Quantifying patient benefit using semi-Markov multi-state models

A. Haris Jameel

A thesis submitted for the degree of

Doctor of Philosophy

School of Mathematical Sciences,

University of Nottingham

First submitted in March 2024

Resubmitted with corrections in September 2024

Supervised by

Dr. Chris Brignell	Dr. Blesson Chacko
Dr. Christopher Fallaize	Dr. Joachim Grevel
Prof. Gilles Stupfler	

In loving memory of

Abdoul Sukur Hamidou
(1943 – 2020)

Mohamed Farook Mohamed Hussain
(1941 – 2021)

J.M.A. Farida Sukur
(1944 – 2023)

You are all dearly missed.

Abstract

In the context of oncological drug trials, the semi-parametric Cox proportional hazards model is traditionally used to establish treatment efficacy based on patient response to treatment. However, the analysis is limited to answering questions about treatment efficacy only, since the focus is usually on a single event of interest (such as significant tumour shrinkage). It would instead be in the interest of patients to address whether a clinically effective drug is potentially beneficial, in terms of whether it can treat cancer while being relatively tolerable compared to alternative treatments. To address this, we propose modelling the entire patient history using a semi-Markov multi-state model so that we can simultaneously consider all possible events that can be experienced by patients. Furthermore, if one defines all possible events to be detrimental to the patient, we can quantify differences in patient benefit by considering each of the active and control treatment arms and the time patients spend in one or more states.

We propose two general statistical procedures to compare patients in each treatment arm. The first procedure is based on differences in expected sojourn time in subsets of states of interest, while the second procedure is based on differences in the survival function of the holding time in specific states. In each case, the test statistic is a function of the maximum likelihood estimates of model parameters. The delta method is used for statistical inference. Properties of the proposed statistical procedures are assessed by means of a simulation study, including analyses of power and effects of model misspecification. The main result is that each test is able to detect significant patient benefit relatively easily, with limiting factors being sample size and high rates of right-censoring.

Finally, a real dataset is analysed and our method is compared to each of the Cox proportional hazards model and the Fine-Gray proportional hazards model. The main conclusion is that our method is more flexible and insightful when considering patient benefit.

Acknowledgements

This research project is the result of a collaboration between University of Nottingham and BAST Inc. Limited. Their support, both financial and non-financial, has been instrumental to my success.

I am grateful that I have reached the end of what has been a long journey fraught with trials and tribulations. I could not have been made it without the help and support of many. First and foremost, I give thanks to my Lord and Master for supporting me despite my many failings. I also thank my loving family, who has been supportive of me moving nearly 7000 miles away to pursue my goals and ambitions. They have provided much comfort and emotional support over the years. I am also indebted to all of my supervisors who have constantly contributed valuable ideas to steer our research forward, and patiently helped me whenever I found myself stuck. In my eyes, they have been far better supervisors to me than I deserve.

My supervisors associated with BAST have provided endless amounts of moral support and guidance, for which I am grateful. Over the years, I have learned a great many things about the value and applications of our research. Dr. Joachim Grevel has been a patient and knowledgeable mentor who has, based on his vast experience in both academia and industry, shared much of his valuable wisdom with me. Dr. Blesson Chacko, despite his hectic work schedule, has always found time to offer both useful modelling insights and professional advice. I cannot thank them enough.

I also give my sincere thanks and gratitude to my academic supervisors. They have been a constant source of guidance, support, and professional advice. Each of them has supported me in their own unique way. Prof. Gilles Stupfler has constantly reminded me to observe rigour when writing mathematics, while Dr. Chris Fallaize has time and again helped me piece together details so that I can see the big (mathematical) picture. Dr. Chris Brignell has asked insightful questions about our research, which has strengthened the exposition of this thesis. Once again, I

cannot thank them enough for their support and dedication.

Finally, I thank my various friends and colleagues who have kept me sane and grounded through times both lonely and stressful. Among my friends and colleagues here in the UK, I thank: Isaac, Patrick, Haris, Elizabeth, Tahsin, Joe, Celene, Niren, Kat, and Jade. Among my friends back home, I thank: Hendrick, Hamzah, Nadirah, Aiman, Chentel, Ryzal, Samuel, and Traci. Among the friends I made online, I thank: HorologicalWoomy, Mushie64, Splatoonatic, tahm, and cinders. I also give special mention to everyone associated with Singapore Institute of Management who supported me in my endeavours to pursue graduate studies in the UK – especially Ms. Sylvia Yeo, Dr. Zhang Jianlin, and Dr. Phil Stephenson.

Last, but not least, I thank Kintaro the cat for being a good boy.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Time-to-event analysis is widely used in clinical oncology drug studies to ascertain the effectiveness of oncology drugs. Competing risks analysis methods are usually employed, namely fitting a cause-specific hazard function associated with an event of interest using the standard semi-parametric Cox proportional hazards model ([Cox, 1972]). The Cox proportional hazards model was first proposed by D.R. Cox, and it was R.L. Prentice *et al.* who built on Cox's work for competing risk modelling using cause-specific hazard functions ([Prentice et al., 1978]). The Cox model has since become one of "the standard bases for analysis, particularly of medical, epidemiological, and demographic data" ([Lawless, 2003, Chapter 7 Bibliographic Notes]). In trying to ascertain efficacy of a drug using the Cox model, the focus is on a particular event of interest (such as response to a drug which leads to a tumour shrinkage). All other events (such as premature discontinuation of drug due to adverse effects, loss to follow-up, death, *etc.*) are treated as right-censoring events [Wolbers et al., 2014, p. 2939], and without properly taking into account the propensity of these other events acting upon the patient during the course of treatment. This, by itself, may be acceptable if the main objective is merely to determine treatment efficacy, all other things being equal. However, such treatments can lead to other undesirable outcomes, such as adverse side effects which lead to premature discontinuation of treatment. For the purposes of this introduction, we may loosely define “benefit” as patients undergoing an effective treatment and being able to tolerate it enough to complete the treatment successfully. In this sense, the use of the Cox model to determine efficacy answers little about patient benefit since we are not attempting to quantify or incorporate information related to the other undesirable forces acting upon the patient (due to their disease and/or undergoing the treatment). Such information is certainly crucial to patients who are trying to decide how to best treat their illness, or if they even should.

The U.S. Department of Health and Human Services, Food and Drug Administration has recently issued a new draft guidance ([Food and Drug Administration, 2022]) to adjust the focus of drug development, with a view on emphasising patient benefit rather than easily quantifiable biological signals. Despite this, the statistical literature on patient well-being and benefit in the context of clinical drug development is scarce. One example that attempts to address patient benefit is [Oberoi et al., 2020], where the authors consider what factors lead to patients being lost to follow-up. However, loss to follow-up is but one undesirable outcome of a clinical trial, and there is a need to consider all of them in tandem to begin discussing potential benefit to patients. Besides being of interest to regulators, understanding patient benefit is also relevant to pharmaceutical companies formulating new treatments, as modelling results may convince payers to adopt a new “beneficial” treatment against already marketed competitors. In addition to the information provided by the standard Cox model, quantifying “patient benefit” as described above would give patients extra information about treatment options (in addition to information about drug efficacy) which they can weigh against their remaining life expectancy with and without treatment.

This thesis seeks to address the lack of statistical methodology to quantify potential benefit to patients so that they, their doctors, and their caretakers can make better-informed decisions about cancer treatments. Our contributions involve the use of semi-Markov multi-state survival models and are detailed in Chapters 3 through 6, with Chapter 2 describing the necessary background material and motivation. A more detailed description of the contents of this thesis can be found in Section 1.2.

Multi-state models account for the entire patient history from the start of a clinical trial up until either (i) they reach an absorbing state such as “death” or “lost to follow-up” or (ii) the end of the observation period (if they are right-censored because an absorbing state is never reached). This is as opposed to just considering one event of interest in a competing risks model, as is currently the case when analysing oncological clinical trial data. By incorporating information about the entire patient history more carefully, we are able to answer questions more general than just whether the treatment under consideration is effective given that the patient completes their treatment.

The multi-state models we consider use an underlying semi-Markov process to model the entire patient history. The main reason for choosing semi-Markov processes for our models is that they allow us to choose sensible distributions to model the biological processes that are common in clinical trials, as opposed to the possibly unrealistic assumption in Markov models that requires exponential-distributed holding times in a given state. Semi-Markov models are often modelled with *intensity transition functions* (ITFs), which generalise cause-specific hazard functions. The focus in this thesis is instead on the *mixture model* approach (which we refer to as the “mixture approach”), popularised by [Larson and Dinse, 1985]. This approach involves modelling semi-Markov models by estimating transition probabilities and parameters of proper probability distributions for a given possible transition. This choice is made due to higher overall interpretability and flexibility. While [Larson and Dinse, 1985] popularised the mixture approach, it has been known for significantly longer than that, as evident from the work by [Weiss and Zelen, 1965]. However, our methods are broadly applicable should the ITFs approach be preferred. Since we recognise that it is the ubiquitous and often preferred way of working with semi-Markov models, references to ITFs are made in the thesis where appropriate. See [Asanjarani et al., 2021] for a detailed comparison of both approaches, including advantages and disadvantages.

Traditionally, the EM algorithm is used to perform maximum likelihood estimation of the model parameters in a semi-Markov model, with the transition probabilities modelled with the logistic function. Standard errors are also estimated within the EM framework. See, for example, [Larson and Dinse, 1985], [Meng and Rubin, 1991] and many others. Bootstrap methods have also been used for constructing confidence intervals (see, for example, [Butler and Bronson, 2012] and [Castelli et al., 2007]). [Asanjarani et al., 2021] consider both the mixture and intensity transition functions approaches and explore their relationships in detail, while showing practical usage of the estimation methods on several datasets. In the context of the mixture method, they use the `SemiMarkov` package ([Król and Saint-Pierre, 2015]) in R to numerically estimate the standard errors in their models. However, we have found the method of estimating the parameters and standard errors using this package to be unclear and in disagreement with results that should be expected from the theory. We propose instead standard errors obtained by the use of the observed Fisher information to estimate the expected Fisher information. This is easily done, *e.g.* via numerical approximation by means of the Hessian obtained from an appropriate numerical optimisation procedure. This is discussed in Section 4.2. Preceding that discussion, we also present in Section 4.1 a new way to write the likelihood function, which allows for an alternative expression of entries of the observed Fisher information matrix. To the best of our knowledge, the likelihood has not been expressed in this way in the literature. In addition to allowing for an alternative closed-form expression for the observed Fisher information matrix, the new expression for the likelihood allows for the possibility to make theoretical calculations for entries of the expected Fisher information matrix.

The work by [Weiss and Zelen, 1965] argues the merits of semi-Markov multi-state models, and discusses various probabilities (such as state occupancy probabilities and first passage time probabilities), as well as the distributions of total sojourn times (which can involve passage through certain states of interest). Since the computation of such quantities can involve convolutions, they propose using Laplace transforms to linearise the integrals and make the computation more convenient. They also fit a semi-Markov multi-state model to acute leukaemia clinical trial data to estimate time taken until certain events of interest and also the distribution function for time until patient death. This thesis also contains discussion of the above quantities, specifically the state occupancy probabilities and average total sojourn times. Furthermore, [Weiss and Zelen, 1965, Section 4] alludes to some of the ideas mentioned in this thesis – for example they write "It is then plausible to base a test of efficiency of a drug on the amount of time that a patient is kept alive. Another alternative would be to base efficiency on the amount of time that a patient is kept in a state of relative comfort." This is very similar to the idea of patient benefit as described in this section and in Section 3.1 of this thesis. Furthermore, [Weiss and Zelen, 1965, Section 6] show calculations for the density function of the total sojourn time until death, and also the density function of the total sojourn time given passage through a state of remission. Hence, they show expressions for the mean and variance of these quantities. This is similar to calculations we present in Section 4.3.

However, we differentiate our work from [Weiss and Zelen, 1965] in several ways. First, [Weiss and Zelen, 1965] present their methods in the context of "drug comparison" as a general idea, and do not emphasise the difference between drug efficacy and patient benefit as we do. Additionally, a large focus of our work is on statistical inference whereas it was not a focus of [Weiss and Zelen, 1965] at all. Specifically, we have suggested a specific meaning of patient benefit and how to quantify it, and proposed hypothesis tests for inference. As mentioned, we have also proposed a new way to write the likelihood which allows for exact calculations of the observed information matrix and possible theoretical calculation of the expected information matrix. Furthermore, [Weiss and Zelen, 1965] derive their results with slightly different starting points and assumptions from us in most cases. See Section 4.3 for an explanation. Finally, the focus of our work is not on the calculation of these useful quantities but on the application to show patient benefit (or lack thereof).

1.2 Description of thesis contents

The structure and contents of this thesis are as follows.

Chapter 2 introduces the necessary background knowledge and theory. The chapter describes the theory of survival analysis and estimation methods, as well as how it is used in common competing risks models used to analyse clinical data. Specifically, the Cox proportional hazards model and Fine-Gray proportional hazards model ([Fine et al., 2001]) are introduced and compared. Then, semi-Markov processes and how they are used in modelling are discussed. Estimation and methods for simulating data are also discussed.

Chapter 3 defines patient benefit and then discusses the Cox and Fine-Gray models in the context of patient benefit. The contribution is conducting and interpreting the results of a simple simulation study as proposed by BAST Inc. Ltd. The main result is that a larger "treatment effect" of an effective drug is required before the Fine-Gray model will show that patients in active treatment have a (significantly) higher probability of being better off. This is in contrast to the Cox model, where even a small treatment effect will give the result that patients in active treatment are doing relatively (and significantly) better. Also, an example is given to show that even mildly effective drugs will be reported by the Cox model to be effective, even if the drug happens to be highly toxic and causes a substantial rate of premature discontinuation.

Chapter 4 shows how we can use semi-Markov multi-state models to quantify patient benefit. The contributions are:

(i) a new way to write the likelihood using the mixture approach which could potentially be used for theoretical calculations and also to write a closed-form expression for the observed Fisher information matrix,

(ii) a discussion of some computational considerations and proposing the use of the negative of the inverse of the numerical Hessian to estimate the observed Fisher information matrix,

(iii) presenting the probability distribution of a new quantity that might be used to compute average total sojourn times given passage through a specific state before reaching particular states of interest,

(iv) Two proposed hypothesis tests, one based on the quantity in (iii) (**Test A**) and the other based on differences in survival function of holding time in a particular state (**Test B**).

The contribution in Chapter 5 is a detailed simulation study and investigates properties of the aforementioned proposed hypothesis tests given a specific setup. There are three similar models under consideration: (i) with a large amount of detectable benefit and no right-censoring (baseline model), (ii) the baseline model but with added right-censoring, and (iii) a model with less detectable benefit and with censoring. In each case, the correct parametric model is fitted, as well as two misspecified models. Empirical type I error rates are presented, as well as some measures of statistical power. The main results are that **Test A** appears to be relatively robust to misspecification and right-censoring while **Test B** is not. Both tests do not do well in the absence of sufficiently large sample size when there is little detectable benefit. Some of the supplementary quantities of interest as per Chapter 4 are also presented.

The contribution in Chapter 6 is a similar analysis to a real dataset in order to decide whether there is any difference to patient benefit between both two treatment arms: a control treatment and modified treatment. The main result is that, in that study, patients in modified treatment who progress after having completed treatment might be dying at a faster rate that those in control treatment. On the other hand, patients who progress after prematurely discontinuing treatment in the control treatment arm appear to be the ones who are dying at a faster rate. In the latter case however, results are less conclusive due to the relatively small (effective) sample size and a high amount of right-censoring associated with progression after discontinuing treatment.

Chapter 7 then concludes the thesis and offers a discussion regarding some of the finer points related to modelling patient benefit using semi-Markov models.

Chapter 2

Multi-state survival models

This chapter starts by introducing the basic ideas behind multi-state models in Section 2.1, followed by a section discussing basic survival analysis in Section 2.2. After this, the most common examples of competing risk models are introduced in Section 2.3, after which semi-Markov processes are discussed in Section 2.4. Finally, methods and considerations for data simulation are discussed in Section 2.5.

2.1 Multi-state models

Multi-state models are used to describe the evolution of states through time for a subject under observation. More formally, a multi-state model can be described as a general continuous-time stochastic process $(X_t)_{t \geq 0}$ on a state space $\{1, 2, 3, \dots\}$ where X_t is defined as the current state of the subject at time $t \geq 0$.

In a general model there may be one or more possible states to transition to, after which further transitions could be possible until an *absorbing state* (if there is one) is reached. Roughly speaking, an absorbing state is one where no further transitions to other states are possible after it is reached. An example is the state "death" for a patient in a clinical trial.

A competing risks model is one where a subject under observation starts out in a given starting state, then transitions no more than once. The destination state is one of several possible absorbing states which represent mutually exclusive events.

See Figure 2.1 for an example that depicts a possible clinical trial. Section 2.3 gives further details.

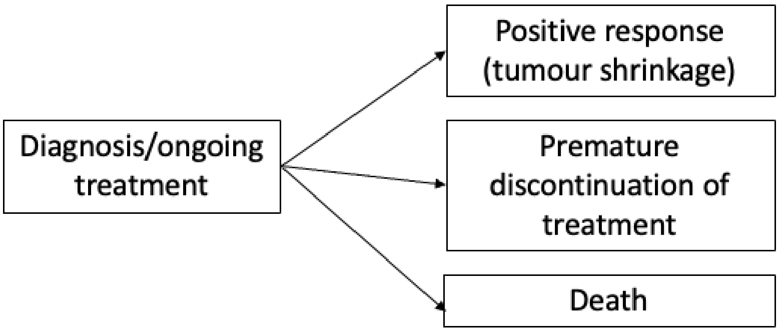


Figure 2.1: An example of a competing risks model depicting a possible clinical trial.

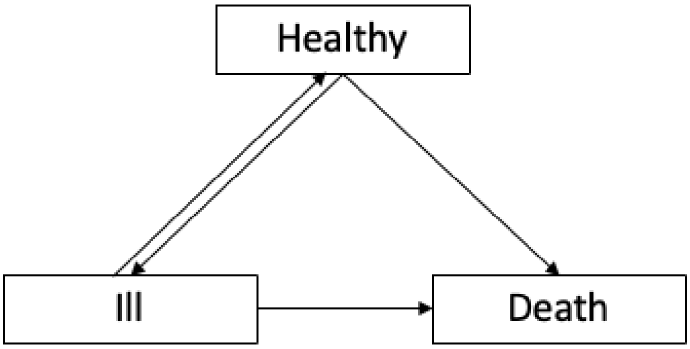


Figure 2.2: The illness-death model as an example of a general multi-state model with bi-directional transitions.

A more general multi-state model might involve several transitions through different states before possibly being absorbed. There is also the possibility to allow for bi-directional transitions between pairs of states. See Figure 2.2 for an example showing a version of the ubiquitous *illness-death model* (see [Touraine, 2019], for example) with the possibility to transition from the states "Healthy" to "Ill" and then back to "Healthy". Section 2.4 discusses such multi-state models where the time between transitions is described by a semi-Markov process.

Example Suppose we label "Healthy", "Ill", "Death" as states "1", "2", and, "3" respectively. Table 2.1 below shows a simple dataset from a version of the illness-death model described in Figure 2.2, where the transition “Ill → Healthy” is not possible.

Individual	From	To	Time (since start of process)
1	1	2	5
1	2	3	11
2	1	3	3
3	1	1	12

Table 2.1: A simple dataset from an illness-death model

Suppose there are three individuals, all starting in state 1. Each row in the dataset represents one possible transition $i \rightarrow j$, with “from” indicating the value of i and “to” indicating the value of j. If “from” and “to” are equal, it means that the individual never transitioned out of the “from” state by the end of the observation period *i.e.* they are *right-censored* (right-censoring is defined in Section 2.2). “Time” denotes the transition time with respect to the start of the process in state 1.

The first individual takes the path $1 \rightarrow 2 \rightarrow 3$ with transitions at times $t = 5$ and $t = 11$ since the start of the process, respectively. The second individual takes the path $1 \rightarrow 3$ at time $t = 3$. The last individual stays in state 1 for $t = 12$ time units and is not observed to leave state 1.

We discuss this example further in Section 2.4.1 and Section 2.4.5.

2.2 Basics of survival analysis

2.2.1 Survival function and hazard function

Throughout this thesis, the event time of an individual is assumed to be continuous. This is a realistic assumption in the context of oncology studies.

Suppose the random variable T denotes the event time of a given individual.

The *survival function* is defined as

$$S(t) = P(T > t) = \int_{t}^{\infty} f(u)\mathrm{d}u = 1 - F(t) \tag{2.1}$$

where f is the probability density function (pdf) of T and $F(t) = P(T \leq t) = \int_{0}^{t} f(u)\mathrm{d}u$ is the corresponding (cumulative) distribution function (cdf). The survival function expresses the probability of the individual being event-free up to time t.

The *hazard function*, denoted $h(t)$, is defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t}. \tag{2.2}$$

The hazard function expresses the rate of incidence of an event in a small interval after time t, given that the individual is event-free up to time t. Using the definition of the numerator in equation (2.2) and rearranging, we can also express the hazard function as

$$h(t) = \left[\lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t)}{\Delta t} \right] \frac{1}{P(T > t)} = \frac{f(t)}{S(t)} \tag{2.3}$$

where the expression in square parentheses is one of the definitions of the pdf of T, $f(t)$.

Since $S(t) = 1 - F(t)$, we have that $\frac{\mathrm{d}}{\mathrm{d}t}S(t) = -f(t)$. Consequently, taking the negative of the (natural) logarithm of $S(t)$ and differentiating with respective to t gives yet another way to express the hazard function since

$$-\frac{\mathrm{d}}{\mathrm{d}t} \log\left(S(t)\right) = \frac{f(t)}{S(t)} = h(t), \tag{2.4}$$

where the last equality is from equation (2.3).

It is also possible to define the *cumulative hazard function*, $H(t)$, where $H(t) = \int_{0}^{t} h(u)\mathrm{d}u$. The cumulative hazard function can be viewed as a measure of "total hazard" up to time t. It is then natural that the survival function can be expressed in terms of the cumulative hazard function. Substituting equation (2.3) into the equation for the cumulative hazard function and using the Fundamental Theorem of Calculus yields

$$H(t) = \int_0^t -\frac{\mathrm{d}}{\mathrm{d}u} \log\left(S(u)\right) \mathrm{d}u = -\big[\log\left(S(t)\right) - \log\left(S(0)\right)\big] = -\log\left(S(t)\right).$$

Multiplying by negative one and exponentiating yields the required expression:

$$S(t) = \exp\left(-H(t)\right). \tag{2.5}$$

Hence,

$$f(t) = -\frac{\mathrm{d}}{\mathrm{d}t} S(t) = h(t) \exp\left(-H(t)\right). \tag{2.6}$$

2.2.2 Estimation

Suppose there are m independent observations of time-to-event data $T_1, T_2, \dots, T_m$ each having a distribution with associated pdf $f(t_j; \boldsymbol{\theta})$ ($j \in \{1, 2, \dots, m\}$) where $\boldsymbol{\theta}$ is a finite-dimensional parameter vector. A given individual's event time T is *right-censored* if we are unable to observe T but know it exceeds a certain value (say C, with its own probability distribution). In other words, for the j^{th} individual, we observe $Y_j = \min(T_j, C_j)$ and $T_j > C_j$ means that the event time is right-censored. In the context of oncology studies, right-censoring can occur if the individual has participated in the study for time C after which the observation period of the study comes to an end without them experiencing any event.

