Multi-state models and joint models: a comparison using AIDS data

20132424

Supervisor: Dr Ardo van den Hout

Department of Statistical Science
University College London

Word count: 9148

July 2021

Contents

1 Introduction

Survival analysis or time-to-event analysis comprises of a set of methods which analyse the length of time until a well defined end point of interest occurs. Examples of end points in medical research include heart attacks, disease remission or most typically, death. In survival analysis, the outcome of interest has both an event and time associated with it. For example, if the event of interest is death, the survival time is the number of months or years until the individual experiences death. This is unlike a typical regression problem where the outcome of interest may be a single continuous variable such as house price. Thus, data of this type have distinguishing features and do not follow typical distributions like other types of data would. Survival data are typically non-negative and skewed. For example a large number of events occuring after an incident, such as patient deaths after an operation, will result in a skewed distribution.

One of the prominent characteristics of survival data is censoring. Censoring is a form of missing data, where time to event is not known or observed. A patient may not experience the event at the end of the study. For example, in the case where the event of interest is a death, an individual may survive the length of the study and not die. Individuals may also be lost to follow-up over the duration of the study or they may withdraw. All of these scenarios result in missing data, where the event time is unknown. Interval censored data arises where the time to event of interest cannot be observed, but is known to lie in an interval obtained from a sequence of follow-up times during the study. Regardless, censoring must be accounted for in the modelling procedure to ensure valid inference.

There are many approaches to survival analysis and they can broadly be split into three main groups. This includes non-parametric, semi-parametric and parametric. Non parametric approaches assume no assumptions about the underlying distribution of data whilst parametric approaches include using a statistical distribution to estimate survival times. The modelling carried out in this thesis all fall under the parametric group.

More specifically, the primary model explored in this thesis is the multi-state model which allows the study of disease states and mortality simultaneously. Multi-state models are widely used in biostatistics and disease modelling, where diseases often present as a series of progressive states. This modelling framework is useful in the case of HIV patients, where drugs and the effect of other covariates can be investigated to inform practitioners of the trajectory of the disease. There has been studies into this specifically, such as modelling HIV's immunological markers to the onset of AIDS [1]. Other examples of studies which employ the use of multi-state models include the rejection of lung transplant recipients [2], diabetic retinopathy [3], and cervical cancer screening [4].

With the growing availability of health data, there have been efforts to pursue personalised medicine. Multi-state models provide a substantial opportunity to gain a deeper understanding of the disease process, and how it can change over time. This thesis aims to introduce this framework and demonstrate its use by applying it to a dataset. First, the data are introduced. Following this, the underlying theory which uphold the multi-state model is introduced before illustrating the model using the dataset. Finally, the model is briefly compared with the joint modelling framework. The comparison of these modelling techniques aims to help the reader consider the possibilities of alternative methods. More importantly, it serves the purpose of highlighting the strengths and weaknesses of both models; in the effort to help inform the reader when considering the use of these different statistical techniques in medical research.

2 Data

Patients who are diagnosed with human immunodeficiency virus (HIV) infection are typically treated with zidovudine therapy. However, second line therapies such as didanosine and zalcitabine are also used if patients display a failure or intolerance to zidovudine therapy. The data used were taken from a randomised trial carried out by the Terry Beirn Community Programs for Clinical Research on AIDS. Patients who were unresponsive to zidovudine therapy were randomly assigned to receive daily didanosine or zalcitabine treatments. CD4 lymphocyte cell counts were recorded at the start of the trial and up to 4 time periods later [5].

In total, there were 467 patients enrolled into the study. 237 patients were administered zalcitabine, and 230 patients were administered didanosine treatments. Patients provided CD4 counts at fixed follow-up times which were at 2, 6, 12 and 18 months. Therefore the time intervals between check-ups were not equidistant. Attendance rates at check-up visits varied from 86% to 95%. There was a minimum of 12 months follow up time. The median follow-up time was 6 months. This is a measure that aims captures how long patients were followed up on average. The value of 6 months was obtained by considering both survival and censored times and computing the median. By the end of the study 188 patients had died. 1,405 measurements were taken out of the 2,335 planned measurements (467 patients multiplied by 5 measurements) [6]. The data are loaded into the programming language `R`, which is the language used throughout this report for statistical computation and graphics. They are loaded from the `JM` package titled `'aids'`.

The CD4 lymphocyte cell count is commonly used as a measurement tool to forecast the progression of HIV disease and to direct antiretroviral treatment. Using this cell count as a surrogate marker can provide insight into the association between treatment, CD4 cell counts and clinical outcome such as disease progression or death. Hence, living states were defined by the CD4 cell counts to model trajectory of the disease. CD4 counts give an indication on the health of the immune system. Lower CD4 counts are generally associated with a weakened immune system, which is typically treated with the necessary HIV treatment. If CD4 counts were less than $5mm^{-3}$, patients were assigned to State 4. If CD4 counts were between $5 - 10mm^{-3}$, State 3 was assigned. If CD4 cell counts were between $10 - 15mm^{-3}$, State 2 was assigned. If CD4 cell counts were above $15mm^{-3}$, State 1 was assigned. State 5 was defined as death. This follows the convention that lower states are typically healthier states and therefore patients in lower states will tend to exhibit higher CD4 counts. Censored states were assigned to -2. The states are depicted in Figure 1. However, it is important to note that although the states have been discretised in this way for this thesis, other choices are possible.

Most observations were that of being in the same state at successive interval times, as evidenced by largest numbers appearing on the diagonals in Table 1. This indicates that the progression of the disease is slow relative to the timing of check-ups. However, forward transitions were also observed. It is interesting to note that these forward transitions were observed more

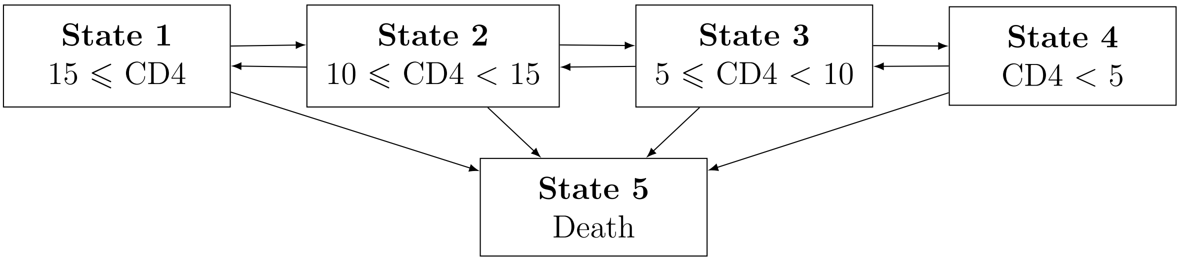

Figure 1 *Five-state model for the levels of CD4 count observed in HIV patients.*

Table 1 *State table for CD4 data: a frequency table showing the number of times each pair of states were observed at successive observation times.*

	To State					
From State	**1**	**2**	**3**	**4**	**5**	**Right Censored**
1	79	39	4	2	3	33
2	25	86	56	7	5	47
3	3	17	179	104	59	72
4	1	3	39	294	121	127

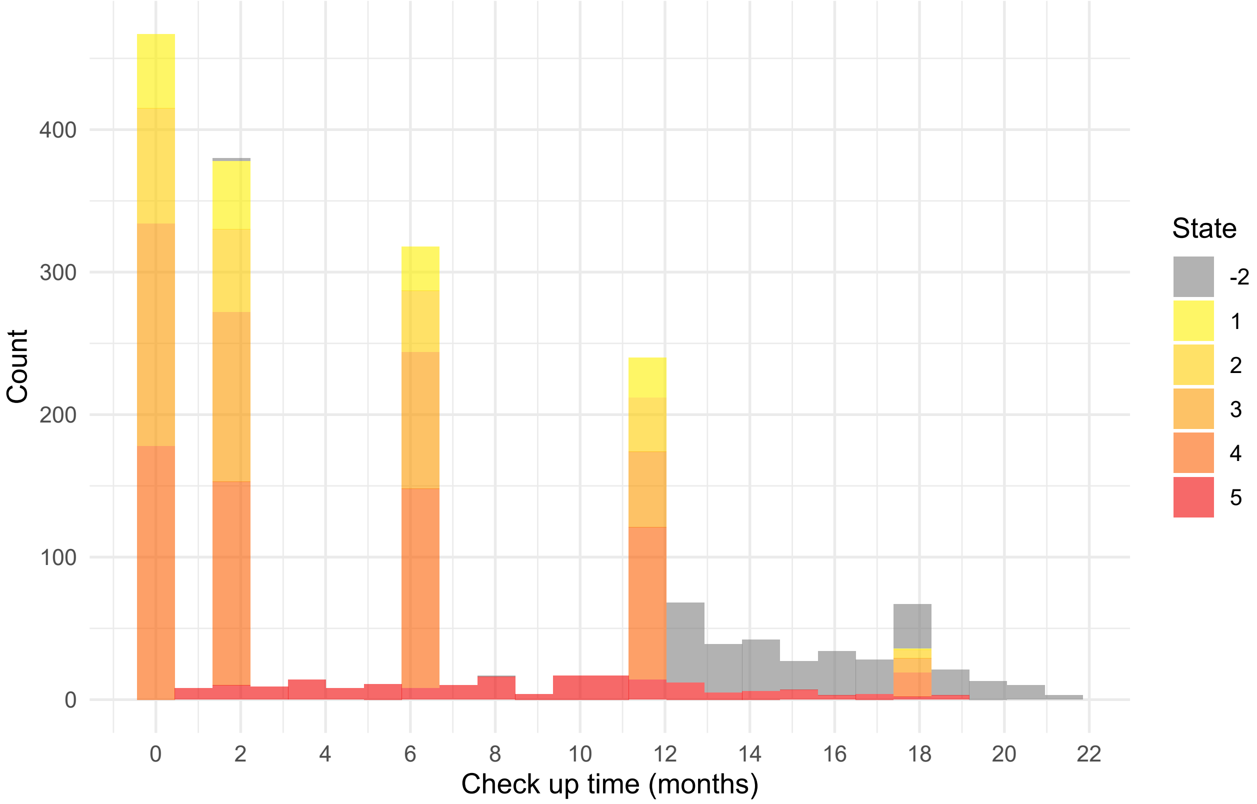

Figure 2 *Graph showing the observed states at various check up times.*

frequently than equivalent backward transitions. This implies that for most HIV patients the nature of disease is a progressive deterioration of health over time, regardless of the type of drug. However, some backward transitions are recorded, albeit less so. This could indicate the positive affect of the administered drugs. Column **5** sums to the total number of deaths. Note that the death state is absorbing, and therefore the frequency for that state will only increase. 121 patients entered the death state after being in State 4, whilst only 3 patients entered the death state after being in State 1. Overall, the state table implies that most patients exhibit a decline in CD4 count and consequently a decline in health. In addition, the most censoring was seen from patients in State 4 and the least censoring in State 1 at the next successive observation time.

