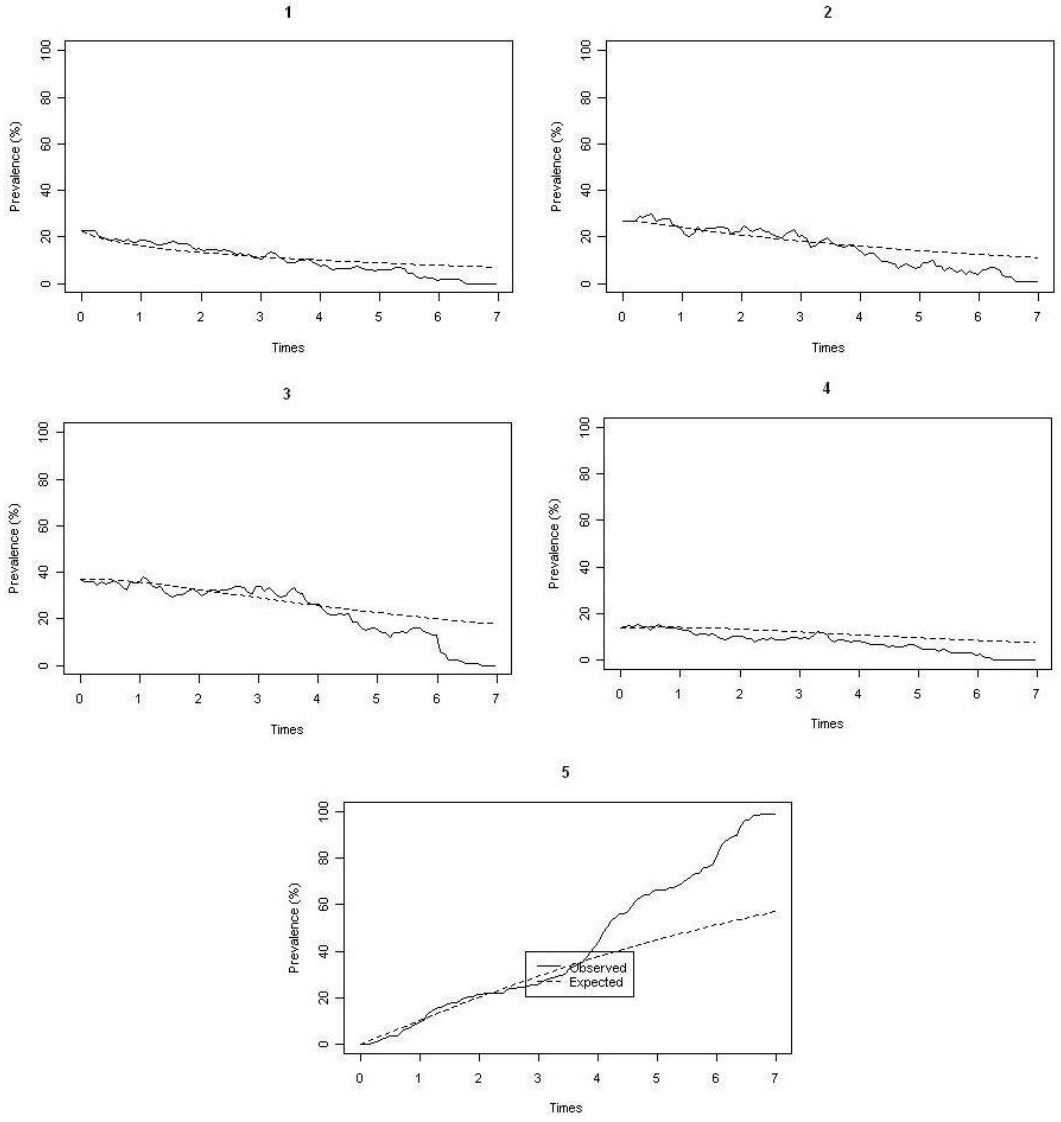

Figure 6.11: A comparison of observed and expected prevalences in the refitted model

State 3 and State 4 are slightly overestimated by the model from about year 4 onwards. Such discrepancies could be due to many factors. One possibility is that the transition rates vary with the time since the beginning of the process. Other factors include the age of the patient or some other omitted covariate, thereby making the Markov model non-homogeneous.

A piecewise model is fitted in the next section to assess the validity of the

assumption of time homogeneity.

6.3 Piecewise Model

In an attempt to improve model fit and assess the assumption of time homogeneity a non-homogenous Markov model was fitted to the data. We fitted a piecewise time homogeneous model with two distinct time intervals such that

$$\mathbf{Q}(t) = \begin{cases} \mathbf{Q}_0 & t < 4 \\ \mathbf{Q}_1 & t \geq 4 \end{cases}$$

Due to the limited information on transitions from state $3 \longrightarrow 5$ and state $4 \longrightarrow 5$ upon splitting the time into the two intervals, we fit a model assuming that q_{35} and q_{45} are fixed, and estimate the remaining parameters.

Table 6.9: Time non-homogeneous Markov model parameter Estimates

Parameter	Piecewise interval ($t < 4$)	Piecewise interval ($t \geq 4$)
q_{12}	1.776 (1.506, 2.094)	1.068 (0.622, 1.833)
q_{21}	1.011 (0.846, 1.208)	0.637 (0.364, 1.114)
q_{23}	1.620 (1.423, 1.844)	1.007 (0.657, 1.544)
q_{32}	0.901 (0.776, 1.046)	0.434 (0.265, 0.710)
q_{34}	0.539 (0.453, 0.640)	0.723 (0.249, 2.098)
q_{35}	0.063 (0.009, 0.426)	0.045 (0.000159, 1269)
q_{43}	0.779 (0.603, 1.005)	1.175 (0.646, 2.140)
q_{45}	0.569 (0.429, 0.753)	0.936 (0.205, 4.268)
$-2 \ln L$	5834.921	

We present a comparison of the transition intensity estimates between the two time intervals, through the computation of a hazard ratio.

It is evident from the ratios of transition intensities that four or more years post enrollment the risk of experiencing immune recovery (that is, reverse transi-

Table 6.10: A comparison of parameter estimates between the two intervals

Transition	Ratio of intensities	95% Confidence interval
$1 \rightarrow 2$	0.60	(0.34, 1.1)
$2 \rightarrow 1$	0.63	(0.35, 1.1)
$2 \rightarrow 3$	0.62	(0.40, 0.97)
$3 \rightarrow 2$	0.48	(0.29, 0.81)
$3 \rightarrow 4$	1.34	(0.44, 4.1)
$3 \rightarrow 5$	0.71	(0.00037, 1300)
$4 \rightarrow 3$	1.51	(0.80, 2.9)
$4 \rightarrow 5$	1.65	(0.27, 10)

tions) is significantly less likely than experiencing immune deterioration.