Putting everything together, we can express a given individual's event time and whether or not it is censored by denoting for the j^{th} individual the pair (Y_j, δ_j) where

$$\delta_j = \begin{cases} 1 \text{ if } T_j \leq C_j \\ \\ 0 \text{ if } T_j > C_j \end{cases} \quad \text{and } Y_j = \begin{cases} T_j \text{ if } \delta_j = 1 \\ \\ C_j \text{ if } \delta_j = 0. \end{cases}$$

This setup assumes *independent random censoring i.e.* all random variables $T_1, T_2, \ldots, T_m,\ C_1, C_2, \ldots, C_m$ are independent. [Lawless, 2003, Section 2.2.1.2] describes this censoring scheme as "often realistic".

Non-parametric estimation

A common non-parametric estimator for the survival function is the *Kaplan-Meier estimator* [Kaplan and Meier, 1958], given by

$$\hat{S}(t) = \prod_{i:y_i \leq t} \left(1 - \frac{d_i}{r_i}\right), \tag{2.7}$$

where d_i is the number of individuals observed to have experienced an event by time t and r_i is the number of individuals at risk of experiencing an event in a small interval before time t. The Kaplan-Meier estimator is also the non-parametric maximum likelihood estimator associated with $S(t)$ when T is discrete.

Related to the Kaplan-Meier estimator is the *Nelson-Aalen estimator* ([Nelson, 1969], [Nelson, 1972], [Aalen, 1978]), which is a non-parametric estimator for the cumulative hazard function:

$$\hat{H}(t) = \sum_{i:y_i \leq t} \frac{d_i}{r_i}.$$

While these and other non-parametric methods are of interest, we mainly focus on parametric methods and use non-parametric estimates for plotting figures to visually assess the fit of our parametric models. As such, we will not go into further detail about the properties of any non-parametric methods discussed in this thesis.

Likelihood function and maximum likelihood estimation

The *likelihood function* associated with a sample $\mathbf{Y}$ of size m is

$$L(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{m} f(y_i;\boldsymbol{\theta})^{\delta_i} S(y_i;\boldsymbol{\theta})^{1-\delta_i}. \tag{2.8}$$

Whenever the event time is observed ($\delta_i = 1$), the contribution to the likelihood is the pdf associated with T, $f(t_i;\boldsymbol{\theta})$. Otherwise, the contribution associated with a right-censored observation (when $\delta_i = 0$) is the probability of observing an event time greater than the observed censoring time, *i.e.* $S(c_i;\boldsymbol{\theta})$.

For a given sample $\mathbf{Y} = \mathbf{y}$ and assuming the log-likelihood function, $l(\boldsymbol{\theta}|\mathbf{y}) = \log\big(L(\boldsymbol{\theta}|\mathbf{y})\big)$, is sufficiently regular we can maximise equation (2.8) with respect to $\boldsymbol{\theta}$ to obtain the maximum likelihood estimate, $\hat{\boldsymbol{\theta}}$. As an estimator, under further regularity conditions (see [Arnab, 2017, Section 22.2.1] for more details), it is asymptotically unbiased and consistent. Furthermore, $\sqrt{m}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to \mathrm{N}\big(\mathbf{0}, \mathbf{I}^{-1}(\boldsymbol{\theta}_0)\big)$ in distribution as $m \to \infty$, where $\boldsymbol{\theta}_0$ denotes the true value of $\boldsymbol{\theta}$ and $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Here, $\mathbf{I}(\boldsymbol{\theta})$ is the *expected Fisher information* matrix with (r, s) entry

$$E\left[\left(\frac{\partial l}{\partial \theta_r}\right)\left(\frac{\partial l}{\partial \theta_s}\right)\right] = E\left[-\frac{\partial^2 l}{\partial \theta_s \partial \theta_r}\right].$$

Note that the above equation is only true whenever we have sufficient regularity in the log-likelihood. See Section 5.5 and Section 5.6 of [van der Vaart, 2000] for more details.

In practice, we use that $\hat{\boldsymbol{\theta}} \sim N\big(\boldsymbol{\theta}_0, \mathbf{I}^{-1}(\boldsymbol{\theta}_0)\big)$ approximately for sufficiently large m and, if necessary, we estimate the expected Fisher information matrix with the *observed Fisher information* matrix, $\hat{\mathbf{I}}(\hat{\boldsymbol{\theta}})$, with (r, s) entry

$$-\frac{\partial^2 l}{\partial \theta_s \partial \theta_r}.$$

If we wish to make statistical inference about univariate functions of $\boldsymbol{\theta}$, *e.g.* $g(\boldsymbol{\theta})$, we can invoke the *delta method* to approximate the asymptotic variance of $g(\hat{\boldsymbol{\theta}})$. See [van der Vaart, 2000, Chapter 3] for more details.

Theorem 2.2.1. *(Delta method) Suppose $\hat{\boldsymbol{\theta}}$ is a maximum likelihood estimator of parameter vector $\boldsymbol{\theta}_0$ with parameter space $\boldsymbol{\Theta}$, which satisfies $\sqrt{m}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \approx N(\mathbf{0}, \mathbf{I}^{-1}(\boldsymbol{\theta}_0))$. Suppose $h : \boldsymbol{\Theta} \to \mathbb{R}$ is a differentiable function with non-zero gradient $\nabla h(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_0$. Then,*

$$\sqrt{m}\left(h(\hat{\boldsymbol{\theta}}) - h(\boldsymbol{\theta})\right) \to N\big(0, \{\nabla h(\boldsymbol{\theta})\}^{\top}\mathbf{I}^{-1}(\boldsymbol{\theta})\{\nabla h(\boldsymbol{\theta})\}\big)$$

in distribution as $m \to \infty$.

2.3 Competing risks models

As discussed at the beginning of Section 2.1, a competing risks model is a special case of a multi-state model where there is an initial state with no more than one transition to one of several different absorbing states. More formally, an individual with event time T faces competing risks if he/she can experience only one of K different mutually exclusive events. One approach to model this is to consider the joint probability distribution of $(T_1, T_2, \ldots, T_K)$ and let $T = \min\{T_1, T_2, \ldots, T_K\}$ *i.e.*. take the event time to be that of the first event to occur. This is known as the *latent variable approach.* The issue with this is that $\{T_1, T_2, \ldots, T_K\}$ is likely not a set of independent random variables, and so we cannot establish (or even estimate from typical data) the correlation between them since we never observe more than one of the K events. In literature, this is referred to as the *non-identifiability problem* – see [Cox, 1959] and [Tsiatis, 1975]. Indeed, Prentice *et al.* questioned whether such a setup should be implemented without proper understanding of the interactions between all the possible event times for each individual ([Prentice et al., 1978, Section 3]). More discussion of this point, especially with respect to data simulation, is discussed in Section 2.5.2.

The preferred approach in the literature is to instead consider the bivariate probability distribution of (T, D) where T is the usual event time and D (taking values in $\{1, 2, \ldots, K\}$ in the absence of censoring) is the type of event associated with the event time. A discussion of this joint distribution and related quantities is in Section 2.3.1. Section 2.3.2 and Section 2.3.3 present the Cox proportional hazards model and Fine-Gray proportional hazards model respectively, since these are the most common models used for modelling the influence of covariates on different types of individuals in the context of competing risks.

2.3.1 Cause-specific hazard function and subdistribution hazard function

The *cumulative incidence function* (CIF) associated with event k, denoted F_k, is defined as the joint probability of event k occuring by time $T = t$. Here, events $\{1, 2, \ldots, K\}$ define a collection of mutually exclusive events. The CIF is given by

$$F_k(t) = P(T \leq t, D = k).$$

It is noted that, for each k, F_k is not a proper distribution function since $\lim_{t\to\infty} F_k(t) = P(D = k) < 1$ for all $k \geq 2$. It is for this reason that the CIF is sometimes referred to as the "subdistribution function". However, if $\{1, 2, \ldots, K\}$ is a collectively exhaustive set of events then $\lim\limits_{t\to\infty} \sum\limits_{i=1}^{K} F_i(t) = 1$. We shall assume this throughout the thesis.

It is most common in the competing risks framework to model the hazard functions associated with event times. We can consider either the *cause-specific hazard* (used in the Cox proportional hazards model, detailed in Section 2.3.2) or the *subdistribution hazard* (used in the Fine-Gray proportional hazards model, detailed in Section 2.3.3). The differences between the two different hazard functions are first explored in order to build intuition and elucidate the underlying differences between both models.

The *cause-specific* (CS) hazard is defined as follows (*cf.* equation (2.2)):

$$h_k^C(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t, D = k | T > t)}{\Delta t}. \tag{2.9}$$

The interpretation of this hazard function is the instantaneous rate of incidence of event k for an individual at time t, given that the individual is event-free up to time t. In other words, an individual is treated as "at risk" of event k at time t unless they have already experienced an event. Since the K events of interest are mutually exclusive, we can write the overall hazard function as the sum of the K cause-specific hazards,

$$h(t) = \sum_{i=1}^{K} h_i^C(t). \tag{2.10}$$

Similarly to Section 2.2, we can re-express the CS hazard function by using the definition of the numerator in equation (2.9). Equation (2.9) can then be written as

$$h_k^C(t) = \left[\lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t, D = k)}{\Delta t} \right] \frac{1}{P(T > t)}.$$

We can define the first term in square parentheses as the *subdensity function* for event k, $f_k(t)$, which satisfies $\frac{\mathrm{d}}{\mathrm{d}t} F_k(t) = f_k(t)$. Hence, we can write the CIF in terms of the CS hazard function as:

$$F_k(t) = \int_0^t h_k^C(u) \exp\left(-H(u)\right) \mathrm{d}u = \int_0^t h_k^C(u) \exp\left(- \sum_{i=1}^{K} H_i^C(u) \right) \mathrm{d}u. \tag{2.11}$$

The integrand after the first equality is another expression for the subdensity function (*cf.* equation (2.6)) while the second equality makes use of equation (2.10).

It is noted that the CIF is a function of all K events of interest (through their respective CS cumulative hazard functions), and not just the k^{th} event (through its CS hazard function). Thus, it is impossible to estimate or make any inference about the CIF for any of the k events without first estimating the CS hazard functions for all K events. The manner in which the CIF is decomposed as above is also the reason that an increase in hazard for event k does not necessarily imply an increase in CIF for that event (see [Haller, 2014, Sub-section 3.3.2]).

To further reinforce this point, suppose $S_k(t) = \exp\left(-H_k^C(t)\right)$ is the cause-specific survival function for event k and we wish to estimate it non-parametrically. It is well-established in the literature that naively using $1 - \hat{S}_k(t)$ yields over-estimates of $F_k(t)$ (here, $\hat{S}_k(t)$ uses the Kaplan-Meier estimator as per equation (2.7) except all the times of events $D \neq k$ are treated as right-censored observations). [Putter et al., 2007] highlight this issue, and offer the intuition that this bias arises due to the violation of the assumption that the censoring distribution is independent of the event times $T_1, T_2, \ldots, T_K$. In particular, right-censored event times associated with $D \neq k$ automatically means that event k will never happen. Since individuals are no longer at risk the moment a competing event occurs (despite the Kaplan-Meier estimator assuming they still are), $1 - \hat{S}_k(t)$ overestimates $F_k(t)$ unless $K = 1$ *i.e.* there are no competing risks.

The cause-specific hazard for event k is the basis for the Cox proportional hazard model. However, as demonstrated above, fitting the model to competing risks data only allows us to make inference about the relative rate of incidence associated with event k, and nothing about the absolute risk (unless we model the other $K - 1$ competing events). The lack of direct relationship between the CS hazard for event k and the associated CIF provides the motivation for the subdistribution hazard.

The subdistribution (SD) hazard is defined as

$$h_k^S(t) = \lim_{\mathit{\Delta} t \to 0} \frac{P(t < T \leq t + \mathit{\Delta} t, D = k | \{T > t\} \cup \{T \leq t, D \neq k\})}{\mathit{\Delta} t}. \tag{2.12}$$

The interpretation of this hazard function is the instantaneous rate of incidence of event k for an individual at time t given that either (i) the individual has survived up to time t, or (ii) the individual has experienced a competing event. Individuals are treated as “at risk” of event k at time t even if they have already experienced some other event $D \neq k$. As a result, the subdistribution hazard is agnostic to the occurrences of other events $D \neq k$.

Once again, we can re-express equation (2.12) by using the definition of the numerator and rearranging. The SD hazard function is

$$h_k^S(t) = \left[\lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t, D = k)}{\Delta t} \right] \frac{1}{P(\{T > t\} \cup \{T \leq t, D \neq k\})}.$$

The first term in square parentheses is, once again, $f_k(t)$. By definition of the denominator term on the right, we have

$$P(\{T > t\} \cup \{T \leq t, D \neq k\}) = 1 - F_k(t).$$

Thus, we can write the SD hazard function as

$$h_k^S(t) = \frac{f_k(t)}{1 - F_k(t)}. \tag{2.13}$$

Similarly to how we obtained equation (2.4), we can take the negative logarithm of $1 - F_k(t)$ and differentiate with respect to t and then substitute equation (2.13) to obtain

$$h_k^S(t) = -\frac{\mathrm{d}}{\mathrm{d}t} \log\left(1 - F_k(t)\right), \tag{2.14}$$

which is analogous to equation (2.4).

The relationship as per equation (2.14) now allows us to make direct inference on the CIF for event k using the SD hazard for event k, without having to estimate anything related to the other competing events. This is what we do when we fit the Fine-Gray proportional hazards model to competing risks data. Thus, we can now make inference directly on the absolute risk of an event k through the CIF for event k (which is a probability).

Further discussion of the two approaches to competing risks modelling, including merits and pitfalls, are discussed in detail in Section 2.3.4.

2.3.2 Cox proportional hazards model

As mentioned before, the Cox proportional hazards regression model can be used to model covariate dependence for event times. Suppose that there is only one event of interest *i.e.* there are no competing risks. Given a finite-dimensional vector of covariates $\mathbf{z}$, the Cox proportional hazards model assumes the underlying hazard function associated with event time T takes the form

$$h(t|\mathbf{z}) = h_0(t) \exp(\mathbf{z}^\top \boldsymbol{\beta}) \tag{2.15}$$

for parameter vector $\boldsymbol{\beta}$ whose values determine covariate influence on the hazard function. This specification is known as *semi-parametric* because the function h_0 is non-parametric while $\exp(\mathbf{z}^\top \boldsymbol{\beta})$ depends on parameter vector $\boldsymbol{\beta}$. The function h_0 is called the *baseline hazard*, because it is the hazard at time t for an individual with $\mathbf{z} = \mathbf{0}$.

The model gets its name because, for two individuals i and j with covariate vectors $\mathbf{z}_i$ and $\mathbf{z}_j$ respectively,

$$\frac{h(t|\mathbf{z}_j)}{h(t|\mathbf{z}_i)} = \exp\left((\mathbf{z}_j - \mathbf{z}_i)^\top \boldsymbol{\beta}\right). \tag{2.16}$$

In other words, the hazard function changes proportionally with changes in $\mathbf{z}$. Equation (2.16), also known as the *hazard ratio*, gives us a nice interpretation. It can be interpreted as the instantaneous rate of incidence of the event of interest in a sub-population with characteristics given by $\mathbf{z}_j$ who are still at risk of the event, relative to the sub-population with characteristics given by $\mathbf{z}_i$. For example, a hazard ratio of 2 means that the instantaneous rate of incidence of the event for individuals with the characteristics given by $\mathbf{z}_j$ is twice that of the individuals with the characteristics given by $\mathbf{z}_i$. More generally, if the hazard ratio is greater than (resp. less than) unity, it can be inferred that the positive change in covariate values increases (resp. reduces) hazard. A unit hazard ratio implies that the hazard does not change with the covariate values. It should also be noted that the hazard ratio as per equation (2.16) is independent of time.

Even if h_0 is assumed to be parametric, it is not necessary to specify its parameters if the only concern is estimating covariate effect. This is due to the fact that the parameter vector $\boldsymbol{\beta}$ is estimated by maximising the *partial likelihood function* (usually numerically) with respect to $\boldsymbol{\beta}$. The following informal derivation provides the heuristics.

Suppose there are m distinct (possibly right-censored) event times $\mathbf{y} = (y_1, y_2, \cdots, y_m)^\top$ and define $\mathbf{z}_i$ as the covariate vector of the individual with event time y_i. Using equations (2.6) and (2.9), we can write the likelihood function as

$$\begin{aligned}
L(\boldsymbol{\theta}, \boldsymbol{\beta}|\mathbf{y}) &= \prod_{i=1}^{m} \left[h_0(y_i|\boldsymbol{\theta}) \exp\left(\mathbf{z}_i^\top \boldsymbol{\beta}\right)\right]^{\delta_i} \exp\left(-H_0(y_i|\boldsymbol{\theta}) \exp\left(\mathbf{z}_i^\top \boldsymbol{\beta}\right)\right) \\
&= \prod_{i=1}^{m} \Bigg\{ \left[\frac{h_0(y_i|\boldsymbol{\theta}) \exp\left(\mathbf{z}_i^\top \boldsymbol{\beta}\right)}{\sum\limits_{j \in R_{y_i}} h_0(y_i|\boldsymbol{\theta}) \exp\left(\mathbf{z}_j^\top \boldsymbol{\beta}\right)}\right]^{\delta_i} \left[\sum_{j \in R_{y_i}} h_0(y_i|\boldsymbol{\theta}) \exp\left(\mathbf{z}_j^\top \boldsymbol{\beta}\right)\right]^{\delta_i} \times \\
&\qquad \exp\left(-H_0(y_i|\boldsymbol{\theta}) \exp\left(\mathbf{z}_i^\top \boldsymbol{\beta}\right)\right) \Bigg\}
\end{aligned}$$

where $\boldsymbol{\theta}$ is the vector of parameters associated with the event-time distribution of T. The second equality results from multiplying and dividing by $\left[\sum_{j \in R_{y_i}} h_0(t) \exp(\mathbf{z}_j^\top \boldsymbol{\beta})\right]^{\delta_i}$. Here, R_{y_i} is the *risk set* at time y_i, which is the set of individuals who have yet to experience an event by time y_i and are still under observation at time y_i. This is as per the definition of the risk set in the CS hazard function in equation (2.9). As per Section 2.2, δ_i denotes the censoring status of the individual associated with event time y_i.

Cox argued in [Cox, 1972] that most of the information about $\boldsymbol{\beta}$ is contained within the first term in the product above, while the other terms mainly contain information about $\boldsymbol{\theta}$. Hence, $\boldsymbol{\theta}$ is treated as a nuisance parameter and the second and third terms are ignored. Hence, the partial likelihood function is

$$PL(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{\exp(\mathbf{z}_i^\top \boldsymbol{\beta})}{\sum_{j \in R_{y_i}} \exp(\mathbf{z}_j^\top \boldsymbol{\beta})} \right]^{\delta_i} . \tag{2.17}$$

The partial likelihood can be viewed as the conditional probability of observing an (non-censored) event at time y_i, given the number of individuals in R_{y_i} at risk of that event at that time. Another feature of the partial likelihood is that it does not depend on the event times itself, but on how many events are at risk of happening at time y_i. Also, censored events do not contribute to the partial likelihood.

The setup above assumed distinct event times *i.e.* no tied event times, although tied event times may occur in practice. Various methods have been proposed to address this, including methods found in [Cox, 1972], [Breslow, 1974], [Efron, 1977], and others. We will not discuss these methods here.

Rigorous proofs of the consistency and asymptotic Gaussian distribution of the Cox model estimator $\hat{\boldsymbol{\beta}}$ were shown some years later by A.A. Tsiatis in [Tsiatis, 1981], while P.K. Andersen and R.D. Gill showed simpler proofs of these results using counting process theory in [Andersen and Gill, 1982]. In practice, we carry out inference similarly to how we would for a maximum likelihood estimator derived from a full likelihood function, by using the observed Fisher information associated with the sample.

Competing risks context

Suppose there are K competing events. Suppose, without loss of generality, that we are interested in event $k \in \{1, 2, \ldots, K\}$. The CS hazard function for event k as per the Cox proportional hazards model is given by

$$h_k^C(t|\mathbf{z}) = h_{0,k}(t) \exp(\mathbf{z}^\top \boldsymbol{\beta}_k) \tag{2.18}$$

where $h_{0,k}$ is the baseline hazard for event k and $\boldsymbol{\beta}_k$ is a vector of coefficients associated with event k.

Define the partial likelihood function for event k given m event times,

$$PL_k(\boldsymbol{\beta}) = \prod_{i=1}^{m} \left[\frac{\exp(\mathbf{z}_i^\top \boldsymbol{\beta}_k)}{\sum\limits_{j \in R_{y_i}} \exp(\mathbf{z}_j^\top \boldsymbol{\beta}_k)} \right]^{\delta_{i,k}} \tag{2.19}$$

where $\delta_{i,k} = 1$ if the individual has event time y_i and experiences event k, and 0 otherwise. Hence, event times associated with any event that is not k are treated as right-censored ([Wolbers et al., 2014, p. 2939]).

Note that the partial likelihood function for all events *i.e.* the equivalent of (2.17) is

$$PL(\boldsymbol{\beta}) = \prod_{i=1}^{m} \prod_{l=1}^{K} \left[\frac{\exp(\mathbf{z}_i^\top \boldsymbol{\beta}_l)}{\sum\limits_{j \in R_{y_i}} \exp(\mathbf{z}_j^\top \boldsymbol{\beta}_l)} \right]^{\delta_{i,l}} = \prod_{l=1}^{K} PL_l(\boldsymbol{\beta}). \tag{2.20}$$

Thus, we can fit the CS hazard for any event k of interest in order to compare the relative rates of incidence between two sub-populations without having to maximise the partial likelihood functions for any $D \neq k$.

As mentioned before, a limitation of the cause-specific hazard function (and therefore the Cox model) is that it only tells us about covariate influence on the relative rates of incidence of an event of interest k, but does not tell us anything about covariate influence on the probability of the same event. This motivates the Fine-Gray model, discussed in the following section.

2.3.3 Fine-Gray proportional hazards model

As mentioned in Section 2.2, the main appeal of modelling subdistribution hazard functions is being able to make use of the one-to-one relationship between the CIF and the subdistribution hazard as per equation (2.13).

Similarly to the Cox model, proportional hazards is also assumed in the Fine-

Gray setup but at the SD hazards level, *i.e.*

$$h_k^S(t|\mathbf{z}) = h_{0,k}^S(t)\exp(\mathbf{z}^\top \boldsymbol{\gamma}_k)$$

where $\boldsymbol{\gamma}_k$ is the parameter vector associated with covariate influence when considering event k. This is similar to the model specified in equation (2.18), with the main difference being the interpretation of the parameter vector $\boldsymbol{\gamma}_k$ and associated hazard ratio. The hazard ratio now gives the effect of covariates on the relative change in probability of event k (as opposed to the effect of the covariate on relative rate of incidence of event k in the Cox model). For example, a (SD) hazard ratio of greater than unity now implies a positive relationship between the covariate vector and the probability of event k. Note however that, unlike before with the CS hazard ratio, we are not able to directly link the magnitude of change of the SD hazard ratio with the magnitude of change of the CIF. All we are able to do is comment that the direction of the change is the same. This is explored in detail in [Austin and Fine, 2017] and is further discussed in Section 2.3.4.