Figure 2 gives an indication of time between observations and the associated states. At the start of the study, the majority of patients were in State 3 and 4, with the least amount of patients observed in State 1. Subsequently, at each successive observation time the distribution of states stayed relatively consistent, with the majority of patients being in the higher states. Censoring occurs after 12 months into the study, which was the minimum length of time for follow-ups. Death can occur at any point in the trial. There did not seem to be a spike in death rates, with Figure 2 suggesting a roughly uniform spread of death rates.

3 Multi-State Models

This section provides the underlying theory and assumptions underpinning the multi-state model. The section is limited to methods that are relevant to the data used. The transitions between the states defined in Figure 1 are of interest, therefore the model is defined as a multi-state model. As the model includes an 'absorbing state', or a death state, it is referred to as a multi-state survival model.

3.1 Censoring

The data are treated as longitudinal data, as it is derived from up to 4 observations over a period of 18 months for each patient. The observations of CD4 count are also interval-censored as the transitions are known to occur in a specific time interval but the exact time of transitions are unknown. This is common in clinical trials and HIV studies as the onset of such diseases are usually based on blood work which is conducted periodically.

Let T be the time to the event of interest. Interval censoring can be defined as the interval I which contains the time T. Therefore it can be written as $I = (L, R]$ [7].

This notation also includes right censoring, in the case where $R = \infty$. This includes patients who drop out of the trial, which is common in the collection of longitudinal data. Death times are known exactly, including when death occurs after the length of the trial. Censoring is also assumed to be non-informative, where the censoring is not informed by the underling process.

3.2 Continous-Time Markov Process

Longitudinal data are often realisations of continous-time processes at arbitrary times, therefore a continuous-time model is assumed. This reflects the ability of the patient to change between states at any time. The modelling framework also assumes an underling Markov process. Transitions between states are assumed to follow a stochastic Markov process, where the transition to the next state is soley dependent on the current state of the patient. Therefore, the previously occupied states are not considered. More formally, a Markov process can be denoted as $\{Y(t), t \geqslant 0\}$ which takes the values in the finite state space, $\mathcal{S}$. The history of the process up till time t is defined as $\mathcal{H}_t = \{Y(v); 0 \leqslant v \leqslant s\}$, where $r, s \in \mathcal{S}$.

The *transition probability* in a Markov multi-state model therefore makes the assumption [8]

$$P(Y(t+u) = s | Y(t) = r, \mathcal{H}_{t-}) = P(Y(t+u) = s | Y(t) = r), \quad \text{where } u > 0. \tag{3.1}$$

Thus, given the state of the process at time t, the future occupied states at any time after t is independent of the entire history, $\mathcal{H}_t$, of the process before time t.

The Markov chain is time homogenous when

$$P(Y(t+u) = s | Y(t) = r) = P(Y(t) = s | Y(0) = r). \tag{3.2}$$

Therefore, when the process enters state r, the way the process develops probabilistically from that point is the same as if the process started in state r at the start, $Y(0)$. Alternatively, the probability of being in state s at time $t+u$ given the current state r at time t depends only on the time interval u. Considering this, the publication by Kalbfleisch and Lawless (1985) is specifically of interest. It discusses the method for obtaining maximum likelihood estimates under a continuous time Markov assumption for longitudinal data [9].

The specific model used in relation to the data is shown in Figure 1. The transitions between states, indicated by the arrows, is dictated by *transition intensity,* $q_{rs}(t)$ for a pair of successive states. The transition intensity represents the conditional probability the event will occur in the time interval t, $t + dt$ as dt tends to 0, given the event has not occurred before. If T is a random variable denoting the time to event, the transition intensity is given as $q_{rs}(t) = \lim_{dt \downarrow 0} {}^{P(t \leqslant T \leqslant t + dt | T > t)}\!/_{dt}$. It can be interpreted as the 'instantaneous risk' of moving from one state to the next and it is always defined as a positive number [10]. Intensities can also be called hazards, both names are used throughout this thesis. Note that this is different from the *transition probability*, which is a probability and therefore is a positive value in $[0, 1]$.

Transition specific intensities can be denoted in a matrix. The transition intensity matrix defined for the five state model specified in Figure 1 is

$$Q(t) = \begin{pmatrix} q_{11t} & q_{12t} & 0 & 0 & q_{15t} \\ q_{21t} & q_{22t} & q_{23t} & 0 & q_{25t} \\ 0 & q_{32t} & q_{33t} & q_{34t} & q_{35t} \\ 0 & 0 & q_{43t} & q_{44t} & q_{45t} \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{3.3}$$

The rows of Q sum to zero. The diagonal entries are defined by $q_{rr} = -\sum_{s \neq r} q_{rs}$. Modelling transition intensities is the crux of multi-state survival modelling. The intensities that are modelled are the off diagonals, which represent the transitions into different states. The negative of the diagonal entries of the matrix represent the rate at which the patient spends in state r and it not modelled here. Alongside modelling the intensities into different states it is also of interest to derive and conduct inference via the transition probabilities.

The transition probabilities are the probabilities of transitioning between two states within a specified time interval. These probabilities are often presented in a matrix, where the rows represent the current state and the columns represent the next state. The *transition probability matrix,* $P(t)$ can be expressed in terms of the transition intensities using the Kolmogorov differential equations [8]. The solution is given as $P(t) = \exp(tQ)$. The matrix $P(t)$ contains all the probabilities where the rows sum to 1 and is given by

$$\hat{\mathbf{P}}(t) = \begin{pmatrix} \hat{p}_{11t} & \hat{p}_{12t} & \hat{p}_{13t} & \hat{p}_{14t} & \hat{p}_{15t} \\ \hat{p}_{21t} & \hat{p}_{22t} & \hat{p}_{23t} & \hat{p}_{24t} & \hat{p}_{25t} \\ \hat{p}_{31t} & \hat{p}_{32t} & \hat{p}_{33t} & \hat{p}_{34t} & \hat{p}_{35t} \\ \hat{p}_{41t} & \hat{p}_{42t} & \hat{p}_{43t} & \hat{p}_{44t} & \hat{p}_{45t} \\ \hat{p}_{51t} & \hat{p}_{52t} & \hat{p}_{53t} & \hat{p}_{54t} & \hat{p}_{55t} \end{pmatrix}. \tag{3.4}$$

The matrix exponential is often not straightforward to calculate and has been discussed in the literature by Moler and Van Loan [11]. The paper discusses various ways to calculate this exponential of a matrix, however for the purpose of this thesis, the *expm* package in *R* is used to compute this [12].

However, the intensities are restricted to being constant over time and thus, the transition intensity q_{rs} from state r → s is equal for all t. This assumption of transition intensities being constant may be unrealistic and restrictive. For example, some older patients in State 4 may be more likely to enter the death state, reflected by a higher value of q_{45}. However some patients may stay in State 4 for a longer time. In this case the transition intensity drops as t gets larger. There are many other scenarios that would result in a poor fit to the data, especially as diseases tend to evolve over time. Therefore, without the ability to model varying transition rates, models may fit poorly to data of this type. It is possible to use a piecewise constant approximation to handle time dependent intensities, which is the method used in this thesis. Flexibility can be introduced by dividing the time period into a number of intervals and estimating transition intensities for each interval, allowing them to differ across intervals [10]. This is discussed in more detail in Section 3.3 and Section 4.

3.3 Intensity Models

Covariates can also be included into the model. This is done by multiplying the log linear regression model with the baseline hazard [13]. For a transition $r \to s$, the model is given by

$$q_{rs}(t|\boldsymbol{x}) = q_{rs}(t) \exp(\boldsymbol{\beta}_{rs}^{\top}\boldsymbol{x}(t)). \tag{3.5}$$

$\boldsymbol{\beta}$ denotes the parameter vector and $\boldsymbol{x}(t)$ denotes the covariate vector. The baseline hazard is given by $q_{rs}(t)$ and can be specified by various parametric distributions. Examples include the exponential or Gompertz distributions, where transition-specific time dependencies can be introduced. The baseline hazards used in this thesis are

$$\begin{aligned}
\text{Exponential}: &\quad q_{rs}(t) = \lambda_{rs} && \lambda_{rs} > 0 && (3.6)\\
\text{Gompertz}: &\quad q_{rs}(t) = \lambda_{rs}\exp(\zeta_{rs}t) && \lambda_{rs} > 0 && (3.7)\\
\text{Weibull}: &\quad q_{rs}(t) = \lambda_{rs}\tau_{rs}t^{\tau_{rs}-1} && \lambda_{rs}, \tau_{rs} > 0 && (3.8)\\
\text{Log-logistic}: &\quad q_{rs}(t) = \frac{\lambda_{rs}\rho_{rs}(\lambda_{rs}t)^{\rho_{rs}-1}}{1+(\lambda_{rs}t)^{\rho_{rs}}} && \lambda_{rs}, \rho_{rs} > 0. && (3.9)
\end{aligned}$$

The exponential distribution leads to a constant hazard function which, as discussed previously is not always suitable. The Gompertz distribution is characterised by two parameters: λ and ζ. The parameter ζ determines the shape of the intensity function with positive values leading to an intensity function which increases with time. The hazard increases from λ at t_0 to ∞ at t_∞ as shown in Figure 3. Subsequently, negative values lead of ζ lead to intensity functions which decreases with time. The Gompertz model is more flexible compared to the exponential model as it allows for intensity rates that are non-constant but monotonic. It can also be seen that the Gompertz distribution is analagous to the exponential distribution when $\zeta = 0$ and therefore has a constant value λ. The Weibull distribution is also displayed in Figure 3. It also treats the hazard as monotonic but not necessarily constant. Similar to the Gompertz model, it is also a two-parameter model, where λ is the location parameter and τ is the shape parameter. The Weibull model also nests the exponential model when τ is equal to 1. An alternative to the Weibull model is the log-logistic model. The log-logistic model is another example of a two-parameter model. It is fairly flexible and allows for non-monotonic hazards which are unimodal.