The extremely wide confidence interval observed for the estimate q_{35} is attributed to the "fixing" of this parameter in the model due to the limited data available on this transition in the second time interval. The addition of six parameters from piecewise model did not represent a significant improvement in log-likelihood from the time homogeneous model (p-value 0.98). The observed and expected prevalences computed from the piecewise model are presented in Appendix D.

6.4 Effect of Covariates on Transition Intensities

A commonly used indicator of disease progression is the HIV incubation period which is characterized by wide inter-individual variability (Sabin et al., 1998). This suggests the existence of cofactors, the effect of which may provide valuable information on HIV pathogenesis and progression.

The influence of demographic and genetic factors on HIV has been reviewed by Langford et al. (2007). A few key findings stated in this review were that

1. Older patients tend to progress more rapidly to an AIDS related illness.

2. Rate of CD4 cell decline is higher in patients over the age of 40 as opposed to those younger than 40.
3. Gender does not have a significant impact on HIV progression.
4. HIV Subtype may have a significant effect on progression but is difficult to confirm due to the many confounding factors including race and socio-economic class.
5. Lower levels of CD4 count, in particular less than 350 cells, are associated with a higher risk of disease progression.

The findings stated above, however, were derived by considering disease progression at entry into a single state that is, AIDS diagnosis. A question of interest however, is whether the effect of these factors differ at different stages of the disease. Such information can only be ascertained through the application of multistate models. Longini et al. (1991) examined the effect of age on HIV progression through the application of an eight state model based on levels of CD4 count. This study, based on a cohort of males from the US Army, found higher rates of CD4 count decline and faster disease progression in men over the age of thirty. This was only observed at lower levels of CD4 count, in particular less than 500 cells.

Alioum et al. (1998) examined the effect of age, gender, mode of transmission and ARV therapy on *clinical* progression of HIV/AIDS. States in this model were based on symptoms of the disease as opposed to the examination of immune deterioration through CD4 cell based stages. In this study it was noted that men tend to progress faster to AIDS diagnosis than females (HR 1.28 (1.16-1.43)) where HR denotes the hazard ratio of progression to AIDS.

We now explore the effect of three factors on the risk of immune deterioration and recovery in the Sinikithemba study. These factors are:

- Age at enrollment
- Gender

- CD4 count at enrollment

To establish an effective comparison with prior results, age and CD4 count were analyzed as a categorical variable. It is of note that upon removal of the three outlying individuals identified earlier, the total number of subjects analyzed is now 333.

Table 6.11: Cofactors analyzed and distribution of subjects within each level

Factor	Level	n (%)
Gender	Male	65 (19.52%)
	Female	268 (80.48%)
Age	≥ 30	195 (58.56%)
	< 30	138 (41.44%)
Baseline CD4 Count	≥ 350	162 (48.65%)
	< 350	171 (51.35%)

The transition intensities now depend on a vector of covariates $\mathbf{z}$. That is,

$$q_{ij}(t \mid \mathbf{z}) = q_{ij0}(t) \exp(\boldsymbol{\beta}'_{ij}\mathbf{z}) \tag{6.1}$$

where $\boldsymbol{\beta}_{ij} = (\beta_{ij1}, \beta_{ij2}, \ldots, \beta_{ijr})'$ is the vector of regression coefficients associated with vector $\mathbf{z}$ for the transition from state i to state j. $q_{ij0}(t)$ represents the baseline transition intensity function.

The effects of gender are presented through the use of hazard ratios, where female is the reference category (Table 6.12). We see that males tend to have slightly lower rates of both immune recovery and deterioration than females, however none of these are statistically significant.

Table 6.12: Effect of gender on transition rates

Transition	Hazard ratio	95% Confidence interval
$1 \rightarrow 2$	1.01	(0.671,1.5)
$2 \rightarrow 1$	0.89	(0.555,1.4)
$2 \rightarrow 3$	0.83	(0.595,1.2)
$3 \rightarrow 2$	0.76	(0.524,1.1)
$3 \rightarrow 4$	0.86	(0.559,1.3)
$3 \rightarrow 5$	0.49	(0.084,2.9)
$4 \rightarrow 3$	0.98	(0.548,1.7)
$4 \rightarrow 5$	0.65	(0.342,1.2)

The effect of age (reference category: age greater than or equal to 30) is presented in Table 6.13. We see that younger subjects experience lower rates of CD4 count decline compared to their older counterparts. We would expect higher rates of immune recovery in younger patients, but this phenomenon was not observed.

Table 6.13: Effect of age on transition rates

Transition	Hazard ratio	95% Confidence interval
$1 \rightarrow 2$	1.01	(0.73, 1.38)
$2 \rightarrow 1$	0.84	(0.59, 1.20)
$2 \rightarrow 3$	0.86	(0.67, 1.10)
$3 \rightarrow 2$	0.97	(0.73, 1.30)
$3 \rightarrow 4$	0.88	(0.64, 1.22)
$3 \rightarrow 5$	0.70	(0.24, 2.01)
$4 \rightarrow 3$	0.76	(0.47, 1.21)
$4 \rightarrow 5$	0.56	(0.35, 0.89)

Since patients enter the study at different levels of CD4 count it is crucial to control for this. We now present the effect of baseline CD4 count on state transitions. These results are presented in Table 6.15. It is clear that patients with a baseline CD4 count greater than 350 have higher rates of immune recovery and lower rates of immune deterioration. This effect is most pronounced on the transition from state 1 to state 2 (Hazard ratio 0.29 ((0.11, 0.75)). We also see that the hazard for recovering from state 2 to state 1 is twice that of patients with lower baseline CD4 count. These results support the argument for earlier initiation of ARV therapy which has been a great concern in recent years.