Model estimation is done by maximising a partial likelihood function as before. However, since the risk set now involves individuals who have experienced a competing event, the exact form of the partial likelihood depends on the model assumptions about censoring. This is because, since we keep all non-censored individuals in the risk set, even individuals who have experienced competing events contribute to the likelihood. [Fine and Gray, 1999] present the score function for three cases: namely (i) complete data (no censoring), (ii) "censoring complete" data, and (iii) incomplete data. "Censoring complete" data refers to data where the only source of censoring comes from administrative censoring *i.e.* individuals are right-censored only if they have not been observed to have an event by the time the observation period is over (and not because they became lost to follow-up). In such a case, the potential censoring time is always observable. In cases related to general censoring, an adapted version of *inverse probability of censoring weighting* [Robins and Rotnitzky, 1992] is used to weight the individuals who have experienced a competing event. Suppose one such individual experienced a competing event at time t_κ. We can weight this individual by the probability that the individual is right-censored, given that we know that the individual's potential censoring time is after time t_κ, *i.e.* we use the weight $G(t) = \frac{P(C>t)}{P(C>t_\kappa)}$. If $w(t)$ denotes the weight for this individual, we have that

$$w(t) = \begin{cases} 1 & \text{if } t < t_\kappa \\ \dfrac{\hat{G}(t)}{\hat{G}(t_\kappa)} & \text{if } t \geq t_\kappa \\ 0 & \text{if right-censored} \end{cases}$$

where $\hat{G}$ is the Kaplan-Meier estimator for the survival function of the censoring time, $P(C > t)$. This weight is time-dependent and decreasing in time.

The following Section 2.3.4 will discuss the merits and pitfalls of each of the Cox and Fine-Gray proportional hazards models.

2.3.4 Discussion about both approaches to competing risks modelling

In the context of drug trials, the results of the Cox model fit are best suited to answer aetiological questions such as "Does the characteristic of 'drug exposure' (the covariate) decrease the relative rate of incidence of disease progression?". However, the Cox model does not offer insight when it comes to changes in relative probability of incidence of an event of interest, which is relevant when trying to address questions related to patient prognosis such as "Does the characteristic of 'drug exposure' reduce the probability of disease progression?". Indeed, questions related to aetiology and prognosis can have very different answers ([Wolbers et al., 2014, p. 2939]). Instead, the Fine-Gray model is better suited to address such questions related to prognosis. Given two individuals, each with a specific set of characteristics, the Fine-Gray model can give an estimate of whether there is a difference in probability of an event of interest occurring between them. It is for this reason that it is argued in [Latouche et al., 2013] that "understanding the effects of covariates on cause-specific hazards and cumulative incidence functions go hand in hand" and that results for both CS hazards and CIFs should be reported together.

The rest of this section brings up several points of discussion related to both of these models, especially in the context of clinical trials.

Discussing the CS hazard ratio alone can be misleading

It has been established in the literature that reporting hazard ratios alone and using its value to make inference about overall risk is misleading and flawed (see, for example, [Sutradhar and Austin, 2018] and [Spruance et al., 2004]). This is due to the fact that there is no clear relationship between the cause-specific hazard for an event and its CIF (as established in Section 2.3.1). [Sutradhar and Austin, 2018] illustrates that, for a given value of the hazard ratio, the actual hazard can take a variety of different values. The intuition given is that the actual hazard depends on the baseline hazard function, which can be small or large in magnitude (and is not reported alongside the hazard ratio since it is not usually estimated).

The Fine-Gray model is less understood and harder to intuitively explain

As mentioned in Section 1.1, the use of the Cox proportional hazards model has become widespread in many areas, especially biostatistics. Hence, the model is well-understood and familiar to most. On the other hand, the Fine-Gray model is not as commonly used and therefore less understood. This is enough of an issue that [Austin and Fine, 2017] surveyed the literature in 2015 for studies which used the Fine-Gray model and "found that many authors provided an unclear or incorrect interpretation of the regression coefficients associated with this model". One of the examples given in the paper is that a publication reported that the SD hazard ratio value of 2.31 meant that "patients with hyponatremia had a 2.31-fold higher risk of cardiovascular events" (which is incorrect, as mentioned in Section 2.3.3). Other examples of publications with similar assertions were also cited. [Austin and Fine, 2017] clarified and discussed why such assertions are, at most, only approximately true.

In a similar vein, a criticism of the Fine-Gray approach is that the risk set is unnatural in the context of actual epidemiological studies [Andersen et al., 2012, p. 868] and thus hard to explain intuitively. This is because individuals who have died of other causes can never be at risk of experiencing event k in practice, and yet such individuals remain in the risk set associated with the model. Consider the case of an oncology study where a patient becomes lost to follow-up at some time t. With respect to the assumptions of the model, it is true that he/she cannot experience (for example) tumour size reduction after that (since we will never observe it). However in reality, although it is never observed, it may be possible that the patient experiences a tumour size reduction after being lost to follow-up. In this way we might find it acceptable to keep such individuals in the risk set (as per the Fine-Gray setup), rather than remove them from the risk set altogether (as per the Cox setup). However, a patient who (for example) dies at some time t before the end of study can never experience tumour size reduction after time t, and so the above justification no longer makes sense.

Despite these issues, it remains the case that the Fine-Gray regression approach allows the convenience of direct estimation and inference of CIFs associated with an event k of interest without having to estimate CIFs associated with the other $K - 1$ events.

Practical aspects of Fine-Gray model

[Austin et al., 2021] reports that, when the Fine-Gray model is used, there are specific scenarios where the sum of estimated CIFs for K competing risks may sum to a number greater than one. The authors advocate for caution when the total probability of all K events is of interest, and suggest instead estimating all the cause-specific hazards using equation (2.10) and then using equation (2.5) to estimate $S(t) = \exp(-\sum_{i=1}^{K} H_k^C(t))$. They found that using this equation, sometimes referred to in literature as the “all-cause survival function”, did not result in this phenomenon.

Perhaps more interestingly, it is known that if the proportional hazards assumption holds for the CS hazard function, it may not hold for the corresponding SD hazard function ([Beyersmann et al., 2009, Section 4.3]). In particular, we can use both equations (2.11) and (2.14) to work out that the CS hazard function is a time-weighted multiple of the SD hazard function:

$$h_k^C(t) = \left[1 + \frac{\sum_{i=1}^K F_i(t) - F_k(t)}{1 - \sum_{i=1}^K F_i(t)}\right] h_k^S(t). \tag{2.21}$$

Using the hazard ratio as per the left-hand side of per equation 2.16 and substituting the above equation (2.21) therefore gives:

$$\frac{h_k^C(t|\mathbf{z}_j)}{h_k^C(t|\mathbf{z}_i)} = \frac{\left[1 + \frac{\sum_{i=1}^K F_i(t|\mathbf{z}_j) - F_k(t|\mathbf{z}_j)}{1 - \sum_{i=1}^K F_i(t|\mathbf{z}_j)}\right] h_k^S(t|\mathbf{z}_j)}{\left[1 + \frac{\sum_{i=1}^K F_i(t|\mathbf{z}_i) - F_k(t|\mathbf{z}_i)}{1 - \sum_{i=1}^K F_i(t|\mathbf{z}_i)}\right] h_k^S(t|\mathbf{z}_i)}. \tag{2.22}$$

Equation (2.22) shows that the CS and SD hazard ratios cannot simultaneously be independent of time. The only scenario where both sets of proportional hazards assumptions are guaranteed to hold is when $K = 1$. This suggests that model misspecification is possible if, for example, the proportional hazards assumption is only true under one framework but the other framework is used instead. In practice, there are diagnostic tests for testing the proportional hazards assumption (see *e.g.* [Grambsch and Therneau, 1994] and [Zhou et al., 2013]).

2.4 Semi-Markov processes

As mentioned in Section 2.1, multi-state models generalise competing risk models by allowing several (possibly bi-directional) transitions until an absorbing state is reached (if one exists). If, in addition, there is an underlying *semi-Markov process* (SMP) governing the state transitions and the times spent in the states then this is a semi-Markov multi-state model. Loosely speaking, a semi-Markov process is an extension of a Markov process. In a Markov process, the probability distribution associated with the next state is determined solely by the current state. We can ignore the time spent in each state since it is well-known that the time spent in each state follows an exponential distribution, which does not retain memory of how long has already been spent in the current state (the "memorylessness" property). Indeed, there are examples in the literature of Markov multi-state models used in clinical studies or for other biomedical purposes (for example, [Jackson et al., 2003], [Ventura et al., 2014], [Smith et al., 2019], [Milic et al., 2021], and more).

In a SMP, the probability distribution of the next state still depends on the current state, but the rate of departures from the current state is allowed to depend on time. For this reason, we can generalise and allow the holding time in each state to be from any appropriate distribution other than exponential. This is a much more flexible and reasonable assumption when considering clinical trials. The semi-Markov process can be viewed as a Markov process when time is ignored *i.e.* only when the exact transition times are considered.

Section 2.4.1 describes the semi-Markov process, with Section 2.4.2 and Section 2.4.3 respectively discussing the mixture approach and intensity transition functions approach to modelling SMPs. Section 2.4.4 describes the relationships of the two approaches in detail. Section 2.4.5 describes maximum likelihood estimation. Finally, Section 2.4.6 shows some simple examples and applications.

2.4.1 Defining the semi-Markov processes

Let $\{J_n\}_{n\geq 0}$ be a homogeneous first-order Markov chain on a state space $\mathcal{S} = \{1, 2, \ldots, l\}$ with associated probability $p_{ij} = P(J_{n+1} = j|J_n = i)$ of transitioning from state i to j $(i \neq j)$. It is assumed that $p_{ii} = 0$ for all i and $p_{ij} = 0$ for all $j \neq i$ whenever i is an absorbing state. Hence, $\sum_{j\neq i} p_{ij} = 1$ if i is not absorbing, and 0 if it is.

Let also $0 = T_0 < T_1 < T_2 < \dots$ be an increasing random sequence of jump times associated with $\{J_n\}_{n\geq 0}$ and, for $n \geq 1$, let $\tau_n = T_n - T_{n-1}$ be the *sojourn (or inter-arrival) time* in J_{n-1} before jumping to J_n. Then, $(X_u)_{u\geq 0}$ is a semi-Markov process (SMP) with states $X_u := J_n$ for $u \in [T_n, T_{n+1})$. Given that the process is in state i at time t, the joint density function associated with reaching state j in a small interval after time t is given by

$$\tilde{f}_{ij}(t) = \lim_{\mathit{\Delta} t \to 0} \frac{1}{\mathit{\Delta} t} P(t < \tau_{n+1} \leq t + \mathit{\Delta} t, J_{n+1} = j | J_n = i). \tag{2.23}$$

Note that each joint distribution is defined by sojourn times, and not the time since the start of the semi-Markov process. The joint distribution associated with the next state is therefore completely divorced from the past history as is the case in a Markov process. It is for this reason that SMPs are sometimes referred to as a “clock reset” process as opposed to a “clock forward” process as in some other stochastic processes (see [Putter et al., 2007, Section 4.2]).

Example (cont'd): Suppose we have a simple dataset as per Table 2.2 from the illness-death model as per Figure 2.2, except here the transition from “ill” to “healthy” is not possible. This is a continuation of the example introduced in Section 2.1. Now, there is an additional column showing the sojourn time, which is the time spent in the last observed state before leaving it, as opposed to the time since the start of the SMP.

Individual	From	To	Time (since start of process)	Sojourn time
1	1	2	5	5
1	2	3	11	6
2	1	3	3	3
3	1	1	12	–

Table 2.2: (Cont'd) A simple dataset from an illness-death model, first described in Section 2.1. There is now an additional column showing the sojourn time in the last observed state.

There are two related ways to characterise an SMP. The ubiquitous approach that is most familiar in the literature is the *intensity transition functions* (ITFs) approach (see, for example, [Meira-Machado et al., 2009, Section 1]). ITFs are analogous to the CS hazard function (equation (2.9)) in that they determine both the next state j conditional on being in state i as well as the time spent in state i. There is also the lesser-known *mixture model* approach as popularised by [Larson and Dinse, 1985], although it has been known for significantly longer than that (see, for example, [Weiss and Zelen, 1965]). This approach involves directly (and separately) modelling the transition probabilities p_{ij} and the sojourn time before transition to state j, conditional on being in state i and transitioning to state j. For brevity and to prevent confusion, we henceforth refer to this method as the "mixture approach".

Our focus is on the mixture approach (Section 2.4.2), although we also discuss the ITFs approach in Section 2.4.3. This is because we recognise ITFs as being widely preferred and commonly used, and would like to clearly express the link between our results (as derived mainly using the mixture approach) and the ITFs approach. Section 2.4.4 concludes the discussion of the two approaches by discussing their relationships.

2.4.2 The mixture approach

We can define a SMP with the sequence $\{(J_n, T_n)\}_{n\geq 0}$ of states and jump times and characterise the SMP by the transition probabilities of the embedded Markov chain as well as the parameters of the sojourn time distributions in each state.

The transition probabilities are often expressed as a *transition probability matrix* $\mathbf{P}$ with the (i, j) entry being p_{ij} and all diagonal entries $p_{ii} = 0$. We now define the distribution function, density function, survival function, and hazard function related to the sojourn time distribution function conditional on transition $i \to j$

being observed. We have

$$\begin{aligned}
F_{ij}(t) &= P(\tau_n \leq t | J_{n-1} = i, J_n = j) \text{ for all } t \geq 0, && (2.24)\\
f_{ij}(t) &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t < \tau_n \leq t + \Delta t | J_{n-1} = i, J_n = j) = \frac{\mathrm{d}}{\mathrm{d}t} F_{ij}(t), && (2.25)\\
S_{ij}(t) &= P(\tau_n > t | J_{n-1} = i, J_n = j) = 1 - F_{ij}(t), && (2.26)\\
\text{and } h_{ij}(t) &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t < \tau_n \leq t + \Delta t | J_{n-1} = i, J_n = j, \tau_n > t). && (2.27)
\end{aligned}$$

All of the above equations are analogous to similar equations seen in Section 2.2.

From equations (2.27) and (2.26), we can obtain

$$\begin{aligned}
h_{ij}(t) &= \frac{f_{ij}(t)}{S_{ij}(t)} && (2.28)\\
\text{and } S_{ij}(t) &= \exp\left(- \int_0^t h_{ij}(u) \mathrm{d}u \right) && (2.29)
\end{aligned}$$

(*cf.* equations (2.3) and (2.5) respectively). Finally, a useful quantity of interest is the probability of staying in state i for at least t time units before leaving it, namely

$$S_i(t) = P(\tau_n > t | J_{n-1} = i) = \sum_{k \neq i} p_{ik} S_{ik}(t). \tag{2.30}$$

This is also known as the survival time of the waiting (or holding) time in state i.

2.4.3 Intensity transition functions

On the other hand, we can characterise a SMP by the parameters associated with intensity transition functions. The intensity transition function for transition $i \to j$ is defined as

$$\tilde{h}_{ij}(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t < \tau_n \leq t + \Delta t, J_n = j | J_{n-1} = i, \tau_n > t). \tag{2.31}$$

As previously mentioned, the ITF is the analogue of the CS hazard function as per equation (2.9). The interpretation of the ITF is now the probability rate of going to a particular state j in a small interval after time t, given that the process is in state i and there are no observed transitions by time t. Similarly to the CS hazard, the intensity transition function is not to be treated like a hazard function associated with a proper probability distribution. Instead, analogous to the relationship between the CS hazard and all-cause hazard as per equation (2.10), summing the intensity transition functions across all $j \neq i$ results in a hazard function associated with a probability distribution. Denote $\tilde{h}_i(t)$ as the hazard function associated with the holding time in state i. Then,

$$\tilde{h}_i(t) = \sum_{k \neq i} \tilde{h}_{ik}(t), \tag{2.32}$$

and so this gives us another way to express the survival function associated with the holding time as per equation (2.30):

$$S_i(t) = \exp\left(- \int_0^t \sum_{k \neq i} \tilde{h}_{ik}(u) \mathrm{d}u \right). \tag{2.33}$$

Since ITFs generalise the ideas met in traditional competing risks analysis, it is natural to ask what the SMP analogue of the cumulative incidence function (CIF) (as per equation (2.11)) is. Define for the SMP,

$$\mathrm{CIF}_{ij}(t) = P(\tau_n \leq t, J_n = j | J_{n-1} = i).$$

Then, using the same ideas as in the derivation of equation (2.11), we have

$$\mathrm{CIF}_{ij}(t) = \int_0^t \tilde{h}_{ij}(u) S_i(u) \mathrm{d}u = \int_0^t \tilde{h}_{ij}(u) \exp\left(- \int_0^u \tilde{h}_i(s) \mathrm{d}s \right) \mathrm{d}u \tag{2.34}$$

where the second equality uses equations (2.32) and (2.33).

2.4.4 Relationships between both approaches

As mentioned at the beginning of Section 2.4, both of these approaches to SMPs are related. Most of the equations in this section and Section 2.4.5 are as per [Asanjarani et al., 2021], but with additional derivations and explanations. In [Asanjarani et al., 2021], the mixture approach is referred to as "Approach I" while the ITFs approach is referred to as "Approach II".

Using the definition of conditional probability on equation (2.31) and manipulating,

$$\begin{aligned}\tilde{h}_{ij}(t) = &\lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t < \tau_n \leq t | J_{n-1} = i, J_n = j, \tau_n > t) \\ &\times \frac{P(\tau_n > t | J_{n-1} = i, J_n = j) P(J_n = j | J_{n-1} = i) P(J_{n-1} = i)}{P(\tau_n > t | J_{n-1} = i) P(J_{n-1} = i)} \\ =& \frac{h_{ij}(t) S_{ij}(t) p_{ij}}{S_i(t)} \qquad (2.35) \\ =& \frac{p_{ij} f_{ij}(t)}{S_i(t)} \qquad (2.36)\end{aligned}$$

where expression (2.35) results from the definition of the transition probability associated with $i \to j$ and equations (2.26) and (2.27). Expression (2.36) results from equation (2.28).

Since we can express $S_i(t)$ with the right-most expression in equation (2.30), it is possible to express an intensity transition function purely in terms of mixture approach quantities. Conversely, it is also possible to express the quantities associated with the mixture approach in terms of intensity transition functions. To do this, first multiply $\tilde{h}_{ij}(t)$ (as per expression (2.36)) by $S_i(t)$ on both sides of the equation and integrate as follows:

$$\int_0^t p_{ij} f_{ij}(u) \mathrm{d}u = \int_0^t \tilde{h}_{ij}(u) S_i(u) \mathrm{d}u. \qquad (2.37)$$

We note that the left side of equation (2.37) is $p_{ij} F_{ij}(t)$. Then, we allow $t \to \infty$.

Since $F_{ij}(t)$ is a proper distribution function, we get

$$p_{ij} = \int_0^\infty \tilde{h}_{ij}(t) S_i(t) \mathrm{d}t = \int_0^\infty \tilde{h}_{ij}(t) \exp\left(- \int_0^t \sum_{k \neq i} \tilde{h}_{ik}(u) \mathrm{d}u \right) \mathrm{d}t \tag{2.38}$$

where the second equality results from equation (2.33). Finally, we can manipulate the equation for the intensity transition function in terms of equation (2.36) to get

$$f_{ij}(t) = \frac{\tilde{h}_{ij}(t) S_i(t)}{p_{ij}} = \frac{\tilde{h}_{ij}(t) \exp\big(- \int_0^t \sum_{k \neq i} \tilde{h}_{ik}(u) \mathrm{d}u \big)}{\int_0^\infty \tilde{h}_{ij}(t) \exp\big(- \int_0^t \sum_{k \neq i} \tilde{h}_{ik}(u) \mathrm{d}u \big) \mathrm{d}t} \tag{2.39}$$

where the second equality results from equations (2.33) and (2.38). Since the sojourn hazard function and distribution functions are both functions of the sojourn time density function, we thus have mixture approach quantities purely as a function of intensity transition functions, as required.

Note that the right side of equation (2.37) is precisely $\mathrm{CIF}_{ij}(t)$ as per equation (2.34). Thus, equation (2.37) gives us an alternative expression for the CIF in terms of mixture approach quantities:

$$\mathrm{CIF}_{ij}(t) = \int_0^t p_{ij} f_{ij}(u) \mathrm{d}u = p_{ij} F_{ij}(t). \tag{2.40}$$

Hence, $\lim_{t \to \infty} \mathrm{CIF}_{ij}(t) = p_{ij}$ (as we also saw in the derivation of equation (2.38)). This gives the useful interpretation that each transition probability associated with the embedded Markov chain of the SMP can be viewed as a "long-term" probability of transitioning from state i to state j (given that the current state is i).

2.4.5 Estimation

First, we introduce notation used for the likelihood functions as per each approach. Suppose we have m individuals under observation during a fixed time interval $[0, \mathcal{T}]$, where $\mathcal{T}$ is possibly random or an observed realisation of a random variable. The states reached by the hth individual are represented by the sequence $\{J_0^{(h)}, J_1^{(h)}, \dots, J_{N^{(h)}}^{(h)}\}$ taking values in the discrete state space $\mathcal{S} = \{1, 2, \dots, l\}$.

Here, $N^{(h)}$ denotes the number of state transitions of the individual up to time $\mathcal{T}$. Sojourn times associated with each of $\{J_0^{(h)}, J_1^{(h)}, \ldots, J_{N^{(h)}}^{(h)}\}$ are represented by $\{\tau_1^{(h)}, \tau_2^{(h)}, \ldots, \tau_{N^{(h)}}^{(h)}\}$. Note that the last state reached by an individual by time $\mathcal{T}$ could be non-absorbing as they might be right-censored there. In this case then we are additionally interested in the time spent in the last observed state, $U^{(h)} = \mathcal{T} - \sum_{i=1}^{N^{(h)}} \tau_i^{(h)}$. We also make the simplifying assumption that all individuals start at the same value of J_0, though this requirement can be relaxed as required.

The key assumption made is that event histories of individuals are independent and so we can write the likelihood $\mathcal{L}$ as

$$\mathcal{L} = \prod_{h=1}^{m} \mathcal{L}^{(h)}. \tag{2.41}$$

Likelihood function for mixture approach

Generalising the ideas in [Larson and Dinse, 1985], [Asanjarani et al., 2021, Section 3.1] gives the likelihood function for each individual as per the mixture approach for the given data $(J_0^{(h)}, J_1^{(h)}, \ldots, J_{N^{(h)}}^{(h)}, \tau_1^{(h)}, \tau_2^{(h)}, \ldots, \tau_{N(h)}^{(h)}, U^{(h)}, \delta^{(h)})$,

$$\mathcal{L}^{(h)} = \left\{ \prod_{k=1}^{N^{(h)}} p_{J_{k-1}^{(h)} J_k^{(h)}} f_{J_{k-1}^{(h)} J_k^{(h)}}(\tau_k) \right\} \left\{ S_{J_{N^{(h)}}^{(h)}}(U^{(h)}) \right\}^{1-\delta^{(h)}}, \tag{2.42}$$

where $\delta^{(h)}$ is a censoring indicator taking value 1 when the individual is observed to reach an absorbing state, otherwise taking value 0. We use equation (2.30) for the second term involving right-censored transitions.