As mentioned previously, the piecewise constant approximation is used to handle time dependent intensities. It is defined as follows. Assume there are K check ups, which can be

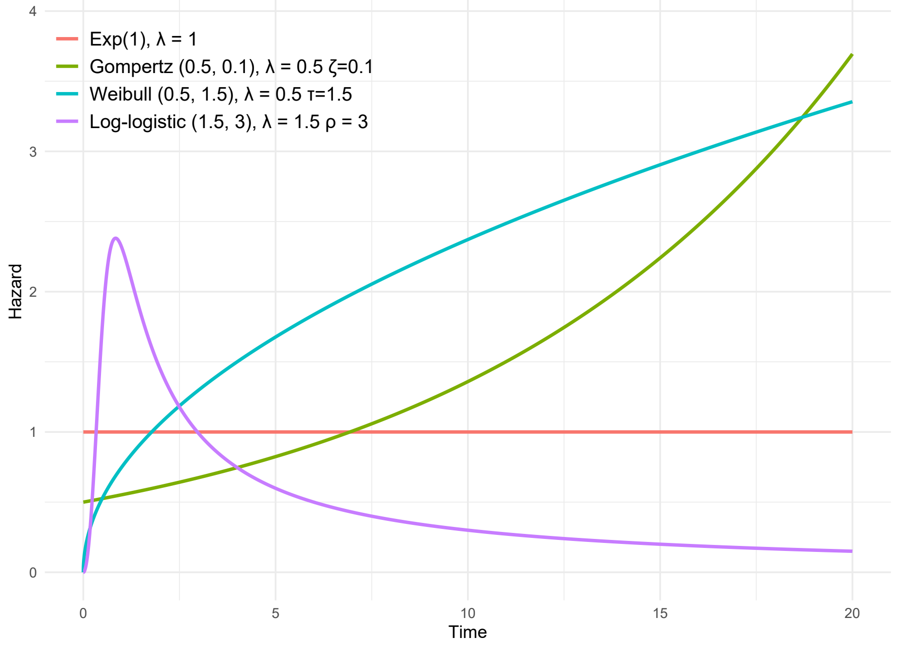

Figure 3 *Graph showing examples of hazards. Red line showing exponential, green line showing Gompertz, blue line showing the Weibull distribution and purple line showing the log-logistic distribution.*

divided into u equidistant time intervals. For example if patient i has 4 check ups, $K = 4$. For each interval let the intensity change in accordance with a parametric shape:

$$q_{rs}(t) = \begin{cases} \lambda_1 & u_1 < t \leqslant u_2 \\ \lambda_2 & u_2 < t \leqslant u_3 \\ \lambda_3 & u_3 < t \leqslant u_4. \end{cases} \tag{3.10}$$

Data must be available for all intervals to be able to estimate all the parameters. The split points can be defined by the data to ensure this. For the purpose of the analysis presented in this thesis, the split points coincide with the follow-up times for each patient. For example, consider the patient i with observed states at 0, 6, 12 and 16.97 months after the trial. The split points are defined as such and therefore the intensities are assumed constant within the intervals (0,6], (6,12], (12,16.97]. As the measurements are recorded in this manner, with progressively longer intervals between measurements, it can be assumed that the levels of CD4 follow a similar trajectory. There may be a noticeable change in levels of CD4 within the first 2 months which decreases over time. Therefore it is implied that the intensity changes in a similar time frame. This is potentially due to an immediate difference in CD4 counts after the first dose of treatment. This piece-wise constant approach is an approximation of the continuous-time intensities and maximum likelihood is used to estimate the parameters. The likelihood function is described in detail in Section 4, where the intensity model is estimated using the follow-up times as split points and approximating the parametric shape for the intensity piecewise constantly. It should be noted that this method is not strictly necessary, piecewise models can be estimated when the split points and follow-up times differ [10][14].

4 Maximum Likelihood Estimation

Let $Y(t)$ be the random variable denoting the states person i can take at time t. There will be several observations of $Y = y$ as person i will have k recorded states denoted as $y_1, y_2, ..., y_k$. Let $\boldsymbol{\theta}$ be a vector containing all the model parameters and $\boldsymbol{x}$ be a vector of covariates. The covariates used in the analysis of the data are time, drug type and previous opportunistic infection. The corresponding parameters are denoted as α, γ and ξ. The intercept parameters are denoted as β.

For person i the likelihood (conditional on the first state) will be defined as follows [10]:

$$L_i(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{x}) = P(Y_K = y_K, Y_{K-1} = y_{K-1}, ..., Y_2 = y_2 | Y_1 = y_1, \boldsymbol{\theta}, \boldsymbol{x}) \tag{4.1}$$

$$= \prod_{k=2}^{K-1} P(Y_k = y_k | Y_{k-1} = y_{k-1}, \boldsymbol{\theta}, \boldsymbol{x}) \times P_c(Y_K | Y_{K-1}, \boldsymbol{\theta}, \boldsymbol{x}). \tag{4.2}$$

(4.1) states the likelihood of patient i as the joint density of all the observations. The last observation time can be known, for example if the patient ended the trial in State 1, 2, 3 or 4. The last observation time can also be as a result of death, State 5. Alternatively, the last state can also be right censored. These different scenarios are accounted for by (4.2), under the Markov assumption discussed in Section 3.2. The assumption of time homogeneity holds over the time interval t_K, t_{K-1}. The first term of (4.2) takes the product of the states at each observation time $y_2, ..., y_{k-1}$ and the *likelihood contribution* P_c accounts for the state at the last observation time y_K.

If the patient is living at the last observation time t_k and the state is known, the *likelihood contribution*, P_c becomes

$$P_c(Y_K = y_K | Y_{K-1} = y_{K-1}, \boldsymbol{x}) = P(Y_K = y_K | Y_{K-1} = y_{K-1}, \boldsymbol{x}). \tag{4.3}$$

If the patient exits the trial, the end state is unknown. Therefore, the patient is known to be alive, but the state upon exit is unknown. Thus, the state is right censored. If right censoring occurs as t_K, all the states except death are summed over. This is because the patient could be in any of these states and thus all living states must be accounted for. For the purpose of the aids dataset, State 1, 2, 3 or 4 are summed over:

$$P_c(Y_K | Y_{K-1}, \boldsymbol{x}) = \sum_{s=1}^{4} P(Y_K = s | Y_{K-1} = y_{K-1}, \boldsymbol{x}). \tag{4.4}$$

If a death is observed at t_K, the summation of all the alive states are multiplied by the hazard intensity from each state s to death D, which indicates an ‘instant death’ from state s:

$$P_c(Y_K = y_K | Y_{K-1} = y_{K-1}, \boldsymbol{x}) = \sum_{s=1}^{4} P(Y_K = s | Y_{K-1} = y_{K-1}, \boldsymbol{\theta}, \boldsymbol{x}) \times q_{sD}(t_{K-1} | \boldsymbol{\theta}, \boldsymbol{x}). \tag{4.5}$$

The likelihood function is formed by combining the contributions for all 467 patients and maximising over $\boldsymbol{\theta}$. Thus, the likelihood function is given by $L = \prod_{i=1}^{437} L_i(\boldsymbol{\theta}|y, x)$. For the models used in this thesis, the likelihood function was coded by first looping over each patient in the trial and extracting the necessary data such as all the observed states, all the observation times, the administered drug and the previous opportunistic infection. Then there was a loop over each individual patient's follow-up and further information was extracted. This included the length of the time interval between follow-ups and the observed states at the start and end of each interval. This information was used in the construction of the likelihood function. For each time interval, the transition intensity matrix, $Q(t)$, was defined using the extracted covariate information. The transition probability matrix, $P(t)$, was then computed for each time interval using the `expm` package in `R`. The likelihood contribution was calculated based on the recorded state at the end of the time interval as discussed above. The log-likelihood was updated at every iteration by adding each contribution.

As mentioned previously, a piecewise constant approximation was used when fitting time dependent models. When iterating over each observation, the start of the time interval was used as the covariate for the effect of time. Therefore, $Q(t)$ is constant from the start of each interval to the end of the interval. Consider patient i with observed states recorded at the start of the trial, 6 and 12 months later. The transition intensity was assumed constant for the time interval from $(0, 6]$ and $(6, 12]$. In the case of the Gompertz model, these constant intensities took the shape of the Gompertz model. That is, the intensities approximated the Gompertz distribution in a stepwise manner. The same is true for the Weibull and log-logistic distributions. Although the transition intensity, $Q(t)$, was assumed constant from the start of the time interval, this is not strictly necessary. $Q(t)$ can be assumed constant from any point, for example the midpoint of the interval.

The approximation is therefore data dependent, which makes the follow-up times an important aspect in fitting these models. This is because the data itself is defining the piecewise constant approximation when carrying out maximum likelihood estimation. Each patient has their own follow-up times which give rise to it's own piecewise constant approximation as used in the likelihood function. If many individuals had long observation times this may cause issues when fitting the piecewise constant approximation. Longer time intervals between follow-ups result in a cruder approximation of the baseline parametric shape. In the case of the aids dataset, the majority of patients were followed up for a minumum of 12 months after the start of the trial and therefore there was sufficient amounts of data collected at follow-up times. Therefore the piecewise constant approximation is appropriate in the construction and maximisation of the likelihood function in this case. Note that the implication of using the piecewise constant approximation with the split points defined by the data is that the model parameters are also data dependent. However the model itself is not piecewise constant and assumes a continuos time dependency.

For the purpose of this thesis, the general optimiser package `ucminf` in `R` is used to maximise over $\boldsymbol{\theta}$ [15]. `ucminf` is an general-purpose unconstrained linear optimisation package. It contains an algorithm which utilises the quasi-Newton approach to optimisation. A line search is used to determine the step size. Convergence is reached if the relative convergence criteria are satisfied. Starting values of -3 were given for each transition intensity that was modelled. Depending on the number of covariates used, the elapsed time for the `ucminf` function to maximise the likelihood function ranged from 20 to 30 minutes. Convergence was reached and stopped by a zero step from the line search. The `R` code used to implement the likelihood function described here is shown in the Appendix.