Table 6.14: Parameter estimates for the model with baseline CD4 count as a covariate

Parameter	Estimate (Baseline CD4 count ≥ 350)	Estimate (Baseline CD4 count < 350)
q_{12}	1.745 (1.482,2.054)	6.072 (2.373,15.540)
q_{21}	1.178 (0.991,1.400)	0.568 (0.237,1.360)
q_{23}	1.291 (1.114,1.495)	3.789 (2.926,4.906)
q_{32}	1.118 (0.935,1.338)	0.755 (0.575,0.991)
q_{34}	0.372 (0.280,0.493)	0.849 (0.707,1.018)
q_{35}	0.058 (0.040,0.084)	0.064 (0.036,0.114)
q_{43}	0.889 (0.365,2.165)	0.826 (0.543,1.258)
q_{45}	0.573 (0.498,0.659)	0.685 (0.549,0.855)

Defining survival as “not initiating ARV therapy”, we may plot the estimated survival functions from each of the four stages of infection stratified by baseline CD4 count. These curves are presented in Figures 6.12 and 6.13.

Table 6.15: Effect of baseline CD4 count on transition rates

Transition	Hazard ratio	95% Confidence interval
$1 \rightarrow 2$	0.29	(0.11, 0.75)
$2 \rightarrow 1$	2.07	(0.85, 5.05)
$2 \rightarrow 3$	0.34	(0.25, 0.46)
$3 \rightarrow 2$	1.48	(1.07, 2.05)
$3 \rightarrow 4$	0.44	(0.31, 0.61)
$3 \rightarrow 5$	0.91	(0.35, 2.35)
$4 \rightarrow 3$	1.08	(0.54, 2.14)
$4 \rightarrow 5$	0.84	(0.58, 1.20)

These curves clearly demonstrate that patients with baseline CD4 count less than 350 progress from all transient states to ARV initiation more rapidly than those with CD4 count greater than 350.

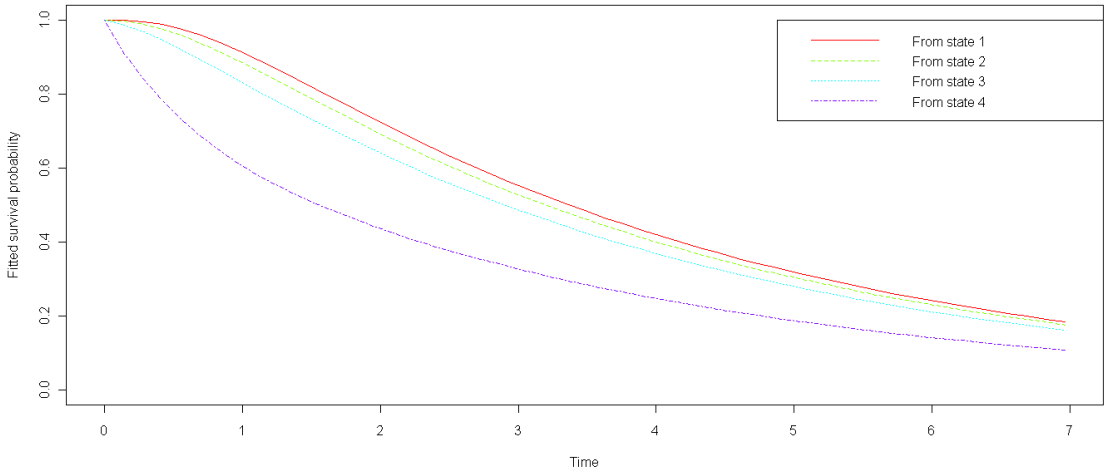

Figure 6.12: Survival curves extracted from the model where baseline CD4 count ≤ 350

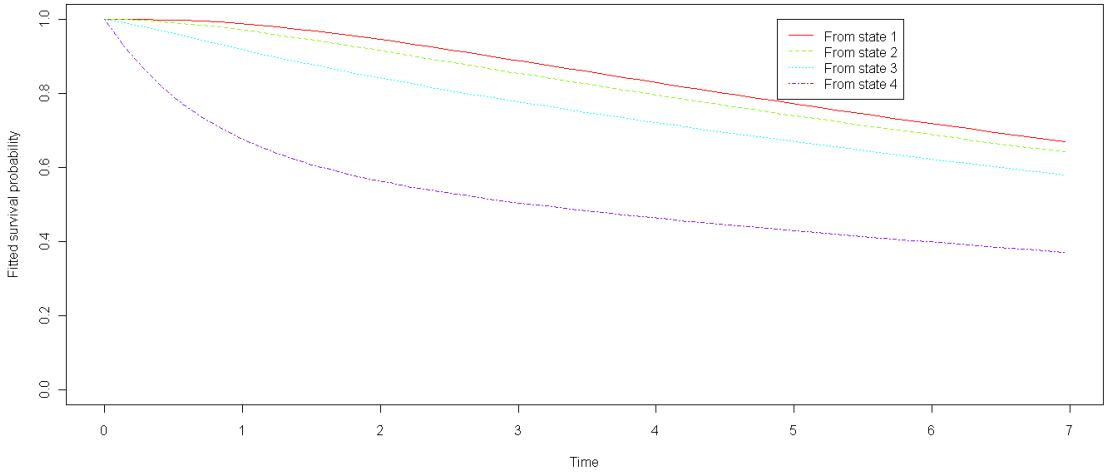

Figure 6.13: Survival curves extracted from the model where baseline CD4 count > 350

Chapter 7

Final Overview and Discussion

This thesis investigated the theory and recent developments in multistate models by drawing together well known existing contributions to the field. The primary objective was to apply this theory to describe the trajectory of HIV in ARV therapy naive South African individuals so as to aid the understanding of HIV progression and discover factors that influence immune deterioration. These findings reaffirm the need to initiate therapy at early stages of the virus which could lead to the development of new treatment strategies and save more lives.

Chapter 2 introduced multistate models and the preliminary concepts required for further chapters. The key components which are states, observation times, actions, rewards, transitions and constraints of the model were described. The two quantities which completely characterize a multistate process, that is transition probabilities and transition intensities, were defined in the general multistate framework. The various types of multistate models, and the assumptions governing each, were briefly introduced.

Chapter 3 explored time to *single event* analysis as it is this concept which forms the foundation of multistate models. Survival functions and the methods for estimation were described as well as the Cox proportional hazards model

and the construction of the partial likelihood. These methods were applied to assess survival in HIV-TB co-infected patients enrolled in the SAPiT trial. The analysis showed that patients on integrated therapy, that is TB and ARV treatment started at the same time interval) have significantly greater survival than those on sequential therapy. A Cox proportional hazards was fitted to the data with covariates age, gender, baseline CD4 count and weight. Patients who had a CD4 count less than 100 at baseline had a significantly poorer prognosis than those with CD4 counts between 350 and 500.