With the likelihood function specified, we use maximum likelihood estimation to estimate a finite-dimensional parameter vector $\boldsymbol{\theta}$ which consists of the transition probabilities and the parameters associated with the distributions of the sojourn time distribution functions. We note that optimising the likelihood function as per equation (2.41) requires constrained optimisation since we have the constraint that the transition probabilities associated with exiting a given non-absorbing state must sum to one.

Example (cont'd): Suppose we have a simple dataset as per Table 2.3 from the illness-death model as per Figure 2.2, except here the transition from "ill" to "healthy" is not possible. This is a continuation of the example introduced in Section 2.1.

Individual	From	To	Time (since start of process)	Sojourn time
1	1	2	5	5
1	2	3	11	6
2	1	3	3	3
3	1	1	12	–

Table 2.3: (Cont'd) A simple dataset from an illness-death model, first described in Section 2.1.

.

Suppose there is reason to believe that the sojourn times associated with each possible transition all follow exponential distributions with respective rate parameters λ_{ij} for each $i \to j$. The sets of parameters associated with each transition $1 \to 2$, $1 \to 3$, and $2 \to 3$ are, respectively, (p_{12}, λ_{12}), (p_{13}, λ_{13}), and (p_{23}, λ_{23}). However, since we must have $p_{12} + p_{13} = p_{23} = 1$, we can simplify the likelihood to be in terms of just the parameters $(p_{12}, \lambda_{12}, \lambda_{13}, \lambda_{23})$:

$$\begin{aligned}\mathcal{L} =& p_{12} f_{12}(5) \times p_{23} f_{23}(6) \times p_{13} f_{13}(3) \times S_1(12)\\ =& p_{12}\lambda_{12}\exp(-5\lambda_{12}) \times \lambda_{23}\exp(-6\lambda_{23}) \times (1-p_{12})\lambda_{13}\exp(-3\lambda_{13}) \times\\ & (p_{12}\exp(-12\lambda_{12}) + (1-p_{12})\exp(-12\lambda_{13}))\end{aligned}$$

where the last term results from equation (2.30). After this, the likelihood can be maximised over the parameter space $\{(p_{12}, \lambda_{12}, \lambda_{13}, \lambda_{23}) : 0 < p_{12} < 1, \lambda_{12} > 0, \lambda_{13} > 0, \lambda_{23} > 0\}$.

Likelihood function for ITFs approach

We can use the fact that the respective integrands in equation (2.37) are equal, to replace $p_{J_{k-1}J_k} f_{J_{k-1}J_k}(\tau_k)$ in equation (2.42). This gives us

$$\mathcal{L}^{(h)} = \left\{ \prod_{k=1}^{N^{(h)}} \tilde{h}_{J_{k-1}^{(h)} J_k^{(h)}}(\tau_k) S_{J_{k-1}^{(h)}}(\tau_k) \right\} \left\{ S_{J_{N^{(h)}}^{(h)}}(U^{(h)}) \right\}^{1-\delta^{(h)}}, \tag{2.43}$$

after which we use equation (2.33) to simplify the likelihood for each individual:

$$\begin{aligned} \mathcal{L}^{(h)} = &\left\{ \prod_{k=1}^{N^{(h)}} \tilde{h}_{J_{k-1}^{(h)} J_k^{(h)}}(\tau_k) \exp\left(- \int_0^{\tau_{J_k^{(h)}}} \tilde{h}_{J_{k-1}^{(h)}}(u) \mathrm{d}u \right) \right\} \\ &\times \left\{ \exp\left(- \int_0^{U^{(h)}} \tilde{h}_{J_{N^{(h)}}^{(h)}}(u) \mathrm{d}u \right) \right\}^{1-\delta^{(h)}}. \end{aligned} \tag{2.44}$$

Now, we use ideas first seen in [Hougaard, 1999] to rewrite equation (2.44) into the form

$$\mathcal{L}^{(h)} = \prod_{\substack{i,j \in \mathcal{S} \\ i \neq j}} \mathcal{L}_{ij}^{(h)}. \tag{2.45}$$

This allows us to separately optimise terms associated with individual transitions $i \to j$ instead of optimising the entire likelihood at once by noting that

$$\mathcal{L} = \prod_{h=1}^{m} \mathcal{L}^{(h)} = \prod_{h=1}^{m} \prod_{\substack{i,j \in \mathcal{S} \\ i \neq j}} \mathcal{L}_{ij}^{(h)} = \prod_{\substack{i,j \in \mathcal{S} \\ i \neq j}} \left\{ \prod_{h=1}^{m} \mathcal{L}_{ij}^{(h)} \right\}. \tag{2.46}$$

In order to achieve the aforementioned separation as per equation (2.45), first define for each h the variables $\tau_{N^{(h)}+1}^{(h)} = U^{(h)}$ and artificial state $J_{N^{(h)}+1}^{(h)} = -1$ to get rid of the censoring indicator in equation (2.44). We obtain

$$\mathcal{L}^{(h)} = \prod_{k=1}^{N^{(h)}+1} \tilde{h}_{J_{k-1}^{(h)} J_k^{(h)}}(\tau_k)^{[k \neq N^{(h)}+1]} \exp\left(- \int_0^{\tau_k^{(h)}} \tilde{h}_{J_{k-1}^{(h)}}(u) \mathrm{d}u \right), \tag{2.47}$$

where $[\cdot]$ is the *Iverson bracket*, taking value 1 when the statement in the bracket is true, 0 otherwise.

We can verify that equation (2.47) is equivalent to equation (2.44) if we note that $\tilde{h}_{ij}(t)$ must be zero by definition whenever state i is absorbing (and so must

$\tilde{h}_i(t)$) and if we define $0^0 \equiv 1$.

Now, the only thing stopping us from grouping terms in the likelihood function by transition $i \to j$ is that the hazard function of the holding time in the exponential term of equation (2.47) requires simplification. To deal with that, first define:

$$\delta_{i\to j}^{k-1,(h)} = \left[\{J_{k-1}^{(h)} = i, J_k^{(h)} = j\}\right] \text{ and } \delta_{i\not\to j}^{k-1,(h)} = \left[\{J_{k-1}^{(h)} = i, J_k^{(h)} \neq j\}\right].$$

Now, we can finally write

$$\begin{aligned}\mathcal{L}_{ij}^{(h)} = \prod_{k=1}^{N^{(h)}+1} &\left\{\tilde{h}_{ij}(\tau_k) \exp\big(- \int_0^{\tau_k^{(h)}} \tilde{h}_{ij}(u)\mathrm{d}u\big)\right\}^{\delta_{i\to j}^{k-1,(h)}} \\ &\times \left\{ \exp\big(- \int_0^{\tau_k^{(h)}} \tilde{h}_{ij}(u)\mathrm{d}u\big)\right\}^{\delta_{i\not\to j}^{k-1,(h)}}. \qquad (2.48)\end{aligned}$$

If we define $\mathcal{L}_{ij}$ as

$$\mathcal{L}_{ij} = \prod_{h=1}^{m} \mathcal{L}_{ij}^{(h)}, \qquad (2.49)$$

then we can use equation (2.49) to optimise the full likelihood as per equation (2.41) by optimising $\mathcal{L}_{ij}$ for each possible $i \to j$.

Example (cont'd): Suppose we have the same data as per Table 2.3 and now want to write the likelihood as per the ITFs approach. Suppose we assume exponential-like intensity transition functions *i.e.* $\tilde{h}_{ij}(t) = \tilde{\lambda}_{ij}$ is constant. Then, according to equations (2.48) and (2.49),

$$\begin{aligned}\mathcal{L}_{12} &= \tilde{\lambda}_{12} \exp\big(- 5\tilde{\lambda}_{12}\big) \times \exp\big(- 3\tilde{\lambda}_{12}\big) \times \exp\big(- 12\tilde{\lambda}_{12}\big) \\ \mathcal{L}_{13} &= \exp\big(- 5\tilde{\lambda}_{13}\big) \times \tilde{\lambda}_{13} \exp\big(- 3\tilde{\lambda}_{13}\big) \times \exp\big(- 12\tilde{\lambda}_{13}\big) \\ \mathcal{L}_{23} &= \tilde{\lambda}_{23} \exp\big(- 6\tilde{\lambda}_{23}\big)\end{aligned}$$

It is then easy to verify that $\mathcal{L}_{12} \times \mathcal{L}_{13} \times \mathcal{L}_{23}$ is equivalent to the likelihood function $\mathcal{L}$ as defined in equations (2.43), (2.44), or (2.47).

Practical aspects of maximum likelihood estimation for each approach

In our work, we have preferred the mixture approach as compared to the ITFs approach for its interpretability and ease in tweaking parameters for the sake of simulation. The main drawback in doing so is that there are many more parameters associated with the likelihood function as per equation (2.42) (mixture approach) as opposed to the likelihood function as per equation (2.47) (ITFs approach). This is due to the fact that we need to specify two sets of parameters for the mixture approach (namely the transition probabilities and parameters of conditional sojourn time distributions), whereas we only need one set of parameters for the ITFs as per the ITFs approach. In particular, [Asanjarani et al., 2021] states that a model with l states may need up to $l^2 - l$ more parameters when the mixture approach is used.

Furthermore, when using the ITFs approach, the separation of the likelihood as per equation (2.45) allows us to ease computational burden by maximising each $\mathcal{L}_{ij}$ (as per equation (2.49)) separately, provided that each transition $i \to j$ does not share parameters with any other transition. It is worth noting that maximising the likelihood as per the mixture approach could be eased by grouping the terms of the likelihood by each state i, and optimising each group separately (again, provided there are no shared parameters between states). This is the approach taken by the authors of the `flexsurv` package [Jackson, 2016]. However, this is still generally more expensive due to the larger number of parameters as compared to the ITFs approach.

Inference

Since we use maximum likelihood estimation to estimate $\boldsymbol{\theta}$, we have similar properties of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ as described in Section 2.2.2.

Section 4.1 shows how the likelihood function as per equation (2.42) can be written in a more general way which might be useful for computing the expected Fisher information matrix analytically, or for computing an exact expression for the observed Fisher information. Definitions from graph theory are used for this purpose. The appeal of closed-form expressions for the entries of the expected Fisher information matrix is of theoretical interest and could be of use for exact calculations of quantities of interest.

It may be considered too tedious to calculate the partial derivatives of the log-likelihood, even for obtaining entries of the observed Fisher information matrix. If so, the numerical estimates of the observed Fisher information matrix work well in practice – see the end of Section 4.1, and also Section 4.2.

2.4.6 Example: Stanford heart transplant data

We consider an example using the Stanford Heart Transplant dataset found in the `survival` package ([Therneau, 2023]) in R. Full details of this dataset are found in [Crowley and Hu, 1977]. There are $m = 103$ individuals in the data, and the data can be modelled using the same illness-death model as seen in previous examples. There are three states: "Alive without transplant (1)", "Alive with transplant (2)" and "Dead (3)". All individuals start at state 1. 69 individuals experience the transition 1 → 2, 30 individuals experience the transition 1 → 3, and 4 individuals are right-censored in state 1. Of the 69 individuals who reach state 2, 45 transition to state 3 while the remaining 24 are right-censored in state 2. All maximisation of likelihood functions was done using the `Rsolnp` package ([Ghalanos and Theussl, 2015]) in R, which uses the optimisation method as described in [Ye, 1987]. The optimisation method uses a sequential quadratic programming (SQP) interior algorithm and is described as a "General Non-linear Optimization Using Augmented Lagrange Multiplier Method" in the package documentation.

For the mixture approach, we can set up a model which can be fully identified with probability p_{12} and sojourn time hazard functions $h_{12}(t)$, $h_{13}(t)$, and $h_{23}(t)$.

Suppose we wish to specify Weibull hazard functions *i.e.*

$$h_{ij}(t) = \frac{a_{ij}}{b_{ij}} \left(\frac{t}{b_{ij}} \right)^{a_{ij}-1}$$

where a_{ij} and b_{ij} are the Weibull shape and scale parameters, respectively, for transition $i \to j$. We compute the likelihood as per equation (2.42) and maximise to obtain the parameter estimates of $(p_{12}, a_{12}, a_{13}, a_{23}, b_{12}, b_{13}, b_{23})$.

For the ITFs approach, suppose we specify Weibull-like intensity transition functions *i.e.*

$$\tilde{h}_{ij}(t) = \frac{\tilde{a}_{ij}}{\tilde{b}_{ij}} \left(\frac{t}{\tilde{b}_{ij}} \right)^{\tilde{a}_{ij}-1}.$$

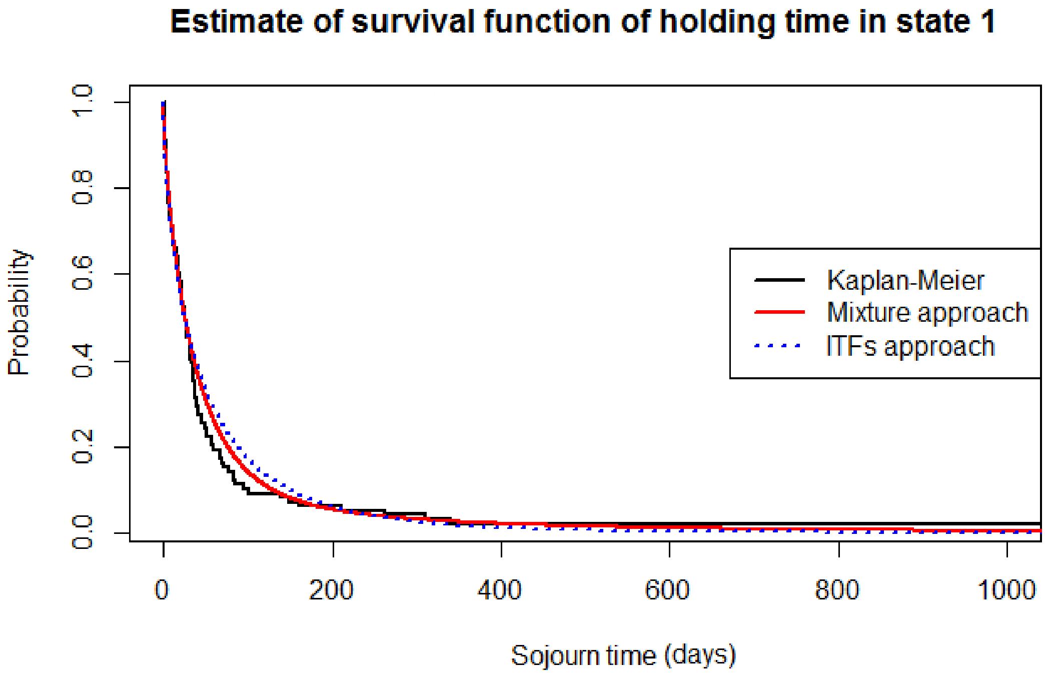
We can then break up the likelihood as per equation (2.48) and maximise to obtain the parameter estimates of $(\tilde{a}_{12}, \tilde{a}_{13}, \tilde{a}_{23}, \tilde{b}_{12}, \tilde{b}_{13}, \tilde{b}_{23})$.

Once we fit the model using both approaches, a fair method to compare both the model fits is to plot estimated survival functions of holding times in each of state 1 and state 2 using the estimated parameters. We can use equation (2.30) and equation (2.33) for $S_i(t)$ specified in terms of mixture approach parameters and ITFs approach parameters, respectively. To visually assess the fits with the data, we can compare each estimate of the survival function of the holding time with an analogue of the Kaplan-Meier estimator as per equation (2.7). For each state i, we have

$$\hat{S}_i(t) = \prod_{\{q:t_q \leq t\}} \left(1 - \frac{d_{i,q}}{m_{i,q}} \right) \tag{2.50}$$

where $d_{i,q}$ is the number of transitions out of state i at a small interval after time t_q, and $m_{i,q}$ is the number of individuals at risk of transition out of state i during a small interval before time t_q.

Figure 2.3 shows plots of the estimates. The black lines are Kaplan-Meier curves as per equation (2.50) while the red line and dotted blue lines are associated



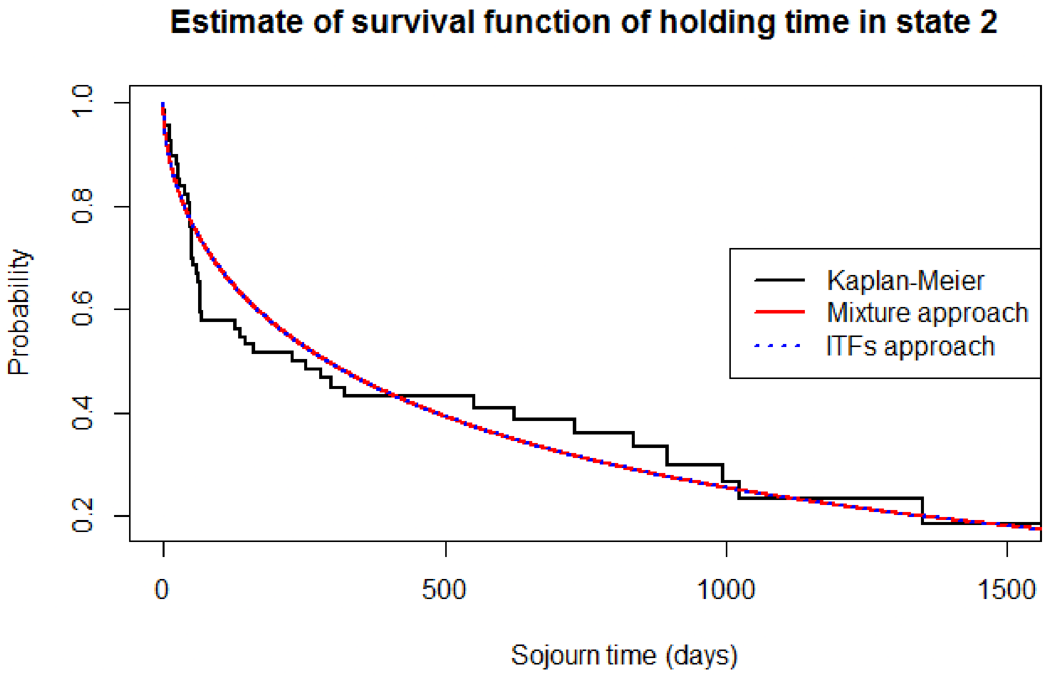


Figure 2.3: Stanford Heart Transplant data fit using both mixture and ITFs approaches: The above figures show estimates of the survival functions of the holding times in state 1 and 2 respectively. The black lines are Kaplan-Meier curves while the red lines are parametric estimates of the respective survival function using equation (2.30). The dotted blue lines are parametric estimates of the respective survival functions using the ITFs approach using equation (2.33). Since state 2 only has one possible transition out of it ($2 \to 3$), the sojourn time hazard function and ITF are both equivalent.

with parametric estimates of the survival functions using the mixture and ITFs approach, respectively. We note that $h_{23}(t) = \tilde{h}_{23}(t)$ because there is only one possible destination state once we exit state 2. The figure of the survival function associated with state 2 indeed shows that estimates of the survival function are essentially identical regardless which approach is taken.

2.5 Simulating data

2.5.1 The mixture approach

To simulate data using the mixture approach, we first need the transition probabilities associated with the Markov chain $\{J_n\}_{n\geq 0}$ on states $\{1, 2, \ldots, l\}$ to decide the next state. Once we have done this, we can simulate an event time associated with the distribution of the conditional sojourn time directly. We have assumed censoring involves first simulating a right-censoring time from a distribution of choice and then checking after every transition whether the censoring time is exceeded. Once the censoring time is exceeded, no further transitions are allowed and the censoring time is the time spent in the last observed (non-absorbing) state. The algorithm is presented below (**Algorithm 1**).

Algorithm 1 Simulating semi-Markov multi-state model data based on mixture approach

```
1: Set values of transition probabilities $p_{ij}$ as well as parameters for $h_{ij}(t)$ for all possible transitions $i \to j$. Decide on the sample size, $m$. For the following steps, start with $k = 1$ and $h = 1$.
2: For individual $h$ at state $J_{k-1}$,
     (i) Set the value of $\mathcal{T}$, which is either fixed or from a distribution of choice. If right-censoring is not required, assume that $\mathcal{T}$ is infinite.
     (ii) Simulate $R$ from Uniform$(0, 1)$ and carry out a multinomial experiment to determine the next state $J_k$ according to probabilities associated with starting at state $J_{k-1}$.
     (iii) Simulate a jump time $T_k$ from the distribution with survival function associated with $h_{J_{k-1}J_k}(t)$.
     (iv) • If $\sum_{i=1}^{k} T_i < \mathcal{T}$ and $J_k$ is not an absorbing state, calculate the sojourn time in the current state as $\tau_k = T_k - T_{k-1}$ and store the values $(J_k, \tau_k)$ (Recall that $T_0 = 0$ by definition). Set $k = k + 1$ and repeat from Step 2(ii).
          • Else if $\sum_{i=1}^{k} T_i < \mathcal{T}$ and $J_k$ is an absorbing state, set $N^{(h)} = k$ and calculate the sojourn time in the current state as $\tau_{N^{(h)}} = T_{N^{(h)}} - T_{N^{(h)}-1}$. Store the values $(J_{N^{(h)}}, \tau_{N^{(h)}})$ and move to Step 3.
          • Else store the value of $U = \mathcal{T} - \sum_{i=1}^{k-1} T_i$ associated with the right-censoring time spent in the most recent state $J_{k-1}$ and move to Step 3.
3: If $h = m$, stop. Else set $h = h + 1$ and repeat from Step 2.
```

Example Suppose we wish to simulate data from an illness-death model with possible states $\{1, 2, 3\}$ and possible transitions $1 \to 2$, $1 \to 3$, and $2 \to 3$. We wish to use **Algorithm 1**. Suppose we want the sojourn times for each transition to be exponential-distributed with (constant) hazard rates $h_{12}(t) = \frac{1}{4}$, $h_{13}(t) = \frac{1}{5}$, and $h_{23}(t) = \frac{1}{3}$ with transition probabilities $p_{12} = 0.3$, $p_{13} = 0.7$. Note we must have $p_{23} = 1$. A single dataset of sample size $m = 1000$ is simulated. The value of $\mathcal{T}$ is chosen to be infinite, *i.e.* the data are non-censored. Figure 2.4 below shows the histograms of sojourn times associated with each of the transitions.

As we might expect based on the chosen probabilities p_{12} and p_{13}, roughly 30% (0.314) of individuals experienced transition $1 \to 2$ while roughly 70% of individuals (0.686) experienced transition $1 \to 3$. The blue lines overlaid in the histograms are the density functions of exponential distributions associated with the sojourn time hazard rates defined above.