5 Data Analysis

5.1 Model Comparison

This section considers the assessment of models which is used in following sections. Specifically the Akaike information criteria (AIC) as a comparison metric is briefly discussed. The AIC is given by

$$AIC = -2\log(L_{max}) + 2k. \tag{5.1}$$

L_{max} is the maximum value of the likelihood estimate and k is the number of parameters in the model. AIC estimates the information loss for a particular model. The model with the least information loss is deemed the higher quality model, and therefore the preferred model is the one with the minimum AIC value. The AIC rewards goodness of fit via the likelihood function but also penalises based on the number of independent parameters. It is therefore used to pick the model with the optimum trade off between model fit and complexity. Models that have more parameters are considered more complex and less parsimonious. However, the AIC is a relative metric and only determines the quality of a model relative to others but not the absolute quality of the model. Therefore it will not determine if the model fits the data well in absolute terms. For this, a goodness of fit metric is required, which is discussed in more detail in Section 5.3.

5.2 Model Specification

5.2.1 Baseline Intensity Model

Models were built from the bottom up beginning with an intercept only exponential intensity model, which was time homogenous given previously in (3.5) and (3.7). The model is given by

$$q_{rs}(t) = \lambda_{rs} = \exp(\beta_{rs}) \tag{5.2}$$

for $(r, s) \in \{(1, 2), (1, 5), (2, 1), (2, 3), (2, 5), (3, 2), (3, 4), (3, 5), (4, 3), (4, 5)\}$. This model was kept parsimonious by fixing the parameters $\beta_{12} = \beta_{23} = \beta_{34} = \beta_{forward}$, $\beta_{21} = \beta_{32} = \beta_{43} = \beta_{backward}$ and $\beta_{15} = \beta_{25} = \beta_{35} = \beta_{45} = \beta_{death}$. That is, the parameters are fixed for all forward, backward and death transitions, giving a total of 3 parameters. The maximum likelihood approach was used to estimate the model parameters, as defined in Section 4. This produced the estimations $\hat{\beta}_{forward} = -2.030$, $\hat{\beta}_{backward} = -3.041$ and $\hat{\beta}_{death} = -3.446$. The corresponding standard errors were $0.071, 0.073$ and 0.108 respectively. The AIC was 3116.84. These estimated $\hat{\beta}$ parameters illustrate the risk of a declining CD4 count. The forward transition intensity is estimated at $\hat{\lambda}_{forward} = \exp(-2.030) = 0.131$. Comparing this to $\hat{\lambda}_{backward} = \exp(-3.041) = 0.048$. illustrates the higher baseline risk associated with moving forward through the states.

This model is not necessarily realistic as it assumed that the intensities are constant and do not change over time. However, the exponential model can be seen as the null model, upon which extended models will build upon.

More flexibility is introduced in the model by releasing constraints on the baseline intensities. In addition, the time since the start of the trial is added as a covariate. Given the number of months since the start of the trial, the time t is transformed by t = number of months since the start of the trial +1. This prevents numerical problems when fitting the Weibull model. Using the piecewise-constant approximation as discussed in Section 4, the Gompertz model is fitted with restrictions on parameters for the effect of time t. Parameters are estimated for the forward transitions, however parameter equality constraints are placed on backward transitions and transitions into the death state. This is to again ensure parsimony, but also as these transitions have less associated data. For example, Table 1 shows only 3 observations of patients entering the death state after being observed in State 1. Therefore, imposing these parameter constraints will help account for the sparse data. The Gompertz model is given by

$$q_{rs}(t) = \exp(\beta_{rs} + \alpha_{rs} t) \tag{5.3}$$

where $\alpha_{21} = \alpha_{32} = \alpha_{43} = \alpha_B$ and $\alpha_{15} = \alpha_{25} = \alpha_{35} = \alpha_{45} = \alpha_D$. The model has 15 parameters and an AIC value of 3038.11. This is noticeably better than model (5.2) which had an AIC of 3116.84.

5.2.2 Regression Models

This model was extended to include the effect of the drug where 0/1 represent zalcitabine/didanosine respectively. It is given by

$$q_{rs}(t) = \exp(\beta_{rs} + \alpha_{rs} t + \gamma_{rs} drug). \tag{5.4}$$

Parameter restrictions were also placed on the effect of the drug. Initially, the same restrictions placed on the time covariate, t, were also placed on the effect of the drug, γ. This resulted in the restrictions $\gamma_{21} = \gamma_{32} = \gamma_{43} = \gamma_B$ and $\gamma_{15} = \gamma_{25} = \gamma_{35} = \gamma_{45} = \gamma_D$. This model had 20 parameters and gave an AIC value of 3042.50. However the standard errors of the parameters for the transitions for the effect of the drug were relatively large. Therefore, further parameter restrictions were added. Firstly, the backward transitions were set to zero, $\gamma_D = 0$ and equality restrictions were placed on the parameter $\gamma_{23} = \gamma_{34}$. This was because the immediate effect of the drugs is of primary interest, and so the transition between State 1 and State 2 was considered. The resulting model had 18 parameters and an AIC value of 3038.97. Note that by adding parameter restrictions a better value of AIC was obtained.

Next, the effect of having a previous opportunistic infection such as an AIDS diagnosis was considered. This covariate was added as *prevOI* where 0/1 denotes a patient having received a prior no AIDS/AIDS diagnosis. The parameter restrictions were the same as the parameter restrictions on the drug, where $\xi_{23} = \xi_{34}$, $\xi_B = \xi_{21} = \xi_{32} = \xi_{43} = 0$ and $\xi_D = \xi_{15} = \xi_{25} = \xi_{35} = \xi_{45}$.

To recap, model is given by

$$\begin{aligned}
q_{rs}(t) &= \exp(\beta_{rs} + \alpha_{rs} t + \gamma_{rs} drug + \xi_{rs} prevOI) \\
&\qquad\qquad \text{for } (r,s) \in \{(1,2),(2,3),(3,4)\} \\
&\qquad\qquad \gamma_{23} = \gamma_{34}, \;\; \xi_{23} = \xi_{34} \\
q_{rs}(t) &= \exp(\beta_{rs} + \alpha_D t + \gamma_D drug + \xi_D prevOI) \\
&\qquad\qquad \text{for } (r,s) \in \{(1,5),(2,5),(3,5),(4,5)\} \\
q_{rs}(t) &= \exp(\beta_{rs} + \alpha_B t) \\
&\qquad\qquad \text{for } (r,s) \in \{(2,1),(3,2),(4,3)\}.
\end{aligned} \tag{5.5}$$

Table 2 *Parameter estimates for five-state model, from the Gompertz model specified in (5.5). Weibull and log logistic parameter estimates also shown as given by (5.6) and (5.7) . Estimated standard errors in brackets. Time scale t is time since start of trial in months + 1. β are the intecept parameters, α is the time parameters, γ is the drug parameters and ξ is the prevOI parameters.*

	Gompertz		*Weibull for* $1 \to 2$		*log-logistic for* $1 \to 2$
β_{12}	-1.946 (0.296)	β_{12}	-1.519 (0.513)	β_{12}	-1.527 (0.345)
β_{15}	-5.803 (0.870)	β_{15}	-5.793 (0.877)	β_{15}	-5.728 (0.836)
β_{21}	-2.090 (0.225)	β_{21}	-1.903 (0.246)	β_{21}	-1.875 (0.247)
β_{23}	-1.799 (0.198)	β_{23}	-1.685 (0.229)	β_{23}	-1.689 (0.229)
β_{25}	-6.051 (1.023)	β_{25}	-6.018 (1.022)	β_{25}	-6.062 (1.041)
β_{32}	-3.065 (0.230)	β_{32}	-2.875 (0.249)	β_{32}	-2.869 (0.248)
β_{34}	-2.057 (0.181)	β_{34}	-2.012 (0.199)	β_{34}	-2.012 (0.199)
β_{35}	-4.835 (0.512)	β_{35}	-4.804 (0.512)	β_{35}	-4.807 (0.513)
β_{43}	-2.744 (0.186)	β_{43}	-2.554 (0.210)	β_{43}	-2.547 (0.209)
β_{45}	-3.441 (0.258)	β_{45}	-3.412 (0.268)	β_{45}	-3.413 (0.268)
α_{12}	-0.050 (0.056)	τ_{12}	-0.302 (0.257)	ρ_{12}	0.137 (0.196)
α_{23}	-0.114 (0.047)	α_{23}	-0.114 (0.047)	α_{23}	-0.113 (0.047)
α_{34}	-0.043 (0.033)	α_{34}	-0.044 (0.033)	α_{34}	-0.044 (0.033)
α_B	-0.177 (0.044)	α_B	-0.181 (0.044)	α_B	-0.184 (0.044)
α_D	-0.028 (0.020)	α_D	-0.028 (0.020)	α_D	-0.028 (0.020)
γ_{12}	0.083 (0.306)	γ_{12}	0.093 (0.305)	γ_{12}	-0.132 (0.417)
$\gamma_{23} = \gamma_{34}$	-0.178 (0.143)	$\gamma_{23} = \gamma_{34}$	-0.177 (0.143)	$\gamma_{23} = \gamma_{34}$	-0.176 (0.143)
γ_D	0.272 (0.148)	γ_D	0.272 (0.148)	γ_D	0.271 (0.148)
ξ_{12}	0.337 (0.359)	ξ_{12}	0.335 (0.358)	ξ_{12}	0.608 (0.538)
$\xi_{23} = \xi_{34}$	0.499 (0.154)	$\xi_{23} = \xi_{34}$	0.499 (0.154)	$\xi_{23} = \xi_{34}$	0.499 (0.154)
ξ_D	0.681 (0.220)	ξ_D	0.681 (0.220)	ξ_D	0.682 (0.220)
AIC	3020.40		3019.37		3019.73

This model had 21 parameters and an AIC value of 3020.40. This model's estimated parameters are stated in Table 2 under the Gompertz column. The effect of time since the start of the trial is associated with an increase in CD4 count and moving backwards through the states as all the $\hat{\alpha}$ parameters are negative. This implies an increase in health. This is potentially due to the effect of the drug helping to produce more CD4 cells thereby increasing health. This makes sense considering the value $\hat{\alpha}_B$ is largest and the value of $\hat{\alpha}_D$ is the smallest.