The time homogeneous Markov model was examined in Chapter 4. The Markov assumption and assumption of time homogeneity were explained and the mathematical properties of transition probability and intensities were stated. The computation of the transition probability matrix was discussed and demonstrated through the use of a practical example. The likelihood construction and iterative procedure for computing parameter estimates were outlined. Informal and formal diagnostics for model assessment, a relatively new concept in the multistate framework, were discussed. The Cox proportional hazards model for survival data was extended to the case of multistate models and the method of partial likelihood estimation was detailed. An alternative method for modeling multistate data, that is the additive hazards model, was briefly discussed.

Chapter 5 examined the parametric and non parametric methods available for the fitting of time dependent Markov models and other less "restrictive models". The parametric procedure, which involves piecewise constant intensities, was discussed as well as the nonparametric Aalen-Johansen estimator. The Aalen Johansen estimate was applied to the simple illness death model. The Semi Markov model which incorporates duration dependence and the hidden Markov and misclassification models were briefly discussed.

An in depth literature review was done. Limitations and strengths of prior studies were discussed. It was discovered that most of these studies examined

HIV progression in the United States and Europe. In Africa, however, the dynamics of HIV progression is rather different as Africa is afflicted by Subtype C HIV which is characterized by rapid progression. Thus more research on the topic with application of models to data from Africa is required.

The methods developed in Chapters 2, 3 and 4 were applied to data on HIV positive individuals in the Sinikithemba study. CD4 counts were taken at three month intervals and the median number of visits was twelve. The multistate approach is particularly useful to this type of problem because the data was characterized by a high degree of interval and right censoring. Furthermore, patients are at different points in disease progression as this was not an acute infection cohort. We have chosen a state structure based on four intervals of CD4 count and ARV initiation as an absorbing fifth state. An exploratory analysis of the data revealed that CD4 count is characterized by a high degree of fluctuation and may increase while a patient is not on ARV Therapy. This was addressed by including reverse transitions into the model. A key finding, consistent with previous research, is that the rate of decline in CD4 count tends to decrease at lower levels. It was also noted that patients enrolling with a CD4 count less than 350 had a far lower chance on immune recovery, and a substantially higher chance of immune deterioration compared to patients with higher CD4 cells. It was also observed that older patients tend to progress more rapidly through the disease than younger patients.

We have found multistate models to be a powerful tool in HIV/AIDS research which can offer a deeper understanding of the natural progression of the disease. The focus of this study was immune deterioration in ARV naive patients. It is of further interest to explore the rates of immune recovery in patients on antiretroviral therapy.

The heterogeneity between individuals cannot always be completely explained by observed covariates. In such cases this residual heterogeneity should be

modeled as a random effect. From a theoretical perspective it is of interest to accommodate frailty in multistate transition models, an area which is relatively new.

The CD4 count which was modeled in Chapter 6 is a marker which is subject to great variability and measurement error. An area of further work would be to formulate a Bayesian model which models two processes : the disease process (as a Markov process) and the measurement process.

Appendix A

Eigenvalues and Eigenvectors

The following definition illustrates how to find eigenvalues for a particular matrix.

Definition

Given a matrix $\mathbf{A}$, a non-zero vector $\mathbf{x}$ is defined to be an eigenvector of the $\mathbf{A}$ if it satisfies the eigenvalue equation $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ for some scalar λ. In this situation, the scalar λ is called an eigenvalue of $\mathbf{A}$ corresponding to the eigenvector $\mathbf{x}$.

The most important information to extract from the above definition is that

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}. \tag{1}$$

We can now form the characteristic equation using the above eigenvalue equation

$$\begin{aligned} \mathbf{A}\mathbf{x} - \lambda\mathbf{I}\mathbf{x} &= 0 \\ (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} &= 0 \end{aligned}$$

If there exists an inverse $(\mathbf{A} - \lambda\mathbf{I})^{-1}$, then both sides can be multiplied by the inverse to obtain the trivial solution $\mathbf{x} = 0$.

Thus we require there to be an inverse by assuming from linear algebra that the determinant equals zero

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0$$

The above equation is known as the characteristic equation, and enables us to find the eigenvalues.

Once the eigenvalues have been found, we return to Equation (1) and compute the eigenvector **x**.

Appendix B

Determinant of an $n \times n$ matrix

Let **A** be an $n \times n$ matrix such that

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \vdots \\ a_{n1} & a_{n2} & & a_{nn} \end{pmatrix}$$

It follows from Finney et al. (2003), we have

$$\det(\mathbf{A}) = |\mathbf{A}| = a_{11}\alpha_{11} + a_{12}\alpha_{12} + \cdots + a_{1n}\alpha_{1n}$$

where the coefficients α_{ij} are given by the relation $\alpha_{ij} = (-1)^{i+j}\beta_{ij}$ where β_{ij} is the determinant of the $(n-1)(\times(n-1)$ matrix that is obtained by deleting row i and column j. α_{ij} is commonly called the cofactor of a_{ij}.