2.5.2 The ITFs approach

Simulating data using the ITFs approach is not as straightforward. We could use the latent variable approach which, as discussed at the beginning of Section 2.3, has been criticised in literature. We instead appeal to the underlying idea of the algorithm proposed in [Beyersmann et al., 2009] for simulating data from a competing risks model (with underlying CS hazard functions). The description of the simulation principles can also be found in [Yu, 2015, Section 1.1.2].

Generalising the calculation made by the authors, we calculate the probability that, given we are in state i at time t and a transition has occurred in a small interval after time t, the destination state is j:

$$P(J_n = j | t < \tau_n \leq t + \mathit{\Delta} t, J_{n-1} = i, \tau_n > t) \to \frac{\tilde{h}_{ij}(t)}{\sum\limits_{k \neq i} \tilde{h}_{ik}(t)} \text{ as } \mathit{\Delta} t \to 0 \qquad (2.51)$$

from equations (2.31) and (2.32). Thus, the simulation approach (**Algorithm 2** below) involves first simulating from the distribution associated with equation (2.33)

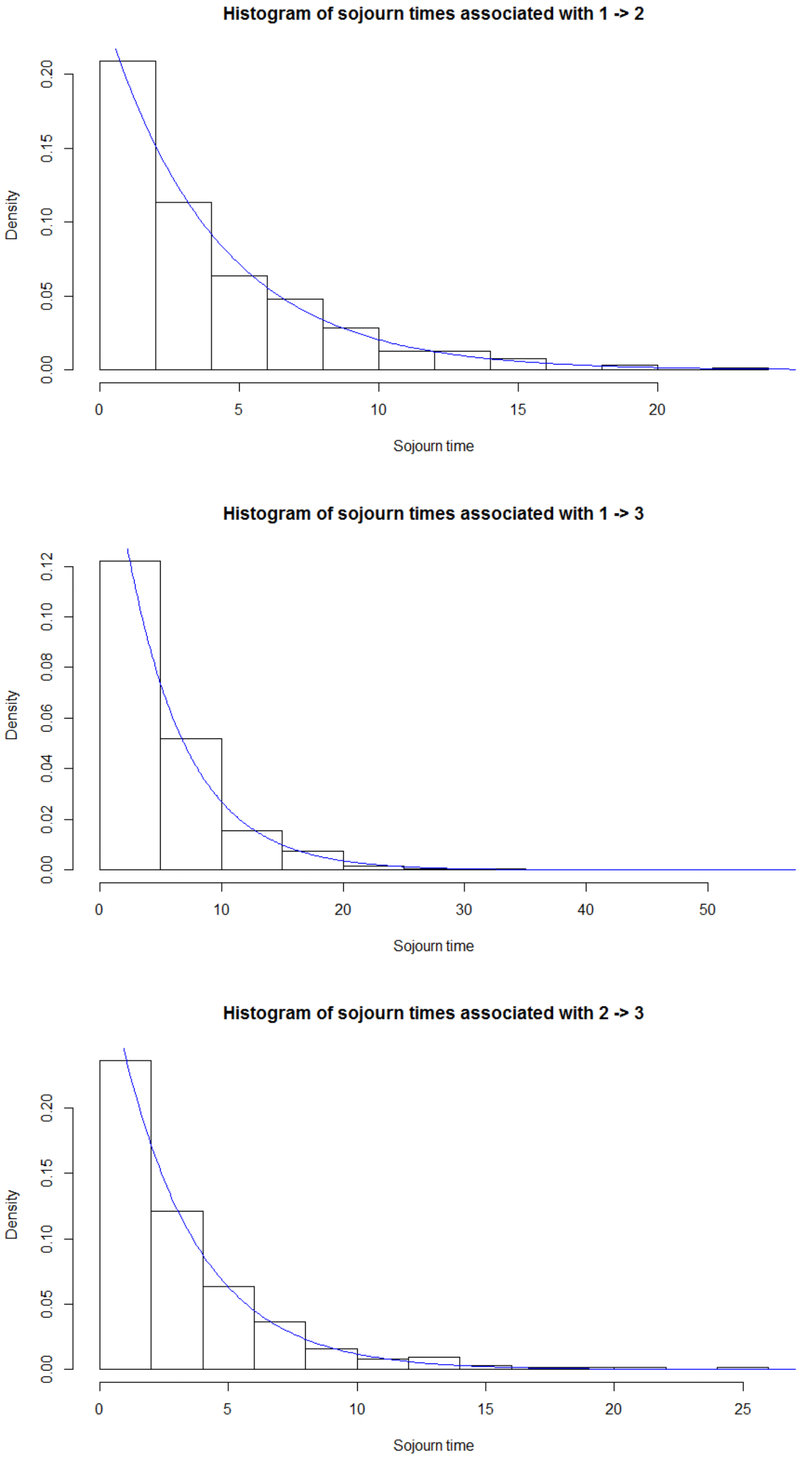


Figure 2.4: Data simulated using mixture approach: Histograms of simulated sojourn times for transitions 1 → 2 (686 individuals) with $p_{12} = 0.3$, 1 → 3 (314 individuals) with $p_{13} = 0.7$, and 2 → 3 (314 individuals) with $p_{23} = 1$, each with exponential distributions with rates $\frac{1}{4}$, $\frac{1}{5}$, and $\frac{1}{3}$ respectively. The blue lines that have been overlaid are the respective density functions $f_{12}(t)$, $f_{13}(t)$, and $f_{23}(t)$.

and then deciding the next state according to a multinomial experiment with probabilities given by equation (2.51).

Algorithm 2 Simulating semi-Markov multi-state model data based on the ITFs approach

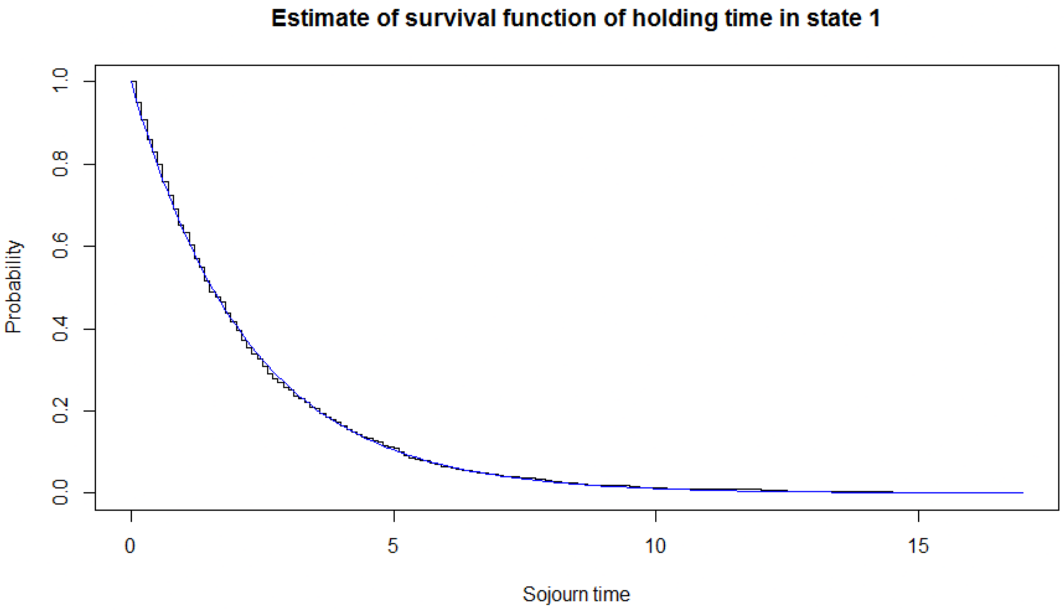
```
1: Set values of parameters associated with $\tilde{h}_{ij}(t)$ for all possible transitions $i \to j$.
   Decide on the sample size, $m$. For the following steps, start with $k = 1$ and $h = 1$.
2: For individual $h$ at state $J_{k-1}$,
     (i) Simulate a jump time $T_k$ from the distribution associated with $S_{J_{k-1}}$
         defined as per (2.33).
     (ii) Simulate $R$ from Uniform$(0, 1)$ and carry out a multinomial experiment
          to determine the next state $J_{k+1}$ according to the probabilities defined in (2.51).
     (iii) • If $T_k < \mathcal{T}$ and $J_k$ is not an absorbing state, calculate the sojourn time
             in the current state as $\tau_k = T_k - T_{k-1}$ and store the values $(J_k, \tau_k)$.
             Set $k = k + 1$ and repeat from Step 2(i).
           • Else if $T_k < \mathcal{T}$ and $J_k$ is an absorbing state, set $N^{(h)} = k$ and calculate
             the sojourn time in the current state as $\tau_{N^{(h)}} = T_{N^{(h)}} - T_{N^{(h)}-1}$. Store
             the values $(J_{N^{(h)}}, \tau_{N^{(h)}})$ and move to Step 3.
           • Else store the value of $U$ associated with the right-censoring time
             spent in the most recent state $J_{k-1}$ and move to Step 3.
3: If $h = m$, stop. Else set $h = h + 1$ and repeat from Step 2.
```

Example Suppose now that we wish to simulate from an illness-death model as per the previous example, but instead wish to use **Algorithm 2**. Suppose we define constant intensity transition functions $\tilde{h}_{12}(t) = \frac{1}{4}$, $\tilde{h}_{13}(t) = \frac{1}{5}$, and $\tilde{h}_{23}(t) = \frac{1}{3}$. with $m = 1000$. Once again, data are non-censored for simplicity.

Since intensity transition functions are not associated with proper probability distributions, a reasonable way to verify the algorithm would be to check that the estimated survival function of the holding time distribution starting from each state i is close to the true survival function given by (2.33). Since the data are non-censored, we can use the empirical distribution function associated with state i to estimate $S_i(t)$.

In this simulated dataset, we find that 553 individuals experience transition $1 \to 2$ and then $2 \to 3$. 447 individuals experience transition $1 \to 3$. Figure 2.5

below shows plots of the estimated survival functions of the holding state in each of state 1 and 2. The blue lines that been overlaid are the respective true functions $S_1(t)$ and $S_2(t)$.



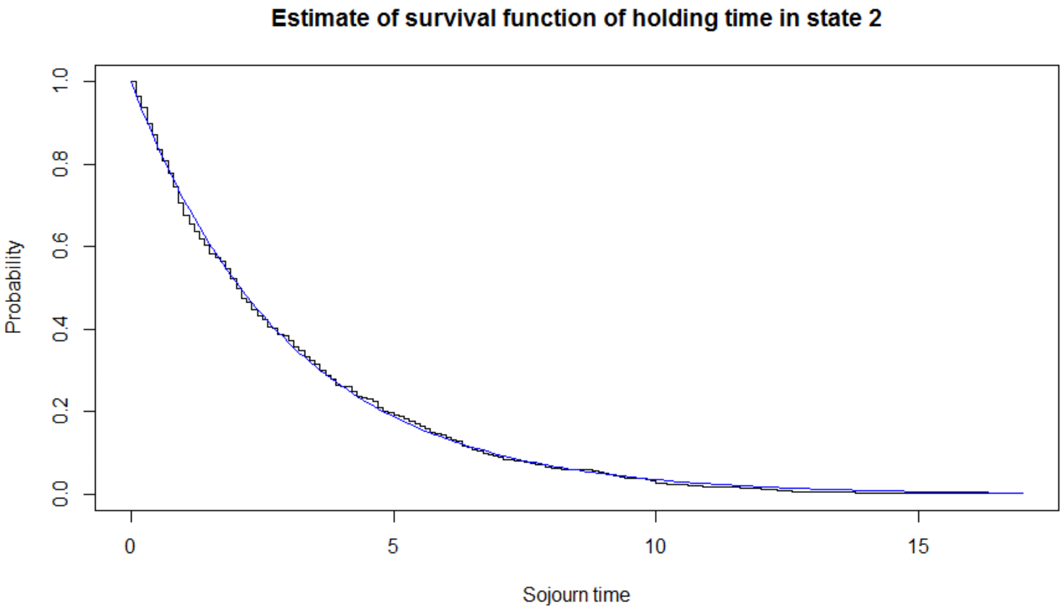


Figure 2.5: Data simulated using ITFs approach: Plots of the estimated survival functions of the holding time in state 1 and state 2 respectively. The survival functions are estimated using empirical survival function. All 1000 individuals experience a transition from state 1, while 553 individuals experience a transition from state 2. The blue lines that have been overlaid are the respective true functions $S_1(t)$ and $S_2(t)$.

2.6 Incorporating covariate effects

The methods by which one can incorporate general covariate effects into semi-Markov multi-state models will not be a focus of this thesis. Our focus will rather be on one categorical covariate of interest, namely $Z = 1$ denoting a patient in active treatment, and $Z = 0$ denoting a patient in control treatment (or any comparable scenario). However, for the sake of completeness and for heuristic purposes, we present an example where one might take a fully parametric approach and specify a "Cox-like" proportional hazards model as per [Meira-Machado et al., 2009]. Further discussion about incorporating multiple covariates in practice is discussed in Section 7.2.

Specifically, in the case of the mixture approach with a covariate vector $\mathbf{Z}$ of length p, we have a sojourn time hazard function of the form

$$h_{ij}(t|\mathbf{Z}) = h_{ij,0}(t) \exp\left(\boldsymbol{\beta}_{ij}^{\top}\mathbf{Z}\right) \tag{2.52}$$

while for the ITFs approach we have

$$\tilde{h}_{ij}(t|\mathbf{Z}) = \tilde{h}_{ij,0}(t) \exp\left(\tilde{\boldsymbol{\beta}}_{ij}^{\top}\mathbf{Z}\right). \tag{2.53}$$

The quantities $h_{ij,0}(t)$ and $\tilde{h}_{ij,0}(t)$ here denote the baseline sojourn time hazard function and baseline intensity transition function respectively. In this setup, each baseline function has its own set of associated parameters $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$ respectively, say. Each of the exponential terms are multiplicative terms showing the effect of covariates, similarly to that seen in the semi-parametric Cox proportional hazards model as per equation (2.18).

The regression coefficients (and hence the hazard ratios) are interpreted differently in the two approaches. The parameter $\boldsymbol{\beta}_{ij}$ affects only the sojourn time hazard for transition $i \to j$ and not the direction of transitions (which is determined solely by p_{ij}). However, since the intensity transition function determines both rate and direction of transitions, $\tilde{\boldsymbol{\beta}}_{ij}$ affects both rate and direction of transitions associated with the SMP.

Computing the likelihood via the ITFs approach is straightforward since one can simply substitute equation (2.53) into equation (2.48) for each individual and proceed. However, we need to make some extra computations to incorporate the hazard function as per equation (2.52) into equation (2.42) correctly. First, we use the relationship as per equation (2.29) and then substitute equation (2.52). After which, we rearrange to get

$$S_{ij}(t|\mathbf{Z}) = \exp\left(-\exp\left(\boldsymbol{\beta}_{ij}^{\top}\mathbf{Z}\right)\int_0^t h_{ij,0}(u)\mathrm{d}u\right) = S_{ij,0}(t)^{\exp\left(\boldsymbol{\beta}_{ij}^{\top}\mathbf{Z}\right)} \tag{2.54}$$

where $S_{ij,0}(t)$ denotes the baseline sojourn time survival function associated with $h_{ij,0}(t)$. Now, we have an expression for the survival function of the holding time in state i given covariates, analogous to equation (2.30),

$$S_i(t|\boldsymbol{Z}) = \sum_{k\neq i} p_{ik}S_{ik}(t|\boldsymbol{Z}) = \sum_{k\neq i} p_{ik}S_{ik,0}(t)^{\exp\left(\boldsymbol{\beta}_{ik}^{\top}\boldsymbol{Z}\right)}. \tag{2.55}$$

Using equation (2.28), we also have an expression for the density function of the sojourn time distribution given covariates,

$$f_{ij}(t|\mathbf{Z}) = h_{ij}(t|\mathbf{Z})S_{ij}(t|\mathbf{Z}) = h_{ij,0}(t)\exp\left(\boldsymbol{\beta}_{ij}^{\top}\mathbf{Z}\right)S_{ij,0}(t)^{\exp\left(\boldsymbol{\beta}_{ij}^{\top}\mathbf{Z}\right)}. \tag{2.56}$$

Now, we can use the likelihood function as per equation (2.42) after substituting equations (2.55) and (2.56).

We note that, after substituting equations (2.55) and (2.56) into equation (2.42), the p_{ij} terms in equation (2.42) do not depend on $\mathbf{Z}$ while the rest of the terms do. On the other hand, [Asanjarani et al., 2021] points out that using the likelihood as per the ITFs approach does allow the probability associated with each transition $i \to j$ to implicitly depend on $\mathbf{Z}$. To address this, [Larson and Dinse, 1985] proposed estimating $p_{ij}(\mathbf{Z})$ using multinomial regression as per [Cox, 1970].

Chapter 3

Competing risk models and patient benefit

Clinical trial data are usually analysed using the Cox proportional hazards model to ascertain how long it takes for a specific treatment of interest to reduce the tumour size of patients, relative to patients in a control group. It has been mentioned in Section 2.3.4 that the Cox proportional hazards model has the limitation that we cannot use the results to make statements about the absolute rate of incidence of events, which would be most useful for answering questions related to patient prognosis. More broadly, it is of interest to ascertain whether most patients can *benefit* from a course of treatment in a more general sense. This is because many kinds of treatments, especially in clinical oncology, can result in undesirable patient outcomes – for example, adverse and serious side effects which lead patients to discontinue treatment. Since such outcomes are undesirable to patients, it is important to adopt modelling approaches which consider all outcomes carefully.

In this chapter, we discuss the basic ideas behind "patient benefit" as we understand it and what they should entail. Section 3.1 describes what patient benefit should entail, while the next two sections use simulation studies to demonstrate these ideas in more detail. In particular, Section 3.2 seeks to show the limitations of the Cox model in the context of patient benefit by comparing it to the Fine-Gray model, which in turn is shown to have limitations. Section 3.3 shows how a drug which is proven effective (according to the Cox model) could bring disbenefit to patients. These two examples naturally lead us in the direction of semi-Markov multi-state models in order to quantify patient benefit.

3.1 Defining patient benefit

Patients could be described as benefiting from a course of cancer treatment if, relatively to patients in a control group (who might be on another type of treatment):

- the treatment under study reduces the rate of transitions to undesirable states which arise as a result of the cancer, and
- the treatment does not dramatically increase the rate of transitions to undesirable states which arise as a result of the treatment or participation in the study.

The first point above encapsulates the "treatment effect" of a drug, which results from the drug's ability to mitigate the cancer. The second point above is related to a drug's toxicity or other factors outside of the cancer itself. Current methods involving the Cox model only consider the first point, when it is necessary to simultaneously consider the second point too. An example in the literature of attempts to consider the second point is [Oberoi et al., 2020], where the authors consider what factors lead to patients being lost to follow-up. Section 3.3 demonstrates why it is crucial to consider both points above together in order to describe patient benefit. It uses simulated data that shows a drug which is found effective (as per the Cox proportional hazards model) but dramatically increases the rate at which patients discontinue prematurely due to adverse side effects, which in turn leads to patients transitioning towards other undesirable states at a faster rate despite the effectiveness of the drug. Informally, patients benefit when the drug is treating the tumour while slowing down the rate of undesirable outcomes.

Tumour shrinkage is usually the main criterion to ascertain drug effectiveness in conventional oncology drug trials, especially in early-phase trials. We note that, as a consequence of the first point in the definition of patient benefit above, the "treatment effect" is no longer defined by tumour shrinkage. Instead, it is defined by the tumour not getting significantly larger (cancer progression). The main reason for this is to take a conservative approach to cancer treatment *i.e.* "No news is good news". More importantly, defining all further states as undesirable to the patient means that patients benefit from treatment if they remain in their current state longer than other patients in the control group. In doing so we extend the idea of [Weiss and Zelen, 1965] to all non-absorbing states in the multi-state model. In other words, we summarise potential patient benefit by determining whether being in active treatment increases the holding times in states of interest compared to being in the control. With this in mind, a natural quantity to study is the survival function of the holding time in state i at time t, $S_i(t) = P(\tau_{n+1} > t | J_n = i)$ as per equation (2.30). Since we can view $S_i(t)$ as the probability of being event-free in state i up to time t, then patients in active treatment at time t in state i are benefiting if they have a higher probability of being event-free up to time t.

However, the survival function of the holding time in particular states may not offer sufficient information or give proper context as to what the drug under study is doing well (or not doing well). Hence, we also propose several other supplementary quantities of interest which can be presented in tandem, such as CIFs, state occupancy probabilities, and other quantities. This is explored in Section 4.3. One of the quantities, the average sojourn time spent in a chosen subset of states, forms the basis of the hypothesis test as proposed in Section 4.4.1.

Since every state in the multi-state model is defined to be undesirable to the patient, one might consider combining two or more states to define a *composite outcome* in a clinical trial. While there are clear benefits, we highlight the importance of modelling the patient history with care so that we may draw reliable conclusions about patient benefit. It is for this reason that we have proposed a hypothesis test and quantities of interest which are unrelated to the survival function of the holding time (as discussed in the previous paragraph). We delay further discussion of composite outcomes until Section 7.1.

To conclude this section, we note that the above notion of patient benefit in this section is defined loosely on purpose. We recognise that, ultimately, our two criteria are qualitative and may not suit every scenario. We believe that the definitions can be adapted to suit the situation, since the main objective most patients want to achieve out of treatment would be to (i) prolong life and (ii) not suffer excessively despite having a prolonged life. The intention of the methods presented in this thesis is to suggest a variety of quantitative criteria to support decision-making when a patient tries to weigh whether it is worth investing part of his/her limited remaining lifespan into a potentially arduous treatment regime.

3.2 Cox model vs Fine-Gray model

Before formally discussing patient benefit, it is necessary to demonstrate the issues associated with current competing risks methods which are commonly used. The Cox model, as previously highlighted, is only useful when considering differences in relative rates of incidence between treatment arms. Even though the following study does not directly make use of the definition of patient benefit as described in Section 3.1, we shall assume a simple "standard" medical trial and show how results can differ when comparing differences between the results of each model.

The following simulation study was designed by BAST Inc. Ltd. The packages `survival` ([Therneau, 2023]) and `cmprsk` ([Gray, 2022]) have been used to fit the Cox and Fine-Gray models respectively.

3.2.1 Setup of study

We consider a simplified oncology study where there are two competing risks, "response" and "dropout". A right-censored observation is one where no event has occurred by the end of the observation period *i.e.* the data are "censoring complete" as described in Section 2.3.3. "Response" signifies the efficacy of treatment while "dropout" is considered to be linked to lack of tolerability *i.e.* intolerable adverse side-effects that lead to patients leaving the study.

Event times for each of response and dropout events are simulated from distributions which have Weibull-like cause-specific hazard functions, with parameter values depending on the particular simulation setup. For each setup, there are $M = 200$ virtual studies. The number of patients, $m \in \{100, 200, 300, 500, 1000, 2000, 4000, 8000\}$, for each setup is also varied.