Different drugs seem to have different effects according to different transitions. For example, $\hat{\gamma}_D$ is positive, indicating a higher risk of death with the didanosine drug. The middle transitions, $\hat{\gamma}_{23}$ and $\hat{\gamma}_{34}$ were negative, indicating a higher risk of decreasing CD4 count with zalcitabine. The standard error for the $\hat{\gamma}_{12}$ estimate is large and therefore inference cannot be made accurately for this transition. However the hazard ratio for the effect of the drug into the death state, $\hat{\gamma}_D$, was exp(0.272) = 1.31, keeping all other covariates constant. Therefore the expected hazard is 1.3 times higher for patients taking the didanosine drug. This could indicate that zalcitabine may help in preventing death, but will not protect aganist the progression of the disease. If zalcitabine is effective in preventing death, patients will live longer and be alive to experience the disease evolve. All the values of the $\hat{\xi}$ parameters were positive, therefore patients with a previous AIDS diagnosis have a higher risk of a decline in health and moving into higher states. The value of $\hat{\xi}_D$ was 0.681, giving a hazard ratio of exp(0.681) = 1.98. Therefore patients who had a previous AIDS diagnosis were nearly twice as likely to experience death.

Using the Weibull and log-logistic model as a baseline hazard was also investigated for the model specified in (5.5) for transitions from State 1 → 2. Therefore the first equation in (5.5) for this specific transition is replaced by

$$q_{12}(t) = \lambda_{rs}\tau_{rs}t^{\tau_{rs}-1} \qquad \text{for } (r,s) \in \{(1,2)\} \tag{5.6}$$

for the Weibull model and

$$q_{12}(t) = \frac{\lambda_{rs}\rho_{rs}(\lambda_{rs}t)^{\rho_{rs}-1}}{1+(\lambda_{rs}t)^{\rho_{rs}}} \qquad \text{for } (r,s) \in \{(1,2)\} \tag{5.7}$$

for the log-logistic model, where $\lambda_{rs} = \exp(\beta_{12} + \gamma_{12}\mathit{drug} + \xi_{12}\mathit{prevOI})$.

The parameters are stated in Table 2. Consider first the Weibull distribution for the transition 1 → 2. The change in distribution reduced the AIC value of the model slightly giving a value of 3019.37. The estimated parameter values are very similar to the model with the Gompertz distribution. The $\hat{\tau}_{12}$ value was given as $\exp(-0.302) = 0.0488$. This value is less than 1, indicating time is associated with a lower risk of decline in health in the transition between State 1 and State 2. This corroborates the previous analysis under the Gompertz distribution. The use of the log linear distribution for transition 1 → 2 increased the AIC slightly to 3019.73, which was above the model containing the Weibull hazard but lower than the AIC of the Gompertz model. The value for $\hat{\rho}$ was 0.137 but had a large standard error of 0.196. However the remaining parameters were very similar to the Gompertz and Weibull model as seen in Table 2. Note however the estimates for the $\hat{\gamma}_{12}$ parameters had large standard error across all three models. This implies that there was not enough data for this transition to provide an estimate for the drug effect between states 1 → 2.

The effect of gender was also considered, with the same parameter restrictions applied for the effect of gender as the effect of *prevOI*. However, this model had 24 parameters and an AIC of 3023.155. Adding the effect of gender therefore increased the AIC and was therefore not chosen as the final model. A full list of the models and the associated AIC values can be seen in Table 3.

Although the parameter values provide some insight, it is often more straightforward to interpret the transition probabilities. Consider a time interval of 2 months since the start of the trial. As the interval is short, it can be assumed that the intensities are constant. For an average patient, given didanosine and with a previous AIDS diagnosis under the Gompertz model specified in (5.5), the transition probability matrix is

$$\hat{\mathbf{P}}\left(\begin{matrix} t_1 = 0 \\ t_2 = 2 \end{matrix}\,\middle|\, \begin{matrix} \mathit{drug} = 1 \\ \mathit{prevOI} = 1 \end{matrix}\right) = \begin{pmatrix} 0.667 & 0.251 & 0.057 & 0.007 & 0.016 \\ 0.142 & 0.528 & 0.259 & 0.050 & 0.020 \\ 0.007 & 0.053 & 0.642 & 0.242 & 0.056 \\ 0.000 & 0.004 & 0.089 & 0.760 & 0.147 \\ 0.000 & 0.000 & 0.000 & 0.000 & 1.000 \end{pmatrix}. \tag{5.8}$$

This matrix is calculated using the transition intensity matrix, where the probability of occupying state s at time $t+u$ conditional on occupying State r at time t is given by the (r,s) entry of the transition probability matrix $\hat{\mathbf{P}}$. This is obtained by using the equation $P(t) = \exp(tQ)$. The highest probabilities are on the diagonals, indicating that a patient with a previous AIDS diagnosis and given didanosine is more likely to stay in the same state. For example the probability of staying in State 1 is 66.7%. If a patient is in State 4, there is a 76.0% probability of saying in State 4, which is the highest probability in the matrix. However, forward transitions have a higher chance than backward transitions. For example, a patient in State 3 has a 24.2% chance of transitioning to State 4 in 2 months, but a 5.3% chance of transitioning to State 2.

For longer periods of time the intensities cannot be assumed constant. Therefore the transition probability matrix cannot be directly estimated from the transition intensity matrix by the equation $P(t) = \exp(tQ)$. This is because the intensity matrix Q varies within the time interval. Therefore the time interval is split up into smaller time intervals, within which the transition intensity is assumed constant. The transition probability matrix is obtained by multiplying these individual matrices in which Q remains constant. Thus the equation $P(t) = \exp(tQ)$ can be used within these intervals. Note that in the fitting of the model, the piecewise constant approximation was used. However once the model is fit, the model itself is not piecewise constant but parametric. Thus the grid for the piecewise constant approximation for prediction can be finer if necessary. For example consider a time interval of 15 years. This grid used for the approximation for prediction is $(0, 5], (5, 10]$ and $(10, 15]$. This could be made finer or expanded, but ultimately is up to choice. Regardless of the grid set-up, the transition intensities may differ across the intervals but remain constant within the interval. The transition probability matrix is computed in short hand notation by $P(15) = P(0, 5)P(5, 10)P(10, 15)$. For an average patient, given didanosine and with a previous AIDS diagnosis under the model specified in (5.5) the transition probability matrix is

$$\hat{\mathbf{P}}\left(\begin{matrix} t_1 = 0 \\ t_2 = 15 \end{matrix}\,\middle|\, \begin{matrix} drug = 1 \\ prevOI = 1 \end{matrix}\right) = \begin{pmatrix} 0.138 & 0.147 & 0.226 & 0.228 & 0.261 \\ 0.083 & 0.098 & 0.203 & 0.264 & 0.352 \\ 0.026 & 0.032 & 0.152 & 0.277 & 0.503 \\ 0.010 & 0.020 & 0.101 & 0.233 & 0.637 \\ 0.000 & 0.000 & 0.000 & 0.000 & 1.000 \end{pmatrix}. \tag{5.9}$$

The highest values are no longer on the diagonals but in higher states. For example, a patient currently in State 3 has a 27.7% chance of progressing to State 4 and a 50.3% chance of entering State 5, the dead state. Consider a patient in State 1 and being treated with didanosine. They have a 26.1% chance of entering the death state after 15 months, while a patient in State 4 has a 63.7% chance of entering the death state after 15 months. Therefore, early intervention seems critical in treatment of HIV patients.

In addition to the transition probability matrix it is also insightful to consider survival plots. This provides a graphical representation of predicting the probability of survival for some time t in the future. Survival is defined as not entering the death state, State 5. This can be obtained directly from the transition probability matrix, where the last column of the matrix indicates transition into the death state. For example consider the transition probability matrix presented in (5.9). A patient entering the trial in State 1 has a 26.1% chance of entering State 5, the death state, after 15 months. Survival is calculated by $1 - 0.261 = 0.739$. Therefore the patient has a 73.9% chance of survival. This value can be extracted and plotted for varying values of time intervals. Figure 4 shows the 15 month survival probability from each living transient state. The probability of survival for a patient in State 4 drops to 36.3% after 15 months, while with a patient in State 1 has 73.9% chance of survival after 15 months. This can be seen graphically in the survival plot.

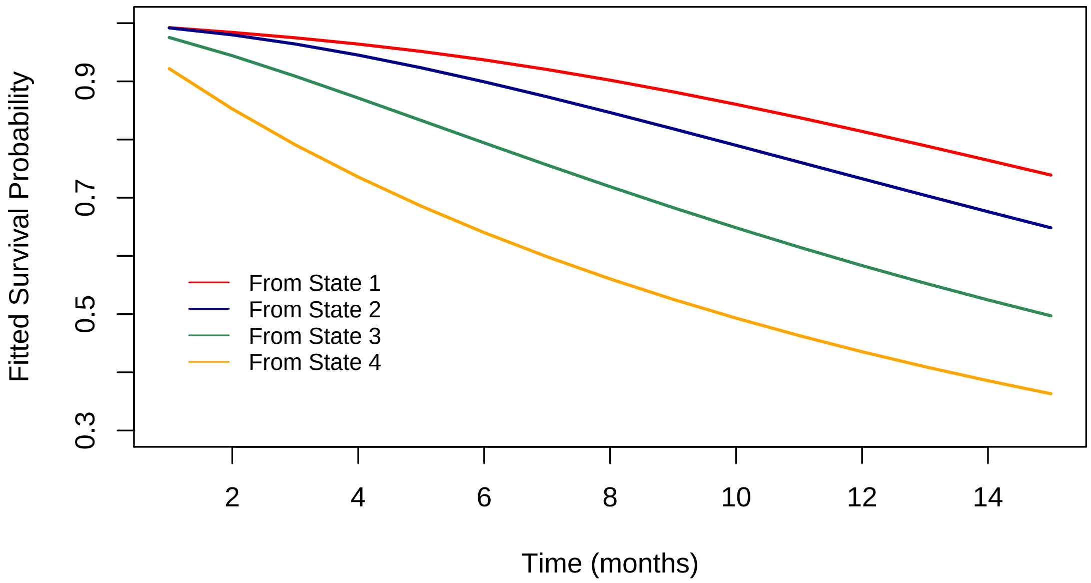

Figure 4 *Graph showing the fitted survival probability from each transient state for a patient being treated with didanosine and with a previous AIDS diagnosis. Survival is defined as not entering the death state.*

Table 3 *Comparison between the different models for the AIDS data, comparing the baseline hazards, parameters and the AIC value.*

Model	*Baseline Hazard*	*Parameters*	*AIC*
$Intercept - only$	exponential	3	3116.84
t	Gompertz	15	3038.11
$t, drug$	Gompertz	20	3042.50
$t, drug$	Gompertz	18	3038.97
$t, drug, prevOI$	Gompertz	21	3020.40
$t, drug, prevOI$	Gompertz and Weibull for $1 \to 2$ transition	21	3019.37
$t, drug, prevOI$	Gompertz and log-logistic for $1 \to 2$ transition	21	3019.73
$t, drug, prevOI, gender$	Gompertz	24	3023.16

5.3 Model Validation

It is also important to validate the models. The Kaplan-Meier estimate is a non parametric statistic which is used to estimate the survival function. Often presented visually, it shows the probability of an event such as survival at a certain time interval, $S(t) = P(T > t)$. With a large enough sample size, it should approach the true survival function for the population. In the absence of censoring the empirical survival function is

$$\widehat{S}(t) = \frac{\text{Number of sujects with T} > t}{\text{Total sample size}}, \tag{5.10}$$

where T denotes the survival time or time to event such as death. Therefore it gives the probability of surviving in a given length of time by considering shorter time intervals. For example if there were 5 deaths in the first month of a 15 month trial which started with 100 subjects. Then the probability of survival is 0.95 in the first month, $S(1) = 0.95$.