Appendix C

Survival Analysis R Code

```
coxph(formula = surv.sapit ~ sapit$GENDER + sapit$AGE + sapit$arm +
    sapit$weigh + sapit$kat)


                       coef exp(coef) se(coef)      z      p
sapit$GENDERMale     0.5299     1.699    0.272  1.947 0.0510
sapit$AGE           -0.0129     0.987    0.017 -0.761 0.4500
sapit$armSequential  0.8179     2.266    0.256  3.192 0.0014
sapit$weighle 60     0.2488     1.282    0.278  0.897 0.3700
sapit$kat0-100       1.2885     3.627    0.602  2.140 0.0320
sapit$kat101-200     0.6295     1.877    0.631  0.997 0.3200
sapit$kat201-350    -0.3379     0.713    0.732 -0.462 0.6400

Likelihood ratio test=37.9  on 7 df, p=3.24e-06  n=640 (2 observations deleted

    Start:  AIC=742.63
survsap ~ GENDER + AGE + kat + arm + weigh

         Df    AIC
- AGE     1 741.22
- weigh   1 741.45
<none>      742.63
```

```
- GENDER  1 744.52
- arm     1 750.72
- kat     3 756.92

Step:  AIC=741.22
survsap ~ GENDER + kat + arm + weigh

         Df    AIC
- weigh   1 740.12
<none>      741.22
- GENDER  1 742.58
- arm     1 749.61
- kat     3 755.20

Step:  AIC=740.12
survsap ~ GENDER + kat + arm

         Df    AIC
<none>      740.12
- GENDER  1 741.55
- arm     1 749.12
- kat     3 754.80
Call:
coxph(formula = survsap ~ GENDER + kat + arm, data = sapit2)


               coef exp(coef) se(coef)      z       p
GENDERMale    0.478     1.613    0.261  1.829 0.06700
kat0-100      1.280     3.597    0.600  2.133 0.03300
kat101-200    0.630     1.878    0.630  1.001 0.32000
kat201-350   -0.361     0.697    0.731 -0.494 0.62000
```

```
armSequential  0.850     2.340    0.255  3.335 0.00085

Likelihood ratio test=36.4  on 5 df, p=8.05e-07  n= 640
```

```
    > cox.zph(cox1)
                   rho chisq     p
GENDERMale    -0.0684 0.294 0.587
kat0-100      -0.1473 1.342 0.247
kat101-200    -0.1434 1.290 0.256
kat201-350    -0.1924 2.303 0.129
armSequential  0.1480 1.365 0.243
GLOBAL             NA 3.841 0.572
```

Appendix D

R output of Multistate Commands

D.1 Computed score residuals

```
    > scoreresid.msm(skstudy.msm, plot=TRUE)
 SK-001  SK-002  SK-003  SK-004  SK-005  SK-006  SK-007  SK-008  SK-009 SK-010
0.00293 0.25838 0.04074 0.02542 0.03226 0.10625 0.04693 0.09229 0.45203 0.19095
 SK-011  SK-012  SK-013  SK-015  SK-016  SK-017  SK-019  SK-020  SK-021  SK-022
0.44591 0.06890 0.26082 0.34909 0.02719 0.06459 0.00559 0.03333 0.02674 0.04927
 SK-023  SK-024  SK-026  SK-028  SK-029  SK-030  SK-031  SK-032  SK-033  SK-034
0.06983 0.53907 0.63668 0.02408 0.65741 0.02013 0.50034 0.04063 0.00348 0.20129
 SK-035  SK-036  SK-037  SK-039  SK-040  SK-041  SK-042  SK-043  SK-044  SK-045
0.15004 0.34189 0.18938 0.02144 0.06683 0.06577 0.04500 0.03185 0.06101 0.11225
 SK-046  SK-047  SK-048  SK-049  SK-050  SK-051  SK-052  SK-054  SK-055  SK-058
0.54013 0.31729 0.04054 0.13154 0.78192 0.26141 0.02049 0.00467 0.23676 0.14360
 SK-060  SK-062  SK-063  SK-066  SK-067  SK-068  SK-069  SK-070  SK-071  SK-073
0.03914 0.06001 0.24729 0.29176 0.18282 0.36871 0.06559 0.13916 0.41163 0.09475
 SK-074  SK-077  SK-078  SK-079  SK-081  SK-083  SK-084  SK-085  SK-086  SK-088
0.04976 0.04397 0.02762 0.37260 0.43021 0.00559 0.02694 0.00117 1.56324 0.01521
 SK-090  SK-093  SK-095  SK-099  SK-100  SK-101  SK-102  SK-104  SK-105  SK-106
0.02938 0.08804 0.02376 0.01955 0.00573 0.44893 0.14868 0.04648 0.04222 0.08469
 SK-107  SK-108  SK-109  SK-112  SK-114  SK-116  SK-117  SK-119  SK-120  SK-122
0.03694 0.05069 0.11712 0.02501 0.34292 0.01754 0.10179 0.01710 0.04227 0.05009
 SK-123  SK-124  SK-126  SK-127  SK-128  SK-129  SK-131  SK-132  SK-134  SK-135
0.73821 0.02768 0.13608 0.01184 0.03172 0.02630 0.05515 0.04222 0.00398 0.19962
 SK-136  SK-137  SK-138  SK-139  SK-140  SK-141  SK-142  SK-143  SK-144  SK-145
0.08948 0.00120 0.23450 0.12434 0.02009 0.27703 0.14088 0.02045 0.01856 0.03397
 SK-146  SK-147  SK-148  SK-150  SK-151  SK-154  SK-155  SK-156  SK-159  SK-160
```

```
0.00742 0.03402 0.01980 0.34431 0.00406 0.04898 0.67666 0.04280 0.00831 0.02738
 SK-162  SK-163  SK-165  SK-166  SK-168  SK-169  SK-170  SK-171  SK-172  SK-173
0.17152 1.16031 0.19860 0.03959 0.01382 0.33411 0.52181 0.27779 0.02552 0.03894
 SK-174  SK-175  SK-176  SK-177  SK-178  SK-179  SK-180  SK-181  SK-182  SK-186
0.00827 0.09396 0.00610 0.04242 0.02636 0.15833 0.19957 0.07947 0.16323 0.06211
 SK-187  SK-188  SK-190  SK-191  SK-192  SK-193  SK-194  SK-195  SK-197  SK-200
0.23574 0.72433 0.01612 0.12092 0.03572 0.06551 0.16909 0.22891 0.07679 0.13988
 SK-201  SK-202  SK-203  SK-206  SK-207  SK-208  SK-212  SK-214  SK-215  SK-216
0.11541 0.26119 0.18142 0.02770 0.01997 0.72364 0.10118 0.06326 0.34287 0.01835
 SK-217  SK-218  SK-220  SK-221  SK-222  SK-223  SK-224  SK-226  SK-227  SK-228
0.03438 0.15476 0.10393 0.11107 0.03429 0.16962 0.09367 0.14431 0.04090 0.03412
 SK-231  SK-232  SK-233  SK-234  SK-236  SK-240  SK-243  SK-244  SK-246  SK-247
0.02735 0.62970 0.01110 0.01488 0.03296 0.00143 0.01273 0.19935 0.14101 0.06462
 SK-248  SK-249  SK-250  SK-252  SK-254  SK-255  SK-256  SK-257  SK-258  SK-259
0.45533 0.11065 0.05438 0.04156 0.06028 0.02648 0.03774 0.38844 0.04241 0.10189
 SK-263  SK-265  SK-267  SK-268  SK-272  SK-273  SK-274  SK-276  SK-281  SK-283
0.16854 0.78246 0.01695 0.04543 0.23378 0.11760 0.02812 0.11922 0.28757 0.06219
 SK-285  SK-287  SK-288  SK-289  SK-291  SK-293  SK-294  SK-295  SK-298  SK-302
0.06796 0.15331 0.00803 0.67043 0.13429 0.01188 0.03821 0.24786 0.04670 0.17144
 SK-305  SK-306  SK-307  SK-308  SK-309  SK-311  SK-312  SK-313  SK-316  SK-318
0.29214 0.23058 0.47089 0.06272 0.02043 0.03883 0.22594 0.62447 0.29869 0.09310
 SK-319  SK-320  SK-321  SK-322  SK-323  SK-324  SK-325  SK-327  SK-328  SK-329
0.03811 0.53299 0.61851 0.07560 0.06729 0.25339 0.19743 0.10546 0.03814 0.23934
 SK-330  SK-331  SK-332  SK-333  SK-334  SK-335  SK-336  SK-337  SK-338  SK-339
0.02975 0.09080 0.01681 0.01573 0.40087 0.04341 0.92570 0.49179 0.30047 0.24636
 SK-340  SK-341  SK-343  SK-344  SK-345  SK-346  SK-347  SK-349  SK-350  SK-351
0.00369 0.28989 0.08256 0.07692 0.00109 0.08441 0.07424 0.15253 0.01780 0.18700
 SK-352  SK-353  SK-355  SK-356  SK-357  SK-358  SK-359  SK-360  SK-361  SK-363
0.10447 0.25656 0.00242 0.09710 0.21901 0.06861 0.47352 0.16581 0.08292 0.09490
 SK-364  SK-365  SK-367  SK-368  SK-370  SK-371  SK-372  SK-373  SK-374  SK-375
0.23844 0.19576 0.13858 0.09159 0.30077 0.09290 0.13208 0.03528 0.15651 0.06836
 SK-376  SK-377  SK-379  SK-380  SK-381  SK-382  SK-383  SK-384  SK-385  SK-386
0.05650 0.05478 0.09873 0.19841 0.31742 0.20562 0.18061 0.21492 0.55954 0.21394
 SK-387  SK-388  SK-390  SK-391  SK-393  SK-394  SK-395  SK-396  SK-397  SK-398
0.05284 0.15797 0.00298 0.23370 0.06292 0.05212 0.23773 0.07442 0.05941 0.11073
 SK-399  SK-400  SK-401  SK-402  SK-403  SK-404  SK-405  SK-406  SK-409  SK-410
0.09791 0.11922 0.04734 0.22429 0.05115 0.48405 0.14617 0.24849 0.22261 0.53928
 SK-411  SK-412  SK-413  SK-415  SK-416  SK-417  SK-418  SK-420  SK-421  SK-422
0.17612 0.48989 0.10683 0.02517 0.06534 0.13149 0.08561 0.19203 0.27926 0.07753
 SK-423  SK-424  SK-425  SK-426  SK-428  SK-430  SK-432  SK-433  SK-434  SK-435
0.05446 0.08918 0.12777 0.10823 0.06242 0.48182 0.11258 0.41426 0.08504 0.09720
 SK-436  SK-437  SK-438  SK-439  SK-440  SK-441  SK-442  SK-443  SK-444  SK-445
0.05298 0.22820 0.03478 0.02043 0.22853 0.00349 0.00866 0.09123 0.09749 0.04188
 SK-446  SK-447  SK-448  SK-449  SK-450  SK-451
0.03199 0.25631 0.27815 0.07165 0.26249     NaN
```