The hazard function for "response" takes the form

$$h(t) = 0.0015t^{0.15}\exp(\gamma Z). \tag{3.1}$$

Here, Z is a dichotomous drug exposure covariate (1 for above and 0 for below the median), simulated from a Bernoulli(0.5) distribution. The coefficient $\gamma \in \{-2, -1.8, \ldots, 2\}$ signifies the influence of Z on the response. It can be viewed as the strength of treatment effect of the drug, with negative γ signifying that the drug has an adverse effect and positive γ signifying a positive effect.

The "dropout" hazard, on the other hand, takes the form

$$h(t) = l \cdot t^{0.25}\exp(Z) \tag{3.2}$$

where $l \in \{0.0002, 0.0005, 0.001\}$ depending on the setup. It is assumed, for simplicity, that γ has no influence on the rate of drop outs.

It is assumed that each study lasts for 730 days (2 years) and the first tumour assessment takes place on the second month of the study ($t = 60$). Hence, a right-censoring time is generated for each patient by simulating from Uniform$(60, 730)$.

The CS hazard functions' parameter values are chosen to emulate a realistic oncological clinical trial, which can last several years (2 years in this case). Additionally, patients are assumed to have gone through at least one tumour assessment. This is why the lower bound of the Uniform distribution is chosen to be 60 days.

The event time and associated event for each patient are simulated according to **Algorithm 2**. After simulating the relevant datasets, the Cox and Fine-Gray models are fitted to each dataset. For each combination of γ and m, we determine a measure of whether the Cox model is identifying a significant drug effect (whenever it exists) in each of the M studies. One way to do this is:

- for $\gamma < 0$ we measure the proportion of M studies which gives both estimated hazard ratio (HR) less than 1 and p-value $p < 0.05$ for a left-tailed hypothesis test
- for $\gamma = 0$, we measure the proportion of M studies with $p < 0.05$ for a two-tailed hypothesis test
- for $\gamma > 0$, we measure the proportion of M studies which gives both estimated HR greater than 1 and $p < 0.05$ for a right-tailed hypothesis test

Then, for each $\gamma > 0$ and m, we use the results of the Fine-Gray model fits to compute the proportion of M studies which give estimated (Fine-Gray) HR greater than 1 and p-value $p < 0.05$ for a right-tailed hypothesis test. This allows a meaningful comparison of both models whenever there is supposed to be a significant positive drug effect detected *i.e.* "whenever the Cox model considers the treatment to be effective". In essence, for a given positive γ, the proportion of significant studies in the Cox model setup tells us the extent of demonstrable drug efficacy (relative to the control group) while the proportion of significant studies in the Fine-Gray model setup tells us the extent by which patients can have a potential increase in probability of a positive response.

Since we are using statistical power as a performance measure, as per [Morris et al., 2019, Section 5], the *Monte Carlo standard error* (MCSE) for each model setup can be estimated by

$$\text{MCSE} = \sqrt{\frac{\widehat{p_{sig}}(1 - \widehat{p_{sig}})}{M}} \tag{3.3}$$

where $\widehat{p_{sig}}$ is the proportion of significant studies as discussed in the previous paragraph. The Monte Carlo standard error is maximised when $p_{sig} = 0.5$, which gives us a value of approximately 3.536% when $M = 200$. As a balance between computational effort and accuracy of results, we deemed this to be sufficient for our purposes.

There were two main setups considered, namely

(i) $l = 0.0002$ in the dropout hazard to establish a baseline for each combination of γ and m

(ii) increasing l to 0.0005 and then again to 0.001 in the dropout hazard for specific combinations of γ and m to ascertain the effect of an increase in rate of dropouts on the findings

We have noted in Section 2.3.4 that there is almost certainly model misspecification if one has data which satisfies proportional cause-specific hazards and fits a proportional subdistribution hazards model (as per this scenario). However, using the methods described in [Beyersmann et al., 2009], we repeated the study with a similar setup using data simulated from a model which satisfies proportional subdistribution hazards and did not find any changes to the overall conclusions. In the interest of brevity, we will only discuss the results as per the setup which uses data simulated from the model with CS hazard functions as per equations (3.1) and (3.2).

3.2.2 Results and findings

In the baseline setups (where $l = 0.0002$), the main findings are as follows:

- For small number of patients ($m < 1000$), we generally need $\gamma > 0.3$ to demonstrate significant positive treatment effect in over 80% of studies using the Cox model but we generally need $\gamma > 0.5$ to demonstrate an increase in probability of positive response in more than 80% of studies using the Fine-Gray model.

- For large number of patients ($m \geq 1000$) any deviation away from $\gamma = 0$ results in essentially 100% of studies demonstrating significant positive drug effect for the Cox model but we generally need $\gamma > 0.3$ to demonstrate an increase in probability of positive response in at least 80% of the studies using the Fine-Gray model.

In essence, a stronger treatment effect is required in every case when data is fitted to the Fine-Gray model, to demonstrate significant patient benefit *i.e.* the Fine-Gray model says that the patient does not have a significant chance of having an increase in probability associated with the treatment. Intuitively, this is due to the need for the drug effect to overcome the dropout effect, since for small γ dropouts are occurring at a higher rate than responses. Such drugs where the treatment effect significantly outweighs the dropout effect are desirable to patients, and so the relatively conservative nature of the Fine-Gray model is a merit when patient benefit is concerned.

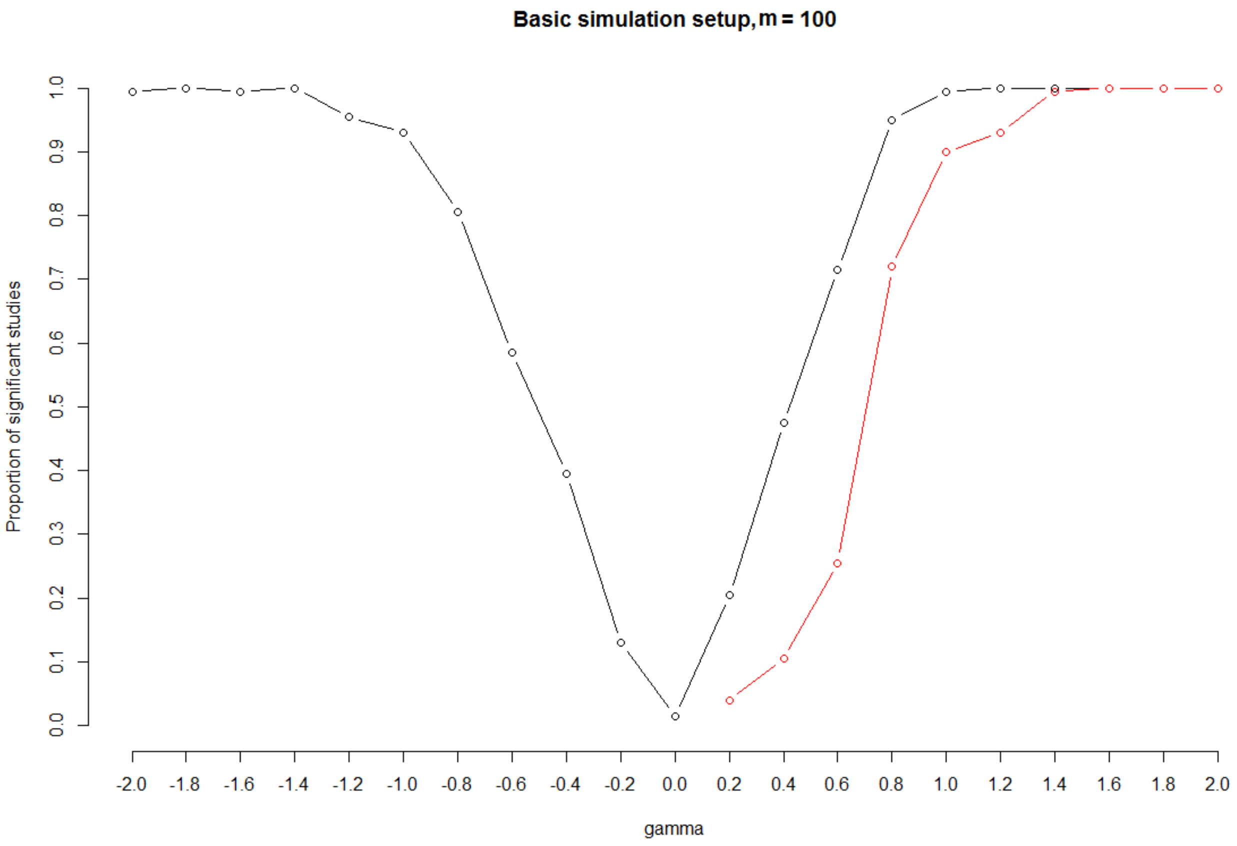
Additionally, and as expected, the so-called "blind spot" region for which the models cannot as easily detect a significant drug effect per the Cox model becomes narrower as the number of patients m increases. A similar observation is made for the "blind spot" for detecting significant patient benefit as per the Fine-Gray model. Thus the influence of exposure to the drug is more easily detected, even for small γ, whenever m is sufficiently large.

Figure 3.1 below shows two plots of "proportion of significant studies" against γ as per the criteria outlined in the beginning of the section. The plots are for $m = 100$ and $m = 8000$.

Increasing the rate of drop outs

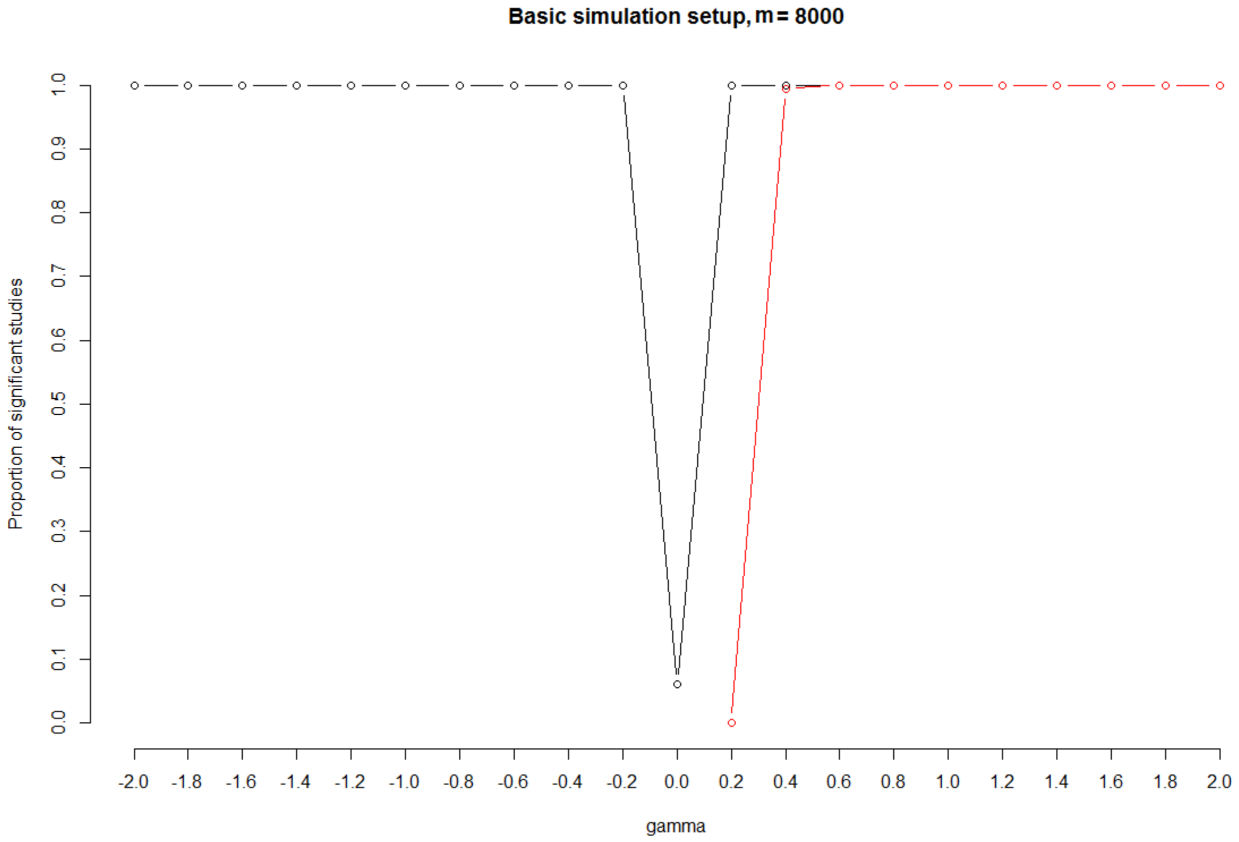
Now, the rate of drop outs is increased through l, by first increasing l to 0.0005 and then to 0.001. To keep things computationally cost-effective, this is only done for $m \in \{100, 4000\}$. The main findings, depicted in Figure 3.2 in a manner similarly to Figure 3.1, are as follows:

- For small number of patients ($m = 100$), increasing the rate of dropouts seems to widen the "blind spot" for both the Cox model and Fine-Gray model.



(a) Studies with basic setup and $m = 100$: The black line is associated with the Cox model, while the red line is associated with the Fine-Gray model. In this case, we need $\gamma > 0.7$ to demonstrate significant positive drug effect in over 80% of studies using the Cox model but need $\gamma > 0.9$ to demonstrate patient benefit in more than 80% of studies using the Fine-Gray model.

Figure 3.1: Basic setup: Results of simulation study comparing Cox and Fine-Gray models ($m = 100$).



(b) Studies with basic setup and $m = 8000$: The black line is associated with the Cox model, while the red line is associated with the Fine-Gray model. In this case, we need $\gamma > 0.1$ to demonstrate significant positive drug effect in over 80% of studies using the Cox model but need $\gamma > 0.3$ to demonstrate patient benefit in more than 80% of studies using the Fine-Gray model. In addition, the region for which either a significant drug effect or significant patient benefit (respectively) cannot be detected has become relatively narrow.

Figure 3.1: Basic setup: Results of simulation study comparing Cox and Fine-Gray models ($m = 8000$).

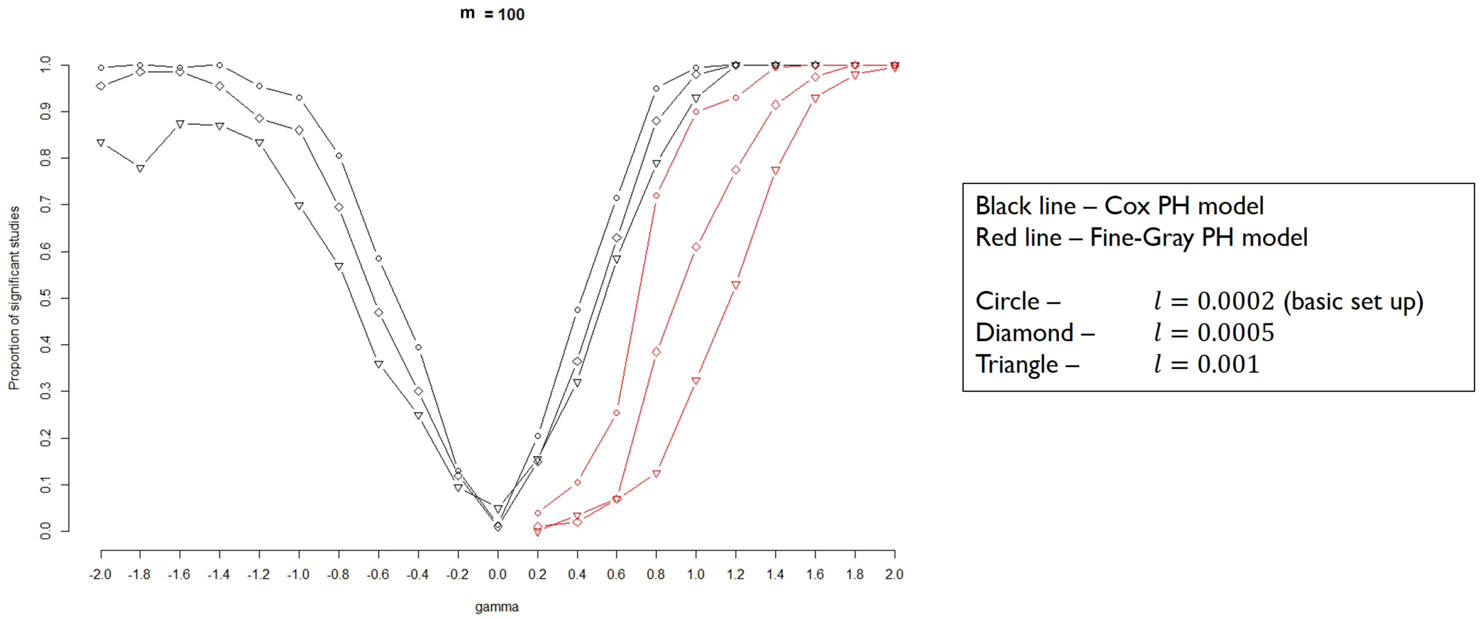
However, the increase in width of the “blind spot” is more dramatic for the case of the Fine-Gray model. For example, to demonstrate significant positive drug effect in over 80% of studies with the Cox model when l increases from 0.0002 to 0.001 requires that γ increase from roughly 0.7 to roughly 0.9. Achieving the same threshold with the Fine-Gray model when l increases from 0.0002 to 0.001 requires that γ increase from 0.9 to 1.5.

- For a large number of patients ($m = 4000$), the increase in “blind spot” for the Cox model as l increases from 0.0002 to 0.001 is negligible, with only a minor change in the region around $\gamma = 0$. Furthermore, we detect positive drug effect in well over 80% of studies under the Cox model regardless when $\gamma > 0.1$. However, there is still significant widening of the “blind spot” for the Fine-Gray model. In this case, to demonstrate patient benefit in over 80% of studies when l increases from 0.0002 to 0.001 requires that γ increase from at least 0.4 to at least 0.8.

Once again, the intuition for the results obtained from the Fine-Gray model stem from the need for the drug effect to overcome the dropout effect, the latter of which is increasing. The Fine-Gray model properly accounts for the increase in drop outs in the risk set to give a measure of absolute change in rate of incidence between both groups of patients, while the Cox model just removes these drop outs from the risk set and treats them as right-censored.

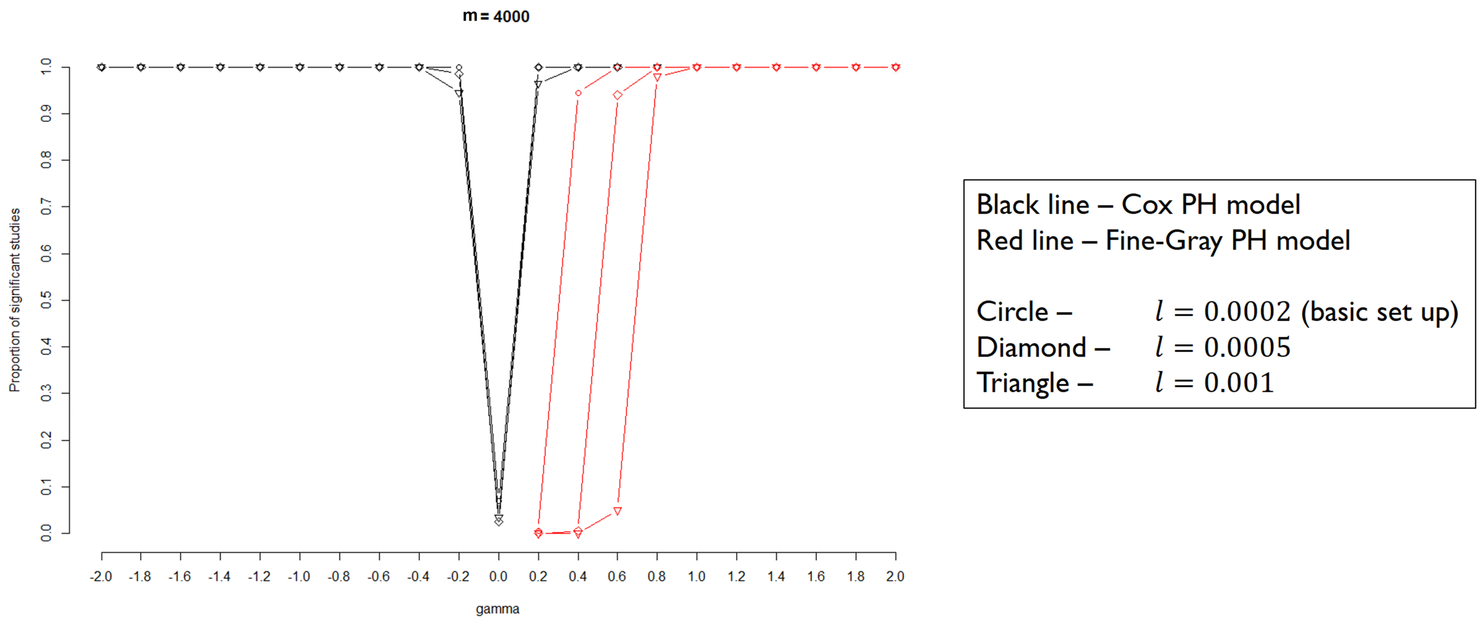
These results demonstrate that the Cox model is useful only for isolating and reporting the effects of the treatment effect on the patient, while ignoring other important factors which influence the patient while they are on treatment. However, there are still issues associated with the Fine-Gray model that are challenging to reconcile, such as the unnatural risk set associated with events such as “death” (mentioned in Section 2.3.4).

Furthermore, more flexibility is required in our model if we are to properly account for possible events that influence patient benefit during the course of treatment. Other extensions of the competing risks setup do ex-



(a) Studies with increased drop out rate and $m = 100$: Increasing the rate of dropouts seems to widen the "blind spot" for both the Cox model and Fine-Gray model. However, the increase in width of the "blind spot" is more dramatic for the case of the Fine-Gray model.

Figure 3.2: Studies with increased drop out rate: Results of simulation study comparing Cox and Fine-Gray models ($m = 100$).



(b) Studies with increased drop out rate and $m = 4000$: The increase in "blind spot" for the Cox model as l increases from 0.0002 to 0.001 is negligible, with only a minor change in the region around $\gamma = 0$. However there is still significant widening of the "blind spot" for the Fine-Gray model, though relatively less than before.

Figure 3.2: Studies with increased drop out rate: Results of simulation study comparing Cox and Fine-Gray models ($m = 4000$).

ist, such as *semi-competing risks* models. For example, [Fine et al., 2001] and [Haneuse and Lee, 2016] discuss how the semi-competing risks setup can be used when the event of interest is a non-absorbing state which can potentially be right-censored by a competing absorbing state (such as "death"). While this setup might suffice for simple clinical trials, it does not address more complex setups where we may have several events of interest that strongly influence patient benefit. For example, we may want to consider "loss to follow-up", "treatment discontinuation" or "death" as events of interest when considering patient benefit. For this reason, we consider more general multi-state models.

We delay discussion of semi-Markov multi-state models to quantify patient benefit until the next chapter, and instead first discuss the survival function of the holding time as a measure to quantify patient benefit in the next section.

3.3 The survival function of the holding time

As mentioned at the beginning of this chapter, we can describe every possible state that the patient can transition to as an undesirable state that should be avoided for as long as possible. In other words, we can think of a drug being beneficial to patients if patients have a higher survival probability of being event-free as compared to patients in control treatment. This makes the survival function of the holding time a natural quantity to consider when assessing patient benefit.