In the presence of censoring, the estimate becomes

$$\widehat{S}(t) = \prod_{i:\, t_i < t} \left(1 - \frac{d_i}{n_i}\right), \tag{5.11}$$

where t_i is a time when at least one event has happened and d_j is the number of events, such as the number of deaths that occurred at t_i. n_i is the number of patients known to have survived (still alive or right censored) up to time t_i. Although not a formal test, the Kaplan-Meier estimate can be used to flag any major issues with the generated models. The Kaplan Meier survival curve is assumed to be a good representation of patient survival in the data and therefore the multi-state model estimates should not be far off.

For the model presented in (5.5), Figure 5 provides a comparison between survivor functions based of the model and Kaplan–Meier estimates. The red line is the survival curve obtained from the multi-state model where the means of the covariates ($time, drug$ and $prevOI$) are used in the estimation. This comparison is baseline-state specific and considers survival from the state the patient entered the trial with. The sample size for each survival curve will be different as there is not an equal spread of states exhibited by the patients at the start of the trial. Figure 2 gives an indication of the spread. At the start of the trial, there seemed to be an equal distribution of patients in States 3 and 4, with the least number of patients entering the trial in State 1. The fit for survival from State 1 and State 3 are in line with the non parametric estimate. This is somewhat interesting considering there was less associated data for patients in State 1. Survival from State 2 falls near the lower bound of the confidence bands. There is some misfit for survival from State 4 with the red line falling above the 90% confidence band, potentially overestimating survival. This could be due to the fact that patients entering the trial in State 4 have different survival characteristics than those patients entering State 4 over the course of the trial.

5.4 Comparison with Joint Models

An alternative to using multi-state models is the joint model approach. There has been considerable interest in joint modelling in recent years, as longitudinal data and survival data arise frequently in practice. This section provides the basics to this approach and considers its use in relation to the AIDS dataset. This section also follows closely the theory presented in book [16] by Rizopoulos (2012), who is also the author of the `JM` package in `R` from which the data are used.

The joint model framework is also especially popular in clinical studies which more often that not deal with longitudinal data. Let Y_1 and Y_2 be the two outcomes of interest, such as the CD4 count and time to death. There are various approaches to obtain a joint density $p(y_1, y_2)$ of $\{Y_1, Y_2\}$. For the purpose of joint modelling, the random effect model framework is used. Joint models combines the probability distribution obtained from a linear-mixed effect model with

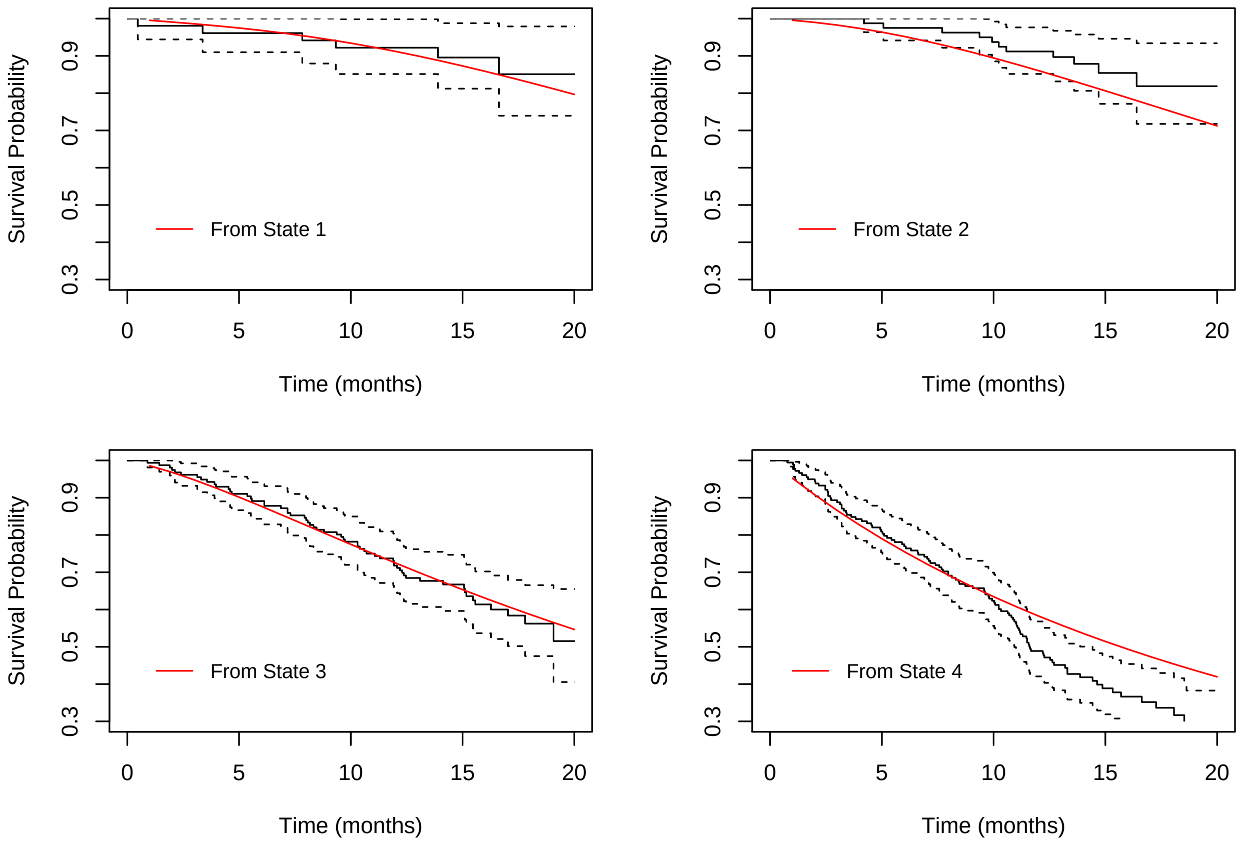

Figure 5 *Comparison of Kaplan–Meier curves (with 95% confidence bands) with estimated survival inferred from multi-state model. Red lines show mean survival from multi-state model and black lines shows the Kaplan–Meier estimate.*

random effects with a relative risk model. This section will first consider both the linear mixed effect model and the Cox proportional hazard model as building blocks to the joint model.

A typical feature of longitudinal data, is repeated measurements recorded per patient over a period of follow up time. Data of this type allows the consideration of a cross-sectional effect and the longitudinal effect. However, repeated measurements obtained on the same patient will often exhibit correlation. Standard statistical methods such as linear regression require the assumption of independent observations, which is typically not the case with data of this type and therefore not suitable.

The linear mixed effect model can be used to make use of all the data available and also account for positive correlations. The underlying premise of this model is that each individual in the population is assumed to have their own subject-specific mean response trajectories over time. Consider the response of a patient i at time t_{ij}. This notation denotes the $j'th$ response for the $i'th$ patient. The response trajectory can be written as

$$Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + \epsilon_{ij}, \qquad i = 1, ..., N, \quad j = 1, .., K \tag{5.12}$$

where β_{0i} is the patient or subject specific intercept and β_{1i} is the subject specific slope. Thefore (5.12) denotes the linear mean trend for the response of patient i. The line $\beta_{0i} + \beta_{1i} t_{ij}$ is described as the subject-specific mean trajectory. ϵ_{ij} is considered random and describes the within-subject variation. The overall mean slope and intercept can be denoted as β_0 and β_1, which account for all the subject specific β_{0i} and β_{1i} parameters. Each β_{0i} term and β_{1i} varies around the population mean intercept and overall mean slope respectively as shown by

$$\beta_{0i} = \beta_0 + b_{0i} \quad \text{and} \quad \beta_{1i} = \beta_1 + b_{1i}. \tag{5.13}$$

The terms b_{0i} and b_{1i} are the random deviations. Therefore β_0 and β_1 are the fixed effects and need to be estimated, b_{0i} and b_{1i} are the random effects. By combining the two models the following model is obtained:

$$Y_{ij} = \underbrace{(\beta_0 + b_{0i})}_{\beta_{0i}} + \underbrace{(\beta_1 + b_{1i})}_{\beta_{1i}} t_{ij} + \epsilon_{ij}. \tag{5.14}$$

The model can also be written in a matrix form:

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \qquad b_i \sim \mathcal{N}(0, D), \;\; \epsilon_i \sim \mathcal{N}(0, \sigma^2 I_{ni}). \tag{5.15}$$

where $\mathbf{b}_i$ is assumed to have a multivariate normal distribution.

This modelling approach is advantageous for a few main reasons. The first is that it is possible to predict the parameter vector $\boldsymbol{\beta}$, which describes how the mean response changes over time. It also allows for the prediction of an individuals trajectory over time via the parameter vectors $\boldsymbol{\beta} + \mathbf{b}_i$. Therefore, a linear mixed model consists of one intercept and slope to model the effect of the given variable on the outcome, for example the effect of CD4 count on the time to death. The inclusion of random effects allow for each patient to have their own individual intercept and slope parameter. It also offers flexibility because a balanced design is not required. Therefore each patient can have a different number of measurements each, not necessarily at the same time.