D.2 Observed and expected prevalence from the time homogeneous Markov model

$Observed

	State 1	State 2	State 3	State 4	State 5	Total
0	62	73	102	38	0	275
0.6965372907	53	81	99	41	17	291
1.3930745814	46	63	87	28	44	268
2.0896118721	38	62	78	24	56	258
2.7861491628	31	52	80	22	61	246
3.4826864535	20	46	71	20	74	231
4.1792237442	12	23	38	11	90	174
4.8757610349	8	12	23	10	98	151
5.5722983256	5	8	21	5	102	141
6.2688356163	2	8	3	0	107	120
6.965372907	0	1	0	0	107	108

$Expected

	1	2	3	4	5	Total
0	62.0	73	102	38.0	1.7e-14	275
0.6965372907	50.4	73	106	41.0	2.1e+01	291
1.3930745814	39.6	61	93	37.1	3.8e+01	268
2.0896118721	33.8	53	83	34.0	5.4e+01	258
2.7861491628	29.1	46	73	30.3	6.7e+01	246
3.4826864535	24.8	40	63	26.4	7.7e+01	231
4.1792237442	17.1	27	44	18.3	6.8e+01	174
4.8757610349	13.6	22	35	14.6	6.6e+01	151
5.5722983256	11.6	19	30	12.5	6.9e+01	141
6.2688356163	9.0	14	23	9.8	6.4e+01	120

```
6.965372907   7.5 12  19  8.1 6.1e+01   108
```

```
$`Observed percentages`
             State 1 State 2 State 3 State 4 State 5
0               22.5   26.55    37.1    13.8     0.0
0.6965372907    18.2   27.84    34.0    14.1     5.8
1.3930745814    17.2   23.51    32.5    10.4    16.4
2.0896118721    14.7   24.03    30.2     9.3    21.7
2.7861491628    12.6   21.14    32.5     8.9    24.8
3.4826864535     8.7   19.91    30.7     8.7    32.0
4.1792237442     6.9   13.22    21.8     6.3    51.7
4.8757610349     5.3    7.95    15.2     6.6    64.9
5.5722983256     3.5    5.67    14.9     3.5    72.3
6.2688356163     1.7    6.67     2.5     0.0    89.2
6.965372907      0.0    0.93     0.0     0.0    99.1
```

```
$`Expected percentages`
                1  2  3    4       5
0            22.5 27 37 13.8 6.2e-15
0.6965372907 17.3 25 36 14.1 7.1e+00
1.3930745814 14.8 23 35 13.8 1.4e+01
2.0896118721 13.1 21 32 13.2 2.1e+01
2.7861491628 11.8 19 30 12.3 2.7e+01
3.4826864535 10.8 17 27 11.4 3.3e+01
4.1792237442  9.8 16 25 10.5 3.9e+01
4.8757610349  9.0 14 23  9.7 4.4e+01
5.5722983256  8.2 13 21  8.9 4.9e+01
6.2688356163  7.5 12 19  8.1 5.3e+01
6.965372907   6.9 11 18  7.5 5.7e+01
```