To illustrate this, we simulate a simple clinical trial dataset using **Algorithm 2** from a competing risks model with proportional cause-specific hazard functions. There are $m = 1000$ patients in the starting state called "Diagnosis" (state 1), and three possible destination states: "Relapse" (state 2), "Death" (state 3) and "Dropout" (state 4). There are 159, 220, and 500 transitions to each state respectively. The rest of the 121 patients are right-censored in state 1. Figure 3.3 below depicts the setup.

The underlying cause-specific hazard functions are that of Cox-like proportional hazard functions as per equation (2.53), with Weibull-like baseline hazard functions

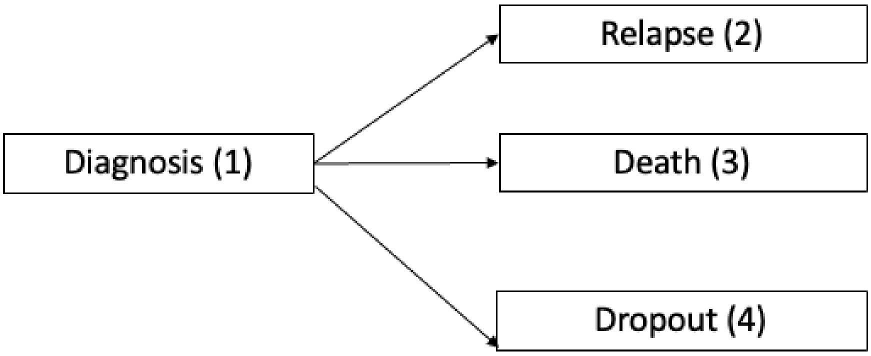


Figure 3.3: Diagrammatic representation of a simple multi-state model for an oncology drug trial

$\tilde{h}_{ij,0}$. We note that the model parameters are chosen such that the drug given to the 500 patients in the active treatment arm ($Z = 1$) is only mildly effective while being highly toxic. This is the reason half of the sample experience transition $1 \rightarrow 4$.

We naively fit a Cox proportional hazards model using the `survival` package [Therneau, 2023] on the data with the event of interest being a relapse, to ascertain if the drug is working. Using similar notation to that in equation (2.53), the reported result is $\hat{\tilde{\beta}} = -0.95$ with associated 95% asymptotic confidence interval $(-1.33, -0.58)$. To contextualise, this is a reliable estimate since the true value of $\tilde{\beta}$ is -0.9 in this case. Based on these results, one could conclude that the drug is effective since, all else being equal, the drug given to patients in active treatment is slowing down the rate of transitions to state 2 as compared to patients in control treatment ($Z = 0$). The first point in our definition of patient benefit as per Section 3.1 is satisfied.

However, this tells us nothing about whether the patient might be benefiting, according to the second point of our definition in Section 3.1. For this, a simple plot of estimated (parametric) survival functions for each treatment arm (Figure 3.4) suffices to show that drug effectiveness is not telling the whole story. At any time t, we can clearly see that patients in active treatment have a lower probability of being event-free. In other words, they have a higher probability of leaving state 1 towards other undesirable states and so are not benefiting. In this particular instance, this is due to the fact that a large number of patients are experiencing transition 1 → 4, with relatively short (holding) event times. The idea of using the survival function of the holding time to quantify patient benefit forms the basis of the hypothesis test proposed in Section 4.4.2.

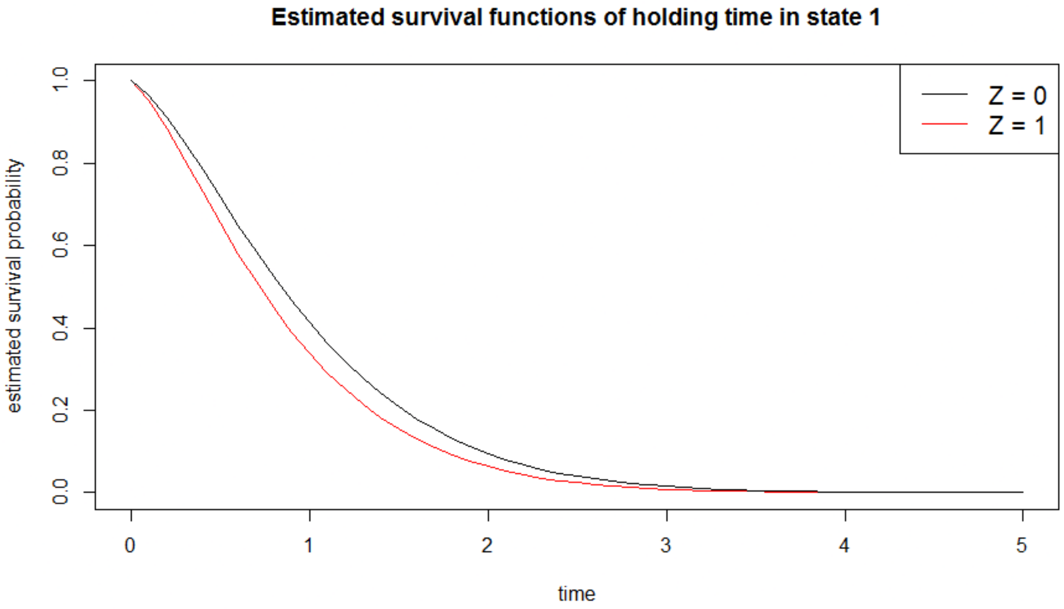


Figure 3.4: The estimated survival functions of the holding time for a parametric model fitted with Weibull-type intensity transition functions using data simulated from a 4-state semi-Markov multi-state model similar to that of Figure 3.3. The red line associated with $Z = 1$ (representing those on a course of particular treatment) is always lower than the black line associated with $Z = 1$, indicating that those on the course of treatment always have a lower probability of being event-free at any time t. This suggests that patients on the treatment are worse off than those who are not on the treatment.

Chapter 4

Quantifying patient benefit using semi-Markov multi-state models

While Chapter 3 has demonstrated some underlying ideas using several examples based on the ITFs approach, we now shift our focus towards the mixture approach. The reason for this is that the mixture approach is more interpretable since the quantities of interest can be expressed in terms of transition probabilities and (conditional) sojourn time hazard functions. This also offers some flexibility *e.g.* for tweaking simulation parameters. While the discussion will be focused on the mixture approach, references to ITFs will be made where appropriate.

At this stage, we make the simplifying assumption that transitions between any two states are uni-directional and that previously visited states cannot be revisted. We also assume that there is at least one absorbing state. In other words, all non-absorbing states are transient. Consequently, we must have a finite number of state transitions before absorption. Furthermore, we also assume that every patient has the same starting state. These assumptions are all reasonable in the context of oncology trials, and can possibly be relaxed and the theory extended if necessary. This point is further discussed in Section 7.3.

Section 4.1 shows how the likelihood function as per the mixture approach (equation (2.42)) can be rewritten to allow for theoretical calculations such as the expected Fisher information, or to provide an exact expression for the observed

Fisher information (after taking partial derivatives). This is a potential alternative to taking numerical approximations of the observed information. Section 4.2 considers some practical considerations when fitting semi-Markov multi-state models, while Section 4.3 introduces some quantities of interest which can give further context when considering patient benefit. Finally, Section 4.4 proposes two new statistical procedures to test for patient benefit.

4.1 A new way to express the mixture approach likelihood function

While the formulation for the likelihood as per equation (2.42) is sufficient for computation in practice, we adopt graph-theoretic notation to write the likelihood in a more general form. Instead of writing the likelihood based on the observed path taken by an individual, we write it in terms of every possible path that could be taken from the starting state by the same individual.

There are examples in the literature representing Markov chains as directed graphs (see [Gingell and Mendivil, 2023], for example) and we take a similar approach below. The reason for using graph-theoretic notation, as alluded to in the previous paragraph, is to incorporate into the likelihood function information about the probabilities of specific "paths" taken by individuals through the evolution of the SMP. This information is crucial to expressing the entries of the Fisher information matrix analytically, but also offers another way to compute the exact observed Fisher information matrix.

As before, denote the homogeneous Markov chain associated with the SMP as $\{J_n\}_{n\geq 0}$ taking values in $\mathcal{S} = \{1, 2, \ldots, l\}$. However, we now formulate the likelihood by making reference to every conceivable transition as opposed to a realisation of the Markov chain $\{J_n\}_{n\geq 0}$. Denote $\mathcal{S}_+ \subset \mathcal{S}$ as the set of states in $\mathcal{S}$ which are not absorbing. Then, denote $\mathcal{V} \subset \mathcal{S}_+ \times \mathcal{S}$ as the ordered pairs of states of the form (i, j) such that transitions are possible *i.e.* with $p_{ij} > 0$. Note that neither $\mathcal{S}_+$ nor $\mathcal{V}$ can be empty since it is assumed that we have a starting state and that we can transition to at least one other state. Throughout we assume, without loss of generality, that the starting state is always $1 \in \mathcal{S}$.

The embedded Markov chain associated with our semi-Markov multi-state model can now be expressed as a directed graph $\mathcal{G} = (\mathcal{S}, \mathcal{V})$ where the state space $\mathcal{S}$ is the set of vertices and $\mathcal{V}$ is the set of directed edges that define all possible transitions. For any (i, j) in $\mathcal{S} \times \mathcal{S}$, the associated matrix of transition probabilities is an $l \times l$ matrix $\mathbf{P}$ with (i, j) entry p_{ij}. Note that $\mathbf{P}$ necessarily has each diagonal entry equal to zero because we have assumed that $p_{ii} = 0$ for all $i \in \mathcal{S}_+$ and also because $p_{ij} = 0$ for all absorbing states $i \in \mathcal{S} \setminus \mathcal{S}_+$. As a consequence, all rows associated with $i \in \mathcal{S} \setminus \mathcal{S}_+$ must have entries all equal to zero. Such a setup makes explicit that individuals must leave towards a different state if they are not trapped in an absorbing state. However, if desired one can always change $\mathbf{P}$ to a proper stochastic matrix by setting $p_{ii} = 1$ for all absorbing states $i \in \mathcal{S} \setminus \mathcal{S}_+$.

We now state several definitions commonly seen in elementary graph theory in order to formally define the notion of a "path" taken by individuals in our setup.

Definition 4.1.1. *(Initial and terminal vertex) Suppose $(i, j) \in \mathcal{V}$. Then, i is the initial vertex and j is the terminal vertex of the (directed) edge (i, j).*

Definition 4.1.2. *(Path from state i to state j) Let $i, j \in \mathcal{S}$ with $i \neq j$. A path from i to j, denoted $r(i, j)$, is (when it exists) a finite sequence of edges (transitions) $\{v_1, v_2, \ldots, v_{n_{ij}}\}$ ($v_k \in \mathcal{V}$ for all $k = 1, 2, \ldots, n_{ij}$) such that (i) the initial vertex of v_1 is i and the terminal vertex of $v_{n_{ij}}$ is j and (ii) the terminal vertex of v_k is the initial vertex of v_{k+1} for $k = 1, 2, \ldots, n_{ij} - 1$. Here, $n_{ij} = |r(i, j)|$ is the number of transitions observed for the the given path $r(i, j)$.*

Definition 4.1.2 makes use of the assumptions of our setup and is slightly different to more general definitions of a path often seen in elementary graph theory. The definition is essentially saying that a path from i to j is made up of a sequence of transitions such that (i) the first state is i, (ii) every destination state after the first transition is a new starting state for the next transition, and (iii) the final state reached is state j.

Example Consider a Markov chain on state space $\mathcal{S} = \{1, 2, 3, 4\}$ and $\mathcal{V} = \{(1, 2), (2, 3), (3, 4)\}$ with $p_{12} = p_{23} = p_{34} = 1$ and all other transitions with probability zero. Then, by Definition 4.1.2, $\{v_1, v_2, v_3\} = \{(1, 2), (2, 3), (3, 4)\}$ describes a path $r(1, 4)$ from state 1 to state 4 with number of transitions $n_{14} = 3$. On the other hand, $\{(1, 2), (3, 4)\}$ does not make up a path from state 1 to state 4.

Since there may be several distinct paths from i to j, we write the list of paths from i to j as $r_1(i, j), r_2(i, j), \ldots, r_{N_{ij}}(i, j)$, where N_{ij} is their number.

Example Suppose $\mathcal{S} = \{1, 2, 3\}$ and we have $p_{12} > 0$, $p_{13} = 1 - p_{12}$, $p_{23} = 1$ with state 3 being absorbing ($p_{31} = p_{32} = 0$). This setup describes the illness-death model similarly to that discussed in Section 2.4.6. In this case, $\mathcal{S}_+ = \{1, 2\}$ and $\mathcal{V} = \{(1, 2), (1, 3), (2, 3)\}$. Since there are two possible paths from state 1 to state 3, we can write, *e.g.* $r_1(1, 3) = \{(1, 3)\}$ and $r_2(1, 3) = \{(1, 2), (2, 3)\}$.

One case we have neglected to account for in our definition of a path is when we have individuals who start in state 1 but are not yet observed to experience the first jump time T_1 due to right-censoring. However, defining a path from state 1 to itself as per Definition 4.1.2 is invalid since $p_{11} = 0$ and therefore $(1, 1) \notin \mathcal{V}$. In such a case, we treat such individuals' path r as unobserved, denoted by $r = \emptyset$.

We also need to introduce the idea of a sub-path.

Definition 4.1.3. *Suppose $i, j \in \mathcal{S}$ and $r(i, j)$ is a path from i to j. A sub-path u of $r(i, j)$ is any other path that makes up a contiguous sub-sequence of $r(i, j)$. We adopt the notation that $u \subseteq r(i, j)$.*

Example If $r(1, 4) = \{(1, 2), (2, 3), (3, 4)\}$ is a path from state 1 to state 4, then $r(1, 2) = \{(1, 2)\}$ and $r(2, 4) = \{(2, 3), (3, 4)\}$ are both sub-paths of $r(1, 4)$.

Note that of course $r(i, j)$ is always a sub-path of itself.

To express the sojourn times for a given transition $(i, j) \in \mathcal{V}$, let τ_{ij} denote a sojourn time in state i before transitioning to state j, conditional on an observed transition from i to j. We define that $\tau_{ij} = 0$ if such a transition is unobserved. Separately, for every $i \in \mathcal{S}_+$, let C_i denote a censoring time in state i, conditional on reaching state i, if the individual is not observed to leave state i by the end of the observation period. We have $C_i = 0$ if the individual is not observed to be censored in state i given he/she has reached it, or if the individual never reaches state i. By definition, at most one of $C_i > 0$ for $i \in \mathcal{S}_+$. The associated censoring indicator for such observations is δ_i. We have that δ_i takes value 1 if the individual is observed to leave state i, or else it takes value 0 otherwise *i.e.* if the individual is observed to reach state $i \in \mathcal{S}_+$ but is not observed to leave it, or if the individual is never observed to reach state i.

Putting it all together, for a given individual h, the data can be expressed as (i) a particular realisation of a path $r = r^{(h)}$ (which is possibly the empty set), (ii) a collection of sojourn times of the form $\tau_{ij} = \tau_{ij}^{(h)}$ for every $(i, j) \in \mathcal{V}$, (iii) a collection of censoring times $C_i = C_i^{(h)}$ for every $i \in \mathcal{S}_+$, and (iv) a collection of censoring indicators $\delta_i = \delta_i^{(h)}$ for every $i \in \mathcal{S}_+$. Using the introduced notation but dropping the superscript (h) for brevity, we write the likelihood for each individual as

$$\begin{aligned}\mathcal{L} = \prod_{\substack{i \in \mathcal{S}_+ \\ i \neq 1}} \Bigg\{ & \Big(p_{1i} f_{1i}(\tau_{1i})\Big)^{\delta_1 [(1,i) \in r]} S_1(C_1)^{(1-\delta_1)} \\ & \times \prod_{j:(i,j) \in \mathcal{V}} \Big(p_{ij} f_{ij}(\tau_{ij})\Big)^{\left(\sum_{z=1}^{N_{1i}} \left[r_z(1,i) \subseteq r\right]\right) \delta_i [(i,j) \in r]} S_i(C_i)^{\left(\sum_{z=1}^{N_{1i}} \left[r_z(1,i) \subseteq r\right]\right)(1-\delta_i)} \Bigg\}. \end{aligned} \tag{4.1}$$

Once again, $[\cdot]$ is the *Iverson bracket*, taking value 1 if the statement within the square brackets is true, 0 otherwise. We adopt the convention that $0^0 := 1$ if this arises when evaluating the likelihood.

The likelihood as formulated above essentially conditions on reaching state $i \neq 1$ before transitioning out to state j for every $i \in \mathcal{S}_+ \setminus \{1\}$ and $j \in \mathcal{S}$. The initial state 1 is treated separately to account for individuals who may not be observed to leave state 1. The power of $p_{1i}f_{1i}(\tau_{1i})$ in the likelihood is associated with an individual starting in state 1 (which is guaranteed by assumption in our setup), then leaving state 1, and then transitioning to state i in one step. Conditional on the observed path r and $i \in \mathcal{S}_+$, these are three independent events with respect to the individual's history up to that point, and therefore the likelihood contribution is $p_{1i}f_{1i}(\tau_{1i})$ only if all three events occur. If the individual is not observed to leave state 1, then the likelihood contribution is given by $S_1(C_1)$. There cannot be contributions associated with both $p_{1i}f_{1i}(\tau_{1i})$ and $S_1(C_1)$ terms since individuals cannot be simultaneously observed to leave state 1 and be right-censored there. For the second line of the equation, the power of $p_{ij}f_{ij}(\tau_{ij})$ is associated with an individual taking one of N_{1i} paths to state i from state 1, then leaving state i, and then transitioning to state j. We have that $[r_z(1, i) \subseteq r] = 1$ only if the particular path $r_z(1, i)$ is a sub-path of the realised path r, and so $\sum\limits_{z=1}^{N_{1i}} [r_z(1, i) \subseteq r]$ is at most 1 since the individual cannot have taken more than 1 distinct sub-path originating in state 1 and ending in state i. Similarly, the power of $S_i(C_i)$ is associated with taking a path from state 1 to state i and whether or not they are observed to leave state i. Again, these are both conditionally independent events with respect to the patient history up to that point, and the likelihood has contributions $S_i(C_i)$ only if both events occurred. Finally, similar to the argument for the first line of the equation, for every $i \in \mathcal{S}_+$ there cannot be likelihood contributions of both $p_{ij}f_{ij}(\tau_{ij})$ and $S_i(C_i)$ simultaneously since an individual cannot take a path from state 1 to state i and then reach state j while simultaneously being censored in state i. For these reasons, the likelihood function as per equation (4.1) reduces to equation (2.42) once a path for the individual is realised, and so the equations are equivalent.

Assuming sufficient regularity in the log-likelihood, we can obtain entries of the observed Fisher information matrix by taking second derivatives of the log-likelihood:

$$
\begin{aligned}
-\frac{\partial^2 \log\mathcal{L}}{\partial\theta_s\partial\theta_r} =& -\sum_{\substack{i\in\mathcal{S}_+\\ i\neq 1}} \delta_1[(1,i)\in r]\frac{\partial^2}{\partial\theta_s\partial\theta_r}\log\Big(p_{1i}f_{1i}(\tau_{1i})\Big) \\
& -(1-\delta_1)\frac{\partial^2}{\partial\theta_s\partial\theta_r}\log\Big(S_1(C_1)\Big) \\
& -\sum_{\substack{i\in\mathcal{S}_+\\ i\neq 1}}\Big(\sum_{z=1}^{N_{1i}}\big[r_z(1,i)\subseteq r\big]\Big)\delta_i[(i,j)\in r]\frac{\partial^2}{\partial\theta_s\partial\theta_r}\log\Big(p_{ij}f_{ij}(\tau_{ij})\Big) \\
& -\sum_{\substack{i\in\mathcal{S}_+\\ i\neq 1}}\sum_{j:(i,j)\in\mathcal{V}}\Big(\sum_{z=1}^{N_{1i}}\big[r_z(1,i)\subseteq r\big]\Big)(1-\delta_i)\frac{\partial^2}{\partial\theta_s\partial\theta_r}\log\Big(S_i(C_i)\Big). \qquad (4.2)
\end{aligned}
$$

It should be possible to compute the expectation of this expression by, for example, conditioning on δ_i and using the law of total expectation. This may require further assumptions about the nature of the censoring process. For example, in this thesis (to simulate data) we have assumed that there is a censoring time from an arbitrary probability distribution generated at the start of the SMP. We then check after every transition whether the sum of sojourn times up to the current state exceeds this censoring time. If it is exceeded, then $C_i > 0$ for some state i where the individual is right-censored. The value of C_i is defined analogously to $U^{(h)}$ as described in the beginning of Section 2.4.5. We could also potentially rewrite equation (4.2) in terms of ITFs parameters by substituting p_{ij}, $f_{ij}(\cdot)$, and $S_i(\cdot)$ with equations (2.38), (2.39), and (2.33) respectively.

In this thesis, all optimisation of the likelihood is done using R software. The `Rsolnp` package ([Ghalanos and Theussl, 2015]) is used to optimise likelihood functions that we have written independently. As previously mentioned, it can be tedious to compute the entries of the Fisher information matrix by using equation (4.2). Instead, we have opted for the convenient option of obtaining the observed information matrix numerically. Specifically, we have used the inverse of the negative of the numerical Hessian matrix (as computed by `Rsolnp` at $\hat{\boldsymbol{\theta}}$) to approximate $\mathbf{I}^{-1}(\hat{\boldsymbol{\theta}})/m$, where $\mathbf{I}(\boldsymbol{\theta})$ is the Fisher information matrix with entries as per equation (4.2). We have found that this gives good results in practice – see Section 4.2.

4.2 Computational considerations

As mentioned in the previous section, it can be tedious to compute the Fisher information matrix with entries as per equation (4.2). The purpose of this section is to demonstrate that using the negative of the numerical Hessian matrix evaluated at $\hat{\boldsymbol{\theta}}$ in place of the observed information gives good results. Note that this section is not intended to exhaustively review what software packages are available for fitting models with the mixture approach.

The negative of the numerical Hessian can be obtained, for example, by writing code for the likelihood function and then choosing an appropriate package that can perform constrained optimisation. We have chosen `Rsolnp` ([Ghalanos and Theussl, 2015]). The other purpose of this section is to discuss the `SemiMarkov` ([Król and Saint-Pierre, 2015]) package used by [Asanjarani et al., 2021] in their estimation. This package reports parameter estimates and their standard errors without the need to write any code for the likelihood function. However, we have found that there are issues with the results obtained using this package, which makes our approach preferable. We elucidate both of the aforementioned points with simulated data. Let us finally mention the package `flexsurv` ([Jackson, 2016]) which allows one to model SMPs using both the mixture and ITFs approach and obtain parameter estimates under a number of model assumptions, with standard errors of estimates reported.