The survival component of the joint model framework typically consists of a relative risk model. The previous model (3.5) presented assumed the survival distribution took a parametric form, such as the Gompertz distribution. The Cox model has the form of

$$q_i(t|\boldsymbol{x}) = q_0(t)\exp(\boldsymbol{\beta}^\top \boldsymbol{x}_i). \tag{5.16}$$

The difference is that the baseline intensity it not directly estimated and not assumed parametric. When adding covariates, distinctions are made between the types of time-dependent covariates [13]. Exogenous, or external covariates, with respect to the outcome process are variables which are not affected by other variables and which the future value is not affected by the survival. Examples include patient age or environmental factors. This is in contrast to endogenous variables, which are internal variables generated by the patient. For the purpose of the dataset used in this thesis, where the CD4 count is treated as a biomarker, the covariate is endogenous. However, issues arise when using endogenous covariates in the Cox model. This is because the Cox model assumes no measurement error, which is often not the case with biomarkers. There is usually biological variation occuring within patients. For example, even if the CD4 count was to be measured twice on the same day it is unlikely to give the same value. The model also assumes the complete path of the CD4 count is fully specified. The assumption it makes is that the value of the biomarker only changes at observation times and is constant in the interval between check-ups. Therefore the intensity at any time point t is associated with the extrapolated value of the CD4 count at the same point. This is not a realistic assumption to make as the levels are likely to change in the time interval between check-ups. Therefore the Cox model cannot be used in isolation to model the disease pathway and risk of death.

Joint models arise from the limiting assumptions that biomarker values are measured without error and constant in time. They allow for both the longitudinal and survival process to be modelled in one model. It can be thought of as two component model where a linear mixed effect model is used to model the longitudinal biomarker and the proportional hazards model is used to model survival. First consider the longitudinal sub-model. Let $y_i(t)$ be the observed value of the biomarker at time t for patient i. It is noted that $y_i(t)$ is not observed for any time t, but at specific time points where the measurement is taken. Let $m_i(t)$ be the true and unobserved value of the biomarker at time t. A liner mixed effect model can be used to model $m_i(t)$ as expressed in equation (5.15). Thus $y_i(t)$ can be expressed as

$$y_i(t) = m_i(t) + \epsilon_i(t). \tag{5.17}$$

This model accounts for the measurement error by assuming that the observed level of the biomarker $y_i(t)$ is equal to the true level $m_i(t)$ plus a random error term. Assume now that $m_i(t)$ is known, and the subject specific values of the true unobserved values of the biomarker at time t are obtained. A standard relative risk model can be defined, where the longitudinal history of the biomarker till time t is considered:

$$q_i(t|\mathcal{M}_i(t)) = q_0(t)\exp(\boldsymbol{\beta}^\top \boldsymbol{x}_i + \alpha m_i(t)), \tag{5.18}$$

where $\mathcal{M}_i(t) = \{m_i(s), 0 \leqslant s < t\}$. The parameter α quantifies the association between the true CD4 biomarker count and the risk of death. It is often typical to leave the baseline hazard q_0 unspecified to avoid the consequences of misspecification. However, within the joint modelling framework q_0 is explicitly defined. A parametric distribution can be utilised, such as the ones discussed previously in (3.6).

Finally, a model is defined for their joint distribution. The function `jointModel` provided within the `JM` package, fits this joint distribution given that the linear mixed effect model and the Cox model is fitted separately first. This function utilises a combination of an EM algorithm and a quasi Newton algorithm if convergence is not achieved to obtain maximum likelihood estimates. More information about this can be found in the `JM` vignette [17].

The main feature of joint models is that they do not assume that the biomarker levels are constant between check-ups. However, it should be noted that full conditional independence is assumed, as the random effects explain all the interdependencies. Therefore the longitudinal outcome of the biomarker is considered independent form the survival outcome. Additionally, the repeated measurements of the biomarker are assumed to be independent of eachother. The censoring times are also assumed to be non informative.

A joint model is fitted using the AIDS data. The code used to fit the model is extracted from [16]. Firstly, the linear mixed model is fit:

$$\begin{aligned} y_i(t) &= m_i(t) + \epsilon_i(t) \\ &= \beta_0 + \beta_1 t + \beta_2(t \times \text{drug}_i) + b_{i0} + b_{i1}t + \epsilon_i(t). \end{aligned} \tag{5.19}$$

The effect of time and the interaction of time and drug (0/1 is zalcitabine and didanosine respectively) are included in the fixed effects part. In the random effects part, an intercept and a time term is included.

For the survival model, the drug effect is included as independent of time and the true and unobserved count of the CD4 are included as time dependent. This estimate is taken from the linear mixed effects model. Thus, the survival model becomes

$$q_i(t) = q_0(t)\exp(\gamma \text{drug}_i + \alpha m_i(t)). \tag{5.20}$$

$q_0(t)$ is assumed piecewise constant, with six knots placed at equidistant intervals of the observed event times. The `jointModel` function was used, which also allows for different basline risk functions such as the Cox proportional hazard model or a Weibull model.

The γ parameter was given as 0.335 (with a standard error of 0.157) and the α parameter was given as -0.288 (with a standard error of 0.036). The α parameter was negative, therefore the joint model finds an strong association between the CD4 cell count and the risk of death. A unit decrease in the CD4 count corresponds to a $\exp(-(-0.288)) = 1.3$ times increase in the risk of death. The γ parameter measures the association between the drug and risk of death. This is a positive number, indicating a higher risk with the didanosine drug. The relative increase in the risk of death from a unit increase of the didanosine drug is $\exp(0.335) = 1.4$. The paramater for the drug on the transition into death from the multi state model was given in Table 2. It was stated as $\gamma_D = 0.272$ where 0/1 represented zalcitabine/didanosine respectively. Therefore the didanosine drug was also associated with a higher risk of death by the multi-state model. The relative risk increase in the risk of death is given by $\exp(0.272) = 1.3$. This is similar to the relative risk computed via the joint model, but slightly lower. These results could indicate zalcitabine may provide more protective effects against the risk of death.

It is also worth considering the joint model from a higher level in order to compare it with the multi-state model. Although both models aim to answer the same questions, they both provide very different ways of modelling. The primary assumption of the multi-state model is the Markov assumption. This is the assumption that future is solely dependent on the current state and some covariates. This is a strong assumption to make and many processes are not strictly Markovian. For example if the transition into another state is dependent on the length of time spent in the current state, the Markov assumption doesn't hold. However it is difficult to lift this assumption due to the interval censored data. It is not possible to know when patients transition between the living states, only when they die. Thus it is difficult to model the length of time spent in a particular state because it is not known when the patient moves into that particular state.

Joint models are based on a different set of assumptions. Firstly, the longitudinal biomarker is modelled using a linear mixed effect model. The model assumes a functional form of the CD4 count, namely that it is a linear relationship. Any model fit with a linear regression is a simplification of the true relationship. The random effects allows for the modelling of individual trajectories via the intercept and slope parameter for each patient. These random effects are also assumed to follow a multivariate normal distribution. This of course is another assumption about things we do not observe. The random effects are assumed to explain all the interdendencies and therefore the assumption of full conditional independence is held. The time to death outcome and the longitudinal outcome is independent of each other. The repeated measurements of the biomarker are also assumed independent each other. Moreover, another assumption is made when extracting information from the linear model and inputting it into the survival model. It is difficult to know what information to take over from the linear model to see if the affects the time to death. This is in contrast with the multi-state model which is a fixed effects model and where there is only one model to model both processes. However along with the Markov assumption, assumptions are also made on how time affects the transition between states. Some of these have been discussed and checked in this thesis, such as the Gompertz and Weibull distributions to model the transition rates between states.

6 Conclusion

The statistical technique of survival analysis has major use in medical research. Many clinical trials consist of longitudinal studies, where patients are followed up over time with repeated monitoring of measurements such as biological markers of a disease. Multi-state models can be used for data of this type, which is interval censored in nature. This was shown in relation to the aids dataset, where the effect of time, drug and previous opportunistic infection were considered when modelling transition intensities. Inference was considered via the use of the likelihood function and the resulting transition probability matrices.

This flexible modelling framework of multi-state models can be utilised in many aspects of healthcare, and states can be discretised based on clinical symptoms or some other scale of the disease. In particular, these models provide interpretable quantities via the estimation of transition rates and the effect of covariates on these transitions. The wider scope of this type of quantitative modelling is the ability to translate research findings to patients' healthcare. Findings can help practitioners understand the disease trajectory to provide effective clinical measures. Moreover, multi-state models also have use outside the realm of healthcare and can extend to fields like engineering and economics. In fact the flexibility of these models can extend practically to any kind of longitudinal failure time data.

The alternative joint modelling approach was also briefly considered. It enables both the repeated biomarker measurements and the survival processes to be modelled simultaneously while accounting for the interrelation between the two. This type of framework takes a considerably different approach to modelling and lies on a different set of assumptions.

This thesis aimed to provide an introduction of the theory and illustration of both modelling frameworks. The presented multi-state models are all fixed effects models, however it is possible to explore frailty models which introduce mixed-effects models. This is useful to explore clusters of patients that share similar hazards and help account for heterogeneity. Further considerations can also include dealing with misclassification of states and missing data which were not discussed here.

References

[1] R. C. Gentleman, J. F. Lawless, J. C. Lindsey, and P. Yan. Multi-state Markov models for analysing incomplete disease history data with illustrations for HIV disease. *Statistics in Medicine*, 13(8):805–821, 1994.

[2] C. H. Jackson and L. D. Sharples. Hidden Markov models for the onset and progression of bronchiolitis obliterans syndrome in lung transplant recipients. *Statistics in Medicine*, 21(1):113–128, 2002.

[3] G. Marshall and R. H. Jones. Multi-state models and diabetic retinopathy. *Statistics in Medicine*, 14(18):1975–1983, 1995.

[4] S. G. Thompson S. W. Duffy C. H. Jackson, L. D. Sharples and E. Couto. Multistate Markov Models for Disease Progression with Classification Error. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 52(2):193–209, 2003.

[5] A. I. Goldman, B. P. Carlin, L. R. Crane, C. Launer, J. A. Korvick, L. Deyton, and D. I. Abrams. Response of CD4 lymphocytes and clinical consequences of treatment using ddI or ddC in patients with advanced HIV infection. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, 11(2):161–169, 1996.