D.3 Observed and expected prevalence from the piecewise Markov model

```
> prevalence.msm(skstudyp.msm)
$Observed
             State 1 State 2 State 3 State 4 State 5 Total
0                 62      73     102      38       0   275
0.6965372907      53      81      99      41      17   291
1.3930745814      46      63      87      28      44   268
2.0896118721      38      62      78      24      56   258
2.7861491628      31      52      80      22      61   246
3.4826864535      20      46      71      20      74   231
4.1792237442      12      23      38      11      90   174
4.8757610349       8      12      23      10      98   151
5.5722983256       5       8      21       5     102   141
6.2688356163       2       8       3       0     107   120
6.965372907        0       1       0       0     107   108

$Expected
                     1        2         3         4            5 Total
0            62.000000 73.00000 102.00000 38.000000 2.672526e-14   275
0.6965372907 49.778120 73.62382 104.99113 41.620015 2.098691e+01   291
1.3930745814 39.053748 61.08357  91.40493 37.835529 3.862222e+01   268
2.0896118721 33.295520 53.23069  81.63171 34.785660 5.505642e+01   258
2.7861491628 28.595072 46.17481  71.65576 31.051745 6.852261e+01   246
3.4826864535 24.380313 39.55811  61.74606 27.001613 7.831391e+01   231
4.1792237442 16.972008 26.49836  43.03666 17.782640 6.971034e+01   174
```

```
4.8757610349 13.617673 19.93026  33.86822 12.855163 7.072869e+01   151
5.5722983256 11.494032 16.54495  28.13478 10.525754 7.430049e+01   141
6.2688356163  8.784387 12.56620  21.27834  7.933176 6.943789e+01   120
6.965372907   7.079614 10.09563  17.03485  6.341898 6.744800e+01   108

$`Observed percentages`
               State 1   State 2  State 3   State 4   State 5
0            22.545455 26.545455 37.09091 13.818182  0.000000
0.6965372907 18.213058 27.835052 34.02062 14.089347  5.841924
1.3930745814 17.164179 23.507463 32.46269 10.447761 16.417910
2.0896118721 14.728682 24.031008 30.23256  9.302326 21.705426
2.7861491628 12.601626 21.138211 32.52033  8.943089 24.796748
3.4826864535  8.658009 19.913420 30.73593  8.658009 32.034632
4.1792237442  6.896552 13.218391 21.83908  6.321839 51.724138
4.8757610349  5.298013  7.947020 15.23179  6.622517 64.900662
5.5722983256  3.546099  5.673759 14.89362  3.546099 72.340426
6.2688356163  1.666667  6.666667  2.50000  0.000000 89.166667
6.965372907   0.000000  0.925926  0.00000  0.000000 99.074074

$`Expected percentages`
                     1         2        3         4             5
0            22.545455 26.545455 37.09091 13.818182 9.718276e-15
0.6965372907 17.105883 25.300282 36.07943 14.302411 7.211997e+00
1.3930745814 14.572294 22.792376 34.10632 14.117735 1.441128e+01
2.0896118721 12.905240 20.632049 31.64020 13.482814 2.133970e+01
2.7861491628 11.624013 18.770248 29.12836 12.622660 2.785472e+01
3.4826864535 10.554248 17.124723 26.72989 11.689010 3.390212e+01
4.1792237442  9.754028 15.228942 24.73371 10.219908 4.006341e+01
4.8757610349  9.018327 13.198846 22.42928  8.513353 4.684019e+01
5.5722983256  8.151796 11.734008 19.95374  7.465074 5.269538e+01
6.2688356163  7.320323 10.471836 17.73195  6.610980 5.786491e+01
```

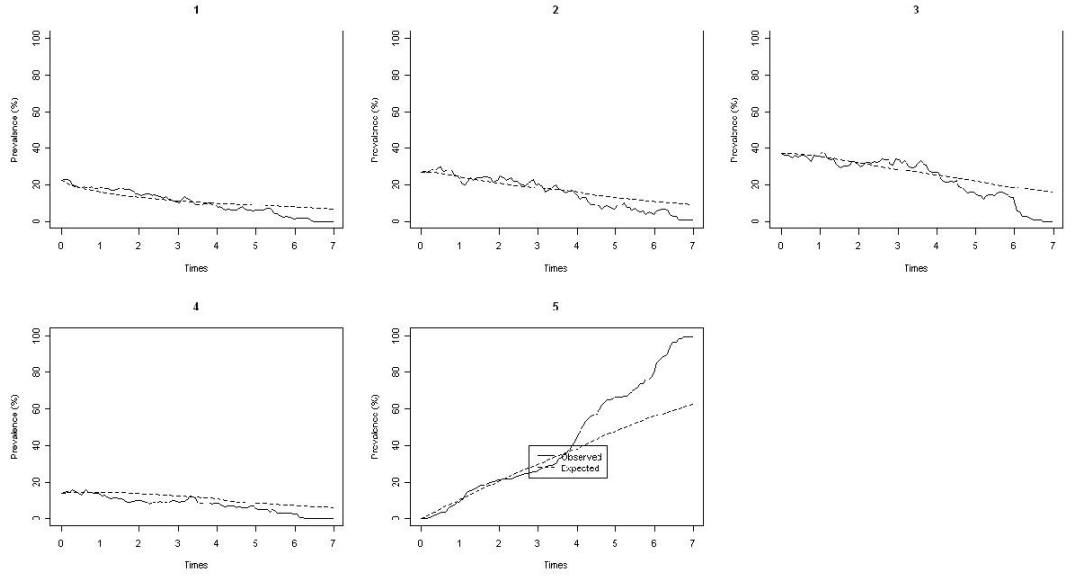

Figure D.1: Prevalence plot computed from the piecewise model

```
6.965372907   6.555198  9.347807 15.77301  5.872128 6.245185e+01
```

Bibliography

Aalen, O. O. (1980). A model for non-parametric regression analysis of counting processes. *Lecture Notes in Statistics* **2,** 1–20.

Aalen, O. O. (2008). *Survival and event history analysis: a process point of view*. Springer.

Abdool Karim, S. S., Naidoo, K., Grobler, A., and Padayatchi, N. (2010). Timing of initiation of antiretroviral drugs during tuberculosis therapy. *New England Journal of Medicine* **362,** 697–706.