We now describe the setup of the simulation. First, we simulate $M = 1000$ datasets from a five-state model based on the mixture approach. We use empirical standard errors as a measure of performance. We have $\mathcal{S} = \{1, 2, 3, 4, 5\}$, $\mathcal{S}_+ = \{1, 2, 3\}$ and

$$\mathcal{V} = \{(1, 2), (1, 3), (2, 3), (2, 4), (3, 4), (3, 5)\}.$$

The underlying graph structure of this model is as shown in Figure 4.1 below. For each possible transition $(i,j) \in \mathcal{V}$, we have exponential sojourn time hazard functions $h_{ij}(t) = a_{ij}$. The parameter vector is $\boldsymbol{\theta} = (p_{12}, p_{23}, p_{34}, a_{12}, a_{13}, a_{23}, a_{24}, a_{34}, a_{35})^\top = (0.47, 0.32, 0.7, 7, 19, 4, 17, 5, 2)^\top$. The choice of graph structure and model parameters is motivated by the desire to depict a simple model which might have some basis in reality. For example, state 1 would be the starting state, while states 4 and 5 could represent absorbing states such as "Death" and "Lost to follow-up". The transient states 2 and 3 could represent states such as "Disease progression" or "Disease recurrence".

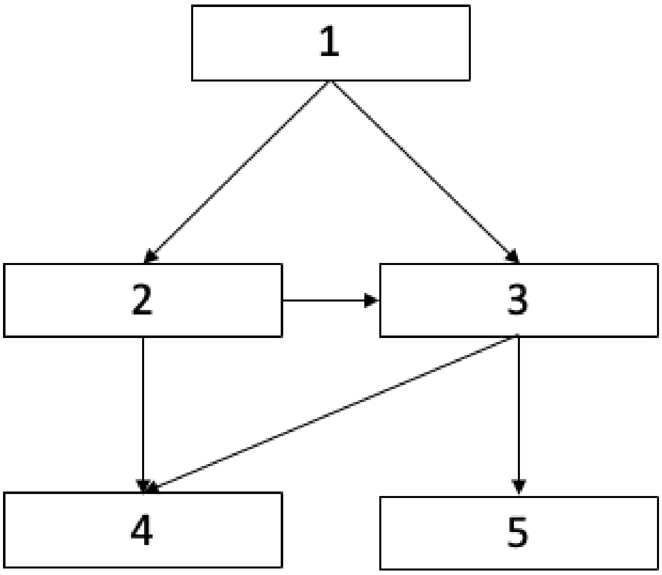


Figure 4.1: The underlying graph structure of the model used for simulation of data. This model is also used for the simulation study in Chapter 5.

There are $m = 10000$ individuals for each of the $M = 1000$ simulated datasets. The data are simulated using **Algorithm 1**. We choose $\mathcal{T}$ from a continuous Uniform(0.5, 1.5) distribution to decide right-censoring time in the last observed state, if applicable. The parameters of the censoring distribution are chosen such that there is a small amount of censoring in state 1, a moderate level of censoring in state 2, and a relatively large amount of censoring in state 3. Specifically, given that individuals reach state i, there is approximately 4%, 8.4%, and 19.5% probability of being censored in state i for $i = 1, 2, 3$ respectively.

According to [Morris et al., 2019, Section 5], the Monte Carlo standard error (MCSE) can be estimated by

$$\text{MCSE} = \sqrt{\frac{\widehat{\text{Var}(\hat{\boldsymbol{\theta}}_i)}}{2(M-1)}} \tag{4.3}$$

where $\text{Var}(\hat{\boldsymbol{\theta}}_i)$ denotes the variance associated with the estimate of the i^{th} component of $\boldsymbol{\theta}$. Based on our results in Table 4.1 (discussed below, and ignoring the anomalous results associated with the `SemiMarkov` package), the MCSE is about 0.7% in the worst case.

Table 4.1 shows the results. The first row shows the true values of the model parameters, while the second and third row give the average of the $M = 1000$ estimates as obtained by `Rsolnp` and `SemiMarkov` respectively. While `SemiMarkov` does well for parameters associated with states 1 and 2, we notice the anomalous results in red for parameters associated with state 3 which do not agree with the true values nor the estimates obtained by using `Rsolnp`.

The fourth row shows the asymptotic standard deviation as calculated by taking the appropriate entry of $\sqrt{\hat{\mathbf{I}}^{-1}(\hat{\boldsymbol{\theta}})/m}$. The fifth row instead estimates the standard errors with the appropriate entry of the square root of the negative of the numerical Hessian obtained with `Rsolnp`, evaluated at $\hat{\boldsymbol{\theta}}$. We note similar results seen in the fourth and fifth rows, justifying the use of the numerical Hessian for computational convenience.

The last row consists of anomalous standard errors as reported by `SemiMarkov`. Despite attempted personal communication with the author, it remains unclear how the standard errors are obtained. To the best of our knowledge, this is not explained in the documentation either. Furthermore, even though `SemiMarkov` appears to use `Rsolnp` as a dependency for optimisation, the results obtained are very different from ours. A further complication is that the standard errors are

Parameters	p_{12}	a_{12}	a_{13}	p_{23}	a_{23}	a_{24}	p_{34}	a_{34}	a_{35}
True value	0.47	7	19	0.32	4	17	0.7	5	2
Mean `Rsolnp` estimate	0.4702	6.9956	19.0014	0.3200	4.0060	17.0132	0.7000	4.9969	2.0031
Mean `SemiMarkov` estimate	0.4702	6.9949	19.0000	0.3200	4.0058	17.0107	0.7433	5.6335	3.1648
Standard errors	p_{12}	a_{12}	a_{13}	p_{23}	a_{23}	a_{24}	p_{34}	a_{34}	a_{35}
Derived using observed FIM	0.0050	0.1064	0.2709	0.0073	0.1184	0.3302	0.0067	0.0951	0.0611
Derived using numerical Hessian reported by `Rsolnp`	0.0051	0.1071	0.2699	0.0073	0.1184	0.3291	0.0067	0.0948	0.0614
Reported by `SemiMarkov`	0.01	0.00	0.00	0.10	1.00	0.00	0.01	0.00	0.01

Table 4.1: Results for parameter estimates and standard errors for a 5-state model with underlying graph structure similarly to that depicted in Figure 4.1. The reported results are the mean of M=1000 estimates in each case. The sojourn time hazard functions are from exponential distributions.

rounded off to 2 decimal places, with seemingly no way to obtain more accurate figures. This is an issue in practice if one wanted to use the standard errors for calculations. For example, standard errors below 0.005 are reported as 0. This can be observed in Table 4.1.

4.3 Related quantities of interest

The survival function of the holding time in a particular state can tell us whether patients in the active treatment arm are worse off due to having a lower probability of being event-free. However, we would need additional information related to each transition out of that state to better understand the full picture. [Asanjarani et al., 2021] makes use of parametric and non-parametric plots of the *cumulative intensity transition function*, which is analogous to the cumulative cause-specific hazard function. The CIF as per equation (2.34) would also work for such diagnostics, as well as provide a form of goodness-of-fit visual check when the plots of parametric and non-parametric functions are compared. An example of the CIFs being used for this purpose can be seen in Section 6.2.3.

This section suggests a few other additional quantities. While the first passage time (Section 4.3.2) and state occupancy probability (Section 4.3.3) are not new concepts, we have described expressions for these quantities in terms of mixture approach parameters. The expression for the conditional distribution of total sojourn times as per Section 4.3.1 forms the basis for the proposed hypothesis test as per Section 4.4.1. While all these quantities are similar to that seen in [Weiss and Zelen, 1965], the results as per Sections 4.3.1 and 4.3.2 have been obtained independently using a somewhat different approach. In a nutshell, instead of using the holding time in each state associated with $S_i(t)$ and the first passage time distribution as starting points, we have made specific assumptions and then worked directly with the knowledge that the total time spent in the SMP for given starting state and destination state(s) makes up a mixture random variable, with components of the mixture being a sum of sojourn times given transitions which make up the path, and component probabilities given by probabilities of specific paths. Note that, like all the parametric quantities in this thesis which can be estimated with the maximum likelihood estimator, it is possible to construct confidence intervals for these quantities using the delta method.

4.3.1 Distribution of total sojourn times given passage through a given state

One such quantity that can be considered alongside the survival function of holding time in particular states is the distribution of the total sojourn time after reaching a given state $i \in \mathcal{S}_+$ and reaching one of several other possible states $j_1, j_2, \ldots, j_q \in \mathcal{S}$ of interest, conditional on reaching one of the $j_1, j_2, \ldots, j_q$ through state i. Here, $i \neq j_\nu$ for any $\nu \in \{1, 2, \ldots q\}$. To formalise this, first define $\mathcal{J} \subseteq \mathcal{S} \setminus \{1\}$ such that, for some fixed $i \in \mathcal{S}_+$, $i \notin \mathcal{J}$. Let η denote the total sojourn time after reaching $i \in \mathcal{S}_+$ and then taking a path to some $j_q \in \mathcal{J}$, conditional on passing through state i to reach any state in $\mathcal{J}$. Define $\tau(r_{w_q}(i, j_q)) = (\tau_{v_1}, \tau_{v_2}, \ldots, \tau_{v_{n_{w_q,ij_q}}})^\top$ as the n_{w_q,ij_q}-vector of sojourn times associated with each path $r_{w_q}(i, j_q)$ that starts in i and ends in a given j_q. Here, $w_q \in \{1, 2, \ldots, N_{ij_q}\}$ for each $j_q \in \mathcal{J}$ and each subscript $v_1, v_2, \ldots$ denotes a member of $r_{w_q}(i, j_q)$ such that τ_{v_r} is shorthand for $\tau_{v_{r,1}, v_{r,2}}$ where $v_{r,s}$ is the s^{th} entry of v_r for $s \in \{1, 2\}$. Define $\mathbf{1}(n_{w_q,ij_q})$ as a vector of n_{w_q,ij_q} with all entries equal to 1. Then, η is a sum of random variables, given by

$$\eta = \sum_{q=1}^{|\mathcal{J}|} \sum_{w_q=1}^{N_{ij_q}} \eta(w_q) \tag{4.4}$$

with each component of the sum $\eta(w_q) = \sum_{q=1}^{|\mathcal{J}|} \sum_{w_q=1}^{N_{ij_q}} \mathbf{1}(n_{w_q,ij_q})^\top \tau(r_{w_q}(i, j_q))$ giving the total sojourn time after reaching state i and then reaching given $j_q \in \mathcal{J}$, conditional on reaching a state in $\mathcal{J}$ through state i. The distribution of η is a mixture, with each component having density and associated weight respectively given by

$$
\begin{aligned}
f_{\eta(w_q)}(u) =& f_{\mathbf{1}(n_{w_q,ij_q})^\top \tau(r_{w_q}(i,j_q))}(u) && (4.5)\\
\text{and } \; p(w_q) =& \left(\sum_{z=1}^{N_{1i}} \mathbf{P}_{(1,i)}^{n_{z,1j}}\right) \mathbf{P}_{(i,j_q)}^{n_{w_q,ij_q}} \Bigg/ \sum_{q=1}^{|\mathcal{J}|} \sum_{w_q=1}^{N_{ij_q}} \left(\sum_{z=1}^{N_{1i}} \mathbf{P}_{(1,i)}^{n_{z,1j}}\right) \mathbf{P}_{(i,j_q)}^{n_{w_q,ij_q}} && (4.6)
\end{aligned}
$$

for each $j_q \in \mathcal{J}$.

We have that $\eta(w_q)$ is a sum of independent random variables, with mixture (conditional) density function $f_{\eta(w_q)}(u)$. It is possible that $f_{\eta(w_q)}(u)$ is easy to obtain in closed form. Otherwise, $f_{\eta(w_q)} = f_{v_1} * \cdots * f_{v_{n_{w_q,ij_q}}}$ where $g_1 * g_2$ denotes the convolution of functions g_1 and g_2. These integrals could be computed numerically, otherwise one could carry out Laplace transforms to linearise the convolution and make it easier to compute the integrals – see, for example, [Weiss and Zelen, 1965]. Hence $\eta = \sum_{q=1}^{|\mathcal{J}|} \sum_{w_q=1}^{N_{ij_q}} \eta(w_q)$ with (conditional) mixture density function

$$
f_\eta(u) = \sum_{q=1}^{|\mathcal{J}|} \sum_{w_q=1}^{N_{ij_q}} p(w_q) f_{\eta(w_q)}(u). \tag{4.7}
$$

Each weight as per equation (4.6) is the conditional probability of observing that particular path. The numerator is associated with taking any of N_{1i} paths from state 1 to state i, and then taking a particular path from state i to j_q. Since state i must be reached from state 1, $\sum_{z=1}^{N_{1i}} \boldsymbol{P}_{(1,i)}^{n_{z,1j}}$ appears in both the numerator and denominator. After factoring these terms out and simplifying, we obtain

$$
p(w_q) = \mathbf{P}_{(i,j_q)}^{n_{w_q,ij_q}} \Bigg/ \sum_{q=1}^{|\mathcal{J}|} \sum_{w_q=1}^{N_{ij_q}} \mathbf{P}_{(i,j_q)}^{n_{w_q,ij_q}} \tag{4.8}
$$

for each $j_q \in \mathcal{J}$. This makes sense intuitively since, given state i is reached, the Markov property ensures that the probability associated with transitions after reaching state i do not depend on the past history. Hence, each of the conditional probabilities depends only on the sub-path taken from state i to j_q instead of the entire path taken from state 1 to state j_q.

Examples A special case of η would be for $i = 1$ and $j_q \in \mathcal{J} = \mathcal{S} \setminus \mathcal{S}_+$ *i.e.* $\mathcal{J}$ is the set of absorbing states. Then, η is the total sojourn time before reaching any absorbing state $j_q \in \mathcal{J}$, and the denominator of $p(w_q)$ in equation (4.8) is simply $\sum_{q=1}^{|\mathcal{J}|} \sum_{w_q=1}^{N_{ijq}} \mathbf{P}_{(1,j_q)}^{n_{w_q,1j_q}} = 1$ because reaching any absorbing state from state 1 is guaranteed. The numerator is simply $\mathbf{P}_{(1,j_q)}^{n_{w_q,1j_q}}$, which is the probability of observing a particular path from state 1 to j_q. A more general example can be seen by considering the 5-state model as per Figure 4.1. We have $\mathcal{S} = \{1,2,3,4,5\}$, $\mathcal{V} = \{(1,2),(1,3),(2,3),(2,4),(3,4),(3,5)\}$. If $i = 2$ and $\mathcal{J} = \{4,5\}$ then, given state 2 is reached, there are three distinct sub-paths that start in state 2 and end up at either state 4 or 5. These sub-paths are $r_1(2,4) = \{(2,4)\}$, $r_2(2,4) = \{(2,3),(3,4)\}$, and $r_1(2,5) = \{(2,3),(3,5)\}$. Hence, each component of η is given by

$$\begin{cases} \tau_{24} & \text{with probability } \ p_{24}/(p_{24} + p_{23}p_{34} + p_{23}p_{35}) \\ \tau_{23} + \tau_{34} & \text{with probability } \ p_{23}p_{34}/(p_{24} + p_{23}p_{34} + p_{23}p_{35}) \\ \tau_{23} + \tau_{35} & \text{with probability } \ p_{23}p_{35}/(p_{24} + p_{23}p_{34} + p_{23}p_{35}). \end{cases} \tag{4.9}$$

If $i = 2$ and $\mathcal{J} = \{3\}$ then $\eta = \tau_{23}$ with probability $p_{23}/p_{23} = 1$ since this is associated with the conditional event of passing through state 2 to reach state 3, of which there is only one way to do so.

One advantage of writing η in this way is that it is relatively straightforward to calculate quantities such as expectation and variance since

$$\mathrm{E}\left(\eta\right) = \sum_{q=1}^{|\mathcal{J}|} \sum_{w_q=1}^{N_{ijq}} p(w_q) E\left[\eta(w_q)\right] \tag{4.10}$$

and

$$\begin{aligned} \mathrm{Var}\left(\eta\right) =& \mathrm{E}\left(\eta^2\right) - \left[\mathrm{E}\left(\eta\right)\right]^2 \\ =& \sum_{q=1}^{|\mathcal{J}|} \sum_{w_q=1}^{N_{ijq}} p(w_q) \mathrm{E}\left[(\eta(w_q))^2\right] - \left[\sum_{q=1}^{|\mathcal{J}|} \sum_{w_q=1}^{N_{ijq}} p(w_q) \mathrm{E}\left[\eta\left(w_q\right)\right]\right]^2 . \end{aligned} \tag{4.11}$$

Equation (4.10) forms the basis for the hypothesis test proposed in Section 4.4.1.

If it is desired to extend the calculations to consider the more general case where we may have several states of interest $i_1, i_2, \cdots \in \mathcal{I}$ to pass through before reaching a state in $\mathcal{J}$, then one could do so if one carefully defines the conditioning event to include all the paths that pass through $i_u \in \mathcal{I}$ $(u = 1, 2, \cdots)$ and also possibly passing through one or more other states in $\mathcal{I}$ before reaching a state in $\mathcal{J}$.

4.3.2 Distribution of the first passage time

Since we have assumed that all non-absorbing states in $\mathcal{S}$ are transient, the *first passage time* from state 1 to state $k \in \mathcal{S} \setminus \{1\}$ can be defined as the time taken to reach state k given the SMP started at time 0 in state 1. Let U_k be the first passage time associated with state k. Using similar ideas seen Section 4.3.1, U_k can be written as a sum of random variables as follows:

$$U_k(z) = \mathbf{1}(n_{z,1k})^\top \tau(r_z(1,k)) \text{ with probability } p(z) = \mathbf{P}^{n_{z,1k}}_{(1,k)}. \qquad (4.12)$$

We note that, because some states $k \in \mathcal{S}$ might never be reached, the distribution of U_k is improper in general *i.e.* $0 < \lim_{u\to\infty} F_{U_k}(u) \leq 1$.

Similarly to how the density function of η is obtained, we can write the subdensity function as

$$f_{U_k}(u) = \sum_{z=1}^{N_{1k}} p(z) f_{U_k(z)}(u). \qquad (4.13)$$

The distribution of the first passage time may be useful information in the context of patient benefit, but may not be of deep interest on its own. However, the expression is used for our derivation of the expression for the state occupancy probability as per Section 4.3.3.

4.3.3 State occupancy probabilities

The *state occupancy probability* associated with state k at time u is defined as the probability of being in state k at time u of the SMP. Using our notation for the SMP as per Section 2.4.1, the state occupancy probability can be denoted $P_k(u) = P(X_u = k|X_0 = 1)$ for $k \in \mathcal{S}$. Since our setup assumes that each state can be visited at most once and that all non-absorbing states are transient, we can write the state occupancy probabilities in terms of the subdensity function of U_k as per equation (4.13) and survival functions of holding times:

$$P_k(u) = \mathbb{1}_{\{1\}}(k)S_1(u) + \big(1 - \mathbb{1}_{\{1\}}(k)\big) \int\limits_0^u f_{U_k}(v)\Big(S_k(u-v)\Big)^{\mathbb{1}_{\mathcal{S}_+\setminus\{1\}}(k)} \mathrm{d}v. \quad (4.14)$$

Here, $\mathbb{1}_A(x)$ is the indicator function taking value 1 if $x \in A$, 0 otherwise.

If $k = 1$, the probability of still being in state 1 given that the SMP started in state 1 is precisely the survival function of the holding time in state 1, which tends to 0 as $t \to \infty$. This is the first term in equation (4.14). Otherwise, $P_k(t)$ involves only the second term and its contribution depends on whether $k \in \mathcal{S} \setminus \mathcal{S}_+$ or $k \in \mathcal{S}_+ \setminus \{1\}$. If $k \in \mathcal{S} \setminus \mathcal{S}_+$, then the state-$k$ occupancy probability is simply $F_{U_k}(t)$. If $k \in \mathcal{S}_+ \setminus \{1\}$ then it means that state k is visited for the first time at time v, followed by a stay in state k for $u - v$ time units before leaving state k. If f_{U_k} is known, then $f_{U_k} * S_k$ is relatively straightforward to compute or estimate.

Example Consider an illness-death model as described in previous examples, with $\mathcal{S} = \{1, 2, 3\}$, $\mathcal{S}_+ = \{1, 2\}$, and $\mathcal{V} = \{(1, 2), (1, 3), (2, 3)\}$. Let the sojourn time hazard function for given $(i, j) \in \mathcal{V}$ be exponential-distributed *i.e.* $h_{ij}(t) = a_{ij}$ for all $(i, j) \in \mathcal{V}$. Let $p_{12} = 0.36, a_{12} = 0.2, a_{13} = 0.3, a_{23} = 0.1$. Figure 4.2 below shows the model and its parameters.

We are able to use equation (4.14) to find

$$\begin{aligned} P_1(t) =& S_1(t) = p_{12}\exp\left(-a_{12}t\right) + (1-p_{12})\exp\left(-a_{13}t\right), && (4.15) \\ P_2(t) =& \frac{p_{12}a_{12}}{a_{12}-a_{23}}\big(\exp\left(-a_{23}t\right) - \exp\left(-a_{12}t\right)\big), && (4.16) \\ \text{and } P_3(t) =& (1-p_{12})\big(1-\exp\left(-a_{13}t\right)\big) + && (4.17) \\ & \frac{p_{12}}{a_{12}-a_{23}}\left(-a_{12}\big(1-\exp(-a_{23}t)\big) - a_{23}\big(1-\exp(-a_{12}t)\big)\right). \end{aligned}$$

We can visualise the state occupancy probabilities for each state by plotting them against time as per Figure 4.3. Figure 4.3a shows the state occupancy probabilities for states 1 (green), 2 (blue) and 3 (red) respectively. We must have $P_1(t) + P_2(t) + P_3(t) = 1$ for any $t \geq 0$. Figure 4.3b shows a different way to visualise this information. The boundary between the green and blue areas is $P_1(t) = S_1(t)$ and the boundary between the blue and red areas is $P_1(t) + P_2(t)$. Hence, the size of each coloured area when considering any vertical strip on the figure represents the relative chance that individuals are likely to be in that state. For example, when considering the interval $(9.8, 10.2)$ (with interval endpoints given by the vertical black lines), the majority of individuals are going to be in state 3. In fact, the total proportion of the red area in the figure is over 50% when considering the interval $(0, 40)$. This tells us that relatively little time is spent in transient states 1 and 2 as compared to absorbing state 3.

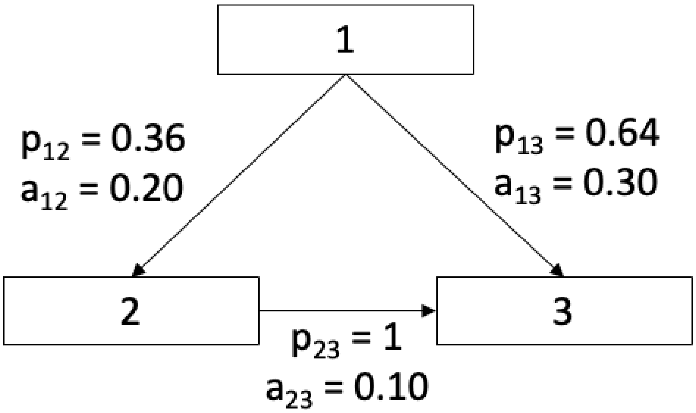