[6] D. I. Abrams, A. I. Goldman, C. Launer, J. A. Korvick, J. D. Neaton, L. R. Crane, M. Grodesky, S. Wakefield, K. Muth, and S. Kornegay. A comparative trial of didanosine or zalcitabine after treatment with zidovudine in patients with human immunodeficiency virus infection. The Terry Beirn Community Programs for Clinical Research on AIDS. *The New England Journal of Medicine*, 330(10):657–662, 1994.

[7] Z. Zhang and J. Sun. Interval censoring. *Statistical Methods in Medical Research*, 19(1):53–70, 2010.

[8] D. R. Cox and H. D. Miller. *The Theory of Stochastic Processes*. Chapman and Hall, 1965.

[9] J. D. Kalbfleisch and J. F. Lawless. The Analysis of Panel Data Under a Markov Assumption. *Journal of the American Statistical Association*, 80(392):863–871, 1985.

[10] A. Van Den Hout. *Multi-State Survival Models for Interval-Censored Data*. Chapman and Hall/CRC, 2017.

[11] C. Moler and C. Van Loan. Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later. *SIAM Review*, 45(1):3–49, 2003.

[12] V. Goulet, C. Dutang, M. Maechler, D. Firth, M. Shapira, and M. Stadelmann. Matrix exponential, log, 'etc'. `https://cran.r-project.org/web/packages/expm/expm.pdf`.

[13] J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, 2nd edition, 2002.

[14] C. Jackson. Multi-State Models for Panel Data: The msm Package for R. *Journal of Statistical Software*, 38(8):1–28, 2011.

[15] H. B. Nielsen and S. B. Mortensen. General-purpose unconstrained non-linear optimization. `https://cran.r-project.org/web/packages/ucminf/ucminf.pdf`, 2016.

[16] D. Rizopoulos. *Joint models for longitudinal and time-to-event data: with applications in R*. Chapman & Hall/CRC, 2012.

[17] D. Rizopoulos. Joint modeling of longitudinal and survival data. https://cran.r-project.org/web/packages/JM/JM.pdf, 2018.

[18] C. Jackson. Multi-state modelling with r: the msm package. https://cran.r-project.org/web/packages/msm/vignettes/msm-manual.pdf, 2019.

7 Appendix

The programming language used to implement and analyse the presnted models is R. The displayed graphics were also all produced in R. R is a free software environment for statistical computing and graphics. More information and instructions for download can be found at https://www.r-project.org/. The appendix provides some samples of the code used in the thesis.

The data are loaded from the JM package [17] as follows:

```
# Load JM library from where data are located
library(JM)

# Data for multi-state model:
dtta <- aids

# Living states by CD4 groups:
CD4rnd <- round(dtta$CD4)
state  <- ifelse(CD4rnd < 5,1,NA)
state  <- ifelse(5 <= CD4rnd & CD4rnd < 10,2,state)
state  <- ifelse(10 <= CD4rnd & CD4rnd < 15,3,state)
state  <- ifelse(15 <= CD4rnd ,4,state)

# Orientation:
# Make higher states less healthy
state <- 5 - state
dtta$state <- state

# Other variables:
dtta$id     <- dtta$patient
dtta$drug  <- as.numeric(dtta$drug == "ddI")
dtta$prevOI <- as.numeric(dtta$prevOI == "AIDS")
dtta$gender <- as.numeric(dtta$gender == "male")

# Death state:
D <- 5
# Censored state:
censored <- -2
# Select variables:
dtta <- dtta[,c("id","state","obstime","CD4","drug",
               "death", "Time", "prevOI", "gender")]

# Add death or censoring:
```

```
subject <- unique(dtta$id)
N <- length(subject)
for(i in 1:N){
       dtta.i <- dtta[dtta$id==subject[i],]
       death <- dtta.i$death[1] == 1
       last.state  <- ifelse(death,D,censored)
       last.time   <- dtta.i$Time[1]
       last.record <- c(subject[i],last.state,last.time,NA,NA,
       as.numeric(death), last.time, NA, NA)
       dtta.i <- rbind(dtta.i,last.record)
       if(i == 1){ddtta <- dtta.i}else{ddtta <- rbind(ddtta,dtta.i)}
       }
dta_cov <- ddtta
```

In Section 4, the likelihood function was defined. The following R code illustrates the implementation of the log-likelihood function. The code shown is for the five state model with Gompertz baseline-intensity specified in (5.5). However, transitions can be easily changed into other baseline hazard distributions. The data are loaded under the data frame dta_cov as defined above:

```
# Load the relevant packages
library(expm)
library(ucminf)

# Prepare data for quick access:
dta_cov.split <- split(dta_cov, dta_cov$id)

# Prepare raw transition data:
o1 <- o2 <- dt <- dr <- ti <- pr <- rep(0, nrow(dta_cov))
n <- 0

for (i in 1:N) {
        # Extracting data for patient i:
        dta_cov.i <- dta_cov.split[[i]]
        O <- dta_cov.i$state
        t <- dta_cov.i$obstime
        drug <- dta_cov.i$drug[1]
        prev <- dta_cov.i$prevOI[1]
       # Loop over individual follow-up:
       for (j in 2:length(O)) {
               n <- n + 1
               dt[n] <- t[j] - t[j - 1] #Time between checkups
               o1[n] <- O[j - 1] #State at start of interval
               o2[n] <- O[j] #State at end of interval
               dr[n] <- drug
               ti[n] <- t[j-1] #Time at start of interval
               pr[n] <- prev
       }
}
```

```
 # Defining function for P matrix:
 Pmatrix <- function(Q, t) {
         expm(Q*t)
}

 loglikelihood <- function(p) {
  # Parameters:
  beta <- p
  if (mon) cat("beta = ", round(beta,digits),"\n")

  # Loop over intervals:
  loglik <- 0
  for (i in 1:n) {
   # Q matrix:
   Q <- matrix(0, D, D)
   # Off diagonal:
   Q[1,2] <- exp(beta[1] + beta[11]*ti[i] + beta[16]*dr[i] + beta[19]*pr[i])
   Q[1,D] <- exp(beta[2] + beta[12]*ti[i] + beta[17]*dr[i] + beta[20]*pr[i])
   Q[2,1] <- exp(beta[3] + beta[13]*ti[i])
   Q[2,3] <- exp(beta[4] + beta[14]*ti[i] + beta[18]*dr[i] + beta[21]*pr[i])
   Q[2,D] <- exp(beta[5] + beta[12]*ti[i] + beta[17]*dr[i] + beta[20]*pr[i])
   Q[3,2] <- exp(beta[6] + beta[13]*ti[i])
   Q[3,4] <- exp(beta[7] + beta[15]*ti[i] + beta[18]*dr[i] + beta[21]*pr[i])
   Q[3,D] <- exp(beta[8] + beta[12]*ti[i] + beta[17]*dr[i] + beta[20]*pr[i])
   Q[4,3] <- exp(beta[9] + beta[13]*ti[i])
   Q[4,D] <- exp(beta[10] + beta[12]*ti[i] + beta[17]*dr[i] + beta[20]*pr[i])

   # Diagonal:
   for (r in 1:(D - 1)){
   Q[r, r] <- -sum(Q[r, ]) #-Sum of the diagonals
   }
   # Pmatrix:
   P <- Pmatrix(Q, dt[i])

   # Likelihood contribution:
   # If patient died, multiply by Q
   if(o2[i] == D){
     contribution <- P[o1[i], 1:(D - 1)] %*% Q[1:(D - 1), D]
   }

   # If patient censored multiply by all living states
   if(o2[i] == censored){
     contribution <- P[o1[i], 1:(D - 1)]%*%rep(1,(D-1))
   }

   # If subject living at end of study and state is known
   if(o2[i] != censored & o2[i]!=D){
     contribution <- P[o1[i], o2[i]]
   }
   # Update likelihood:
   loglik <- loglik + log(contribution)
   }

        # Return:
        if (mon) cat("-2*Loglik = ", -2*loglik,"\n")
        -loglik
}
```

The maximum likelihood estimation was implemented using `ucminf` [15]. The following code produces estimates of the parameters and the associated standard errors:

```
mon <- FALSE #Monitoring
start_time <- Sys.time()
p.start <- rep(-3, 21)
control <- list(reltol = 1e-04, maxit = 500)
max_drug_time_prev4 <- ucminf(par = p.start, fn = loglikelihood,
                                hessian = 1)
end_time <- Sys.time()
time.run <- end_time - start_time
print(time.run)
cat("Starting values =", p.start, "\n")
cat("Convergence code =", max_drug_time_prev4$convergence, "\n")
cat("AIC =", 2 * max_drug_time_prev4$value +
        2 * length(max_drug_time_prev4$par), "\n")
p <- max_drug_time_prev4$par
fisher <- solve(max_drug_time_prev4$hessian)
p.se <- sqrt(diag(fisher))
print(cbind(p = round(p, digits), se = round(p.se, digits)), quote = FALSE)
```

The equivalent Gompertz model was also estimated using the msm package [18] by Jackson (2019). This package has a function called `pmatrix.piecewise.msm` which was used in calculating transition probabilities. For example Graph 4 showing fitted transition probability matrix for (5.5) was obtained using this function, as shown by following code:

```
covariates <- list(list(drug=1, prevOI=1),
                        list(drug=1, prevOI=1), list(drug=1, prevOI=1),
                        list(drug=1, prevOI=1), list(drug=1, prevOI=1))
times <- c(0 ,5, 10, 15)

# Initialising empty vectors for living states
state1 <- c()
state2 <- c()
state3 <- c()
state4 <- c()
```

```
 # Extracting survival probability
 for (i in 1:15) {
         m <- pmatrix.piecewise.msm(model_drug_time_prev4, 0, i,
                 times, covariates)
         state1[i] <- 1 - m[1, 5]
         state2[i] <- 1 - m[2, 5]
         state3[i] <- 1 - m[3, 5]
         state4[i] <- 1 - m[4, 5]
}

 # Plotting survival curves from each living state
 x <- c(1:15)
 plot(x, state1, type='l', xlim = c(1, 15), ylim = c(0.3,1),
         ylab="Fitted Survival Probability",
         xlab = "Time (months)", col="red", lwd=1.8)
 lines(state2, col="blue4", lwd=1.8)
 lines(state3, col="seagreen", lwd=1.8)
 lines(state4, col="orange1", lwd=1.8)
 legend(1, 0.6, legend=c("From State 1", "From State 2",
         "From State 3", "From State 4"),
         col=c("red", "blue4", "seagreen", "orange1"),
         cex=0.8, lty=1, bty = "n")
```