Aguirre-Hernández, R. and Farewell, V. T. (2002). A Pearson type goodness-of-fit test for stationary and time-continuous Markov regression models. *Statistics in Medicine* **21,** 1899–1911.

Alioum, A., Leroy, V., Commenges, D., Dabis, F., and Salamon, R. (1998). Effect of gender, age, transmission category, and antiretroviral therapy on the progression of HIV infection using Multistate Markov Models. *Epidemiology* **9,** 605–612.

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

Borgan, Ø. (1997). Three contributions to the Encyclopedia of Biostatistics:The Nelson-Aalen, Kaplan-Meier and Aalen-Johansen estimators.

Bwayo, J., Nagelkerke, N., Moses, S., Embree, J., Ngugi, E., Mwatha, A., Kimani, J., Anzala, A., Choudhri, S., and Achola, J. (1995). Comparison of the declines in CD4 counts in HIV-1-seropositive female sex workers and women

from the general population in Nairobi, Kenya. *Journal of Acquired Immune Deficiciency Syndromes and Human Retrovirology* **10,** 457–461.

Collett, D. (2003). *Modelling survival data in medical research*. Chapman and Hall.

Cox, D. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society* **34,** 187–220.

Crowe, S., Hoy, J., and Mills, J. (1996). *Management of the HIV-infected patient*. CUP Archive, Health and Fitness.

Elston, R. C., Olson, J. M., and Palmer, L. (2002). *Biostatistical Genetics and Genetic Epidemiology*. John Wiley and Sons.

Finney, R., Thomas, G., Weir, M., and Giordano, F. (2003). *Thomas' calculus*. Addison Wesley.

Fiocco, M., Putter, H., and van Houwelingen, H. C. (2008). Reduced-rank proportional hazards regression and simulation-based prediction for multi-state models. *Statistics in Medicine* **27,** 4340–4358.

Geng, J., Xu, D., Gong, J., and Li, W. (1998). Assessing hepatitis A virus epidemic stochastic process in eight cities in China in 1990. *International Journal of Epidemiology* **27,** 320–322.

Gentleman, R. C., Lawless, J. F., Lindsey, J. C., and Yan, P. (1994). Multi-state Markov Models for analysing incomplete disease history data with illustrations for HIV disease. *Statistics in Medicine* **13,** 805–821.

Hosmer, D. W. and Royston, P. (2002). Using Aalen's linear hazards model to investigate time-varying effects in the proportional hazards regression model. *The Stata Journal* **2,** 331–350.

Hougaard, P. (2002). *Multivariate Survival Data*. Springer.

Jackson, C. (2007). Multi-state modelling with R: the msm package.

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

Karlsson, M. (2000). Hidden Markov models.

Kay, R. (1982). The analysis of transition times in multistate stochastic processes using proportional hazard regression models. *Communications in Statistics-Theory and Methods* **11,** 1743–1756.

Kay, R. (1986). A Markov Model for Analysing Cancer Markers and Disease States in Survival Studies. *Biometrics* **42:4,** 855–865.

Kulkarni, V. G. (1995). *Modelling and Analysis of Stochastic Systems*. Chapman and Hall.

Langford, S., Ananworanich, J., and Cooper, D. (2007). Predictors of disease progression in HIV infection: a review. *AIDS Research and Therapy* **4,** 11.

Lee, E. T. and Wang, J. W. (2003). *Analysis of Multivariate Survival Data*. John Wiley and Sons.

Longini, I. M., Clark, W. S., Byers, R. H., Ward, J. W., Darrow, W. W., Lemp, G. F., and Hethcote, H. W. (1989). Statistical analysis of the stages of HIV infection using a Markov model. *Statistics in Medicine* **8,** 831–843.

Longini, I. M., Clark, W. S., Gardner, L. I., and Brundage, J. F. (1991). The Dynamics of CD4+ T-Lymphocyte decline in HIV-Infected Individuals: A Markov Modeling Approach. *Journal of Acquired Immune Deficiency Syndromes* **4,** 1141–1147.

Malone, J. L., Simms, T. E., Gray, G. C., Wagner, K. F., Bruge, R. J., and Burke, D. S. (1990). Sources of Variability in Repeated T-helper Lymphocyte counts from Human Immunodeficiency Virus Type 1-Infected patients: Total Lymphocyte Count Fluctuations and Diurnal Cycle are important. *Journal of Acquired Immune Deficiency Syndromes* **3,** 144–151.

Marshall, A. W. and Goldhamer, H. (1955). An Application of Markov Processes to the study of the Epidemiology of Mental Disease. *Journal of the American Statistical Association* **50,** 99–129.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall, second edition.

Meira-Machado, L., de Uña Álvarez, J., Cadarso-Suárez, C., and Andersen, P. K. (2007). Multi-state models for the analysis of time to event data. Technical report, University of Copenhagen.

Meira-Machado, L., de Uña Álvarez, J., Cadarso-Suárez, C., and Andersen, P. K. (2009). Multi-state models for the analysis of time-to-event data. *Statistics in Medicine* **18,** 195–222.

Mullins, D. C. (1996). A Simplified Approach to Teaching Markov Models. *American Journal of Pharmaceutical Education* **60,** 42–47.

Pérez-Ocón, R., Ruiz-Castro, J. E., and Gámiz-Pérez, M. L. (2001). A piecewise Markov process for analysing survival from breast cancer in different risk groups. *Statistics in Medicine* **20,** 109–122.

Putter, H., Fiocco, M., and Geskus, R. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine* **26,** 2389–2340.

Sabin, C. A., Mocroft, A., Cozzi Lepri, A., and Phillips, A. N. (1998). Cofactors and marker of disease progression in human immunodeficiency virus infection. *Journal of the Royal Statistical Society* **161,** 177–189.

Schaubel, D. E., Morrison, H. I., Desmeules, M., Parsons, D., and Fenton, S. S. (1998). End-stage renal disease projections for Canada to 2005 using Poisson and Markov models. *International Journal of Epidemiology* **27,** 274–281.

Statistics (2007). Random processes. UKZN.

Titman, A. C. (2007). *Model diagnostics in multi-state models of biological systems*. PhD thesis, University of Cambridge.

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

UNAIDS (2008). Epidemiological fact sheet on HIV and AIDS, core data on epidemiology and response, South Africa. Technical report.

UNAIDS (2009). UNAIDS epidemic update. Technical report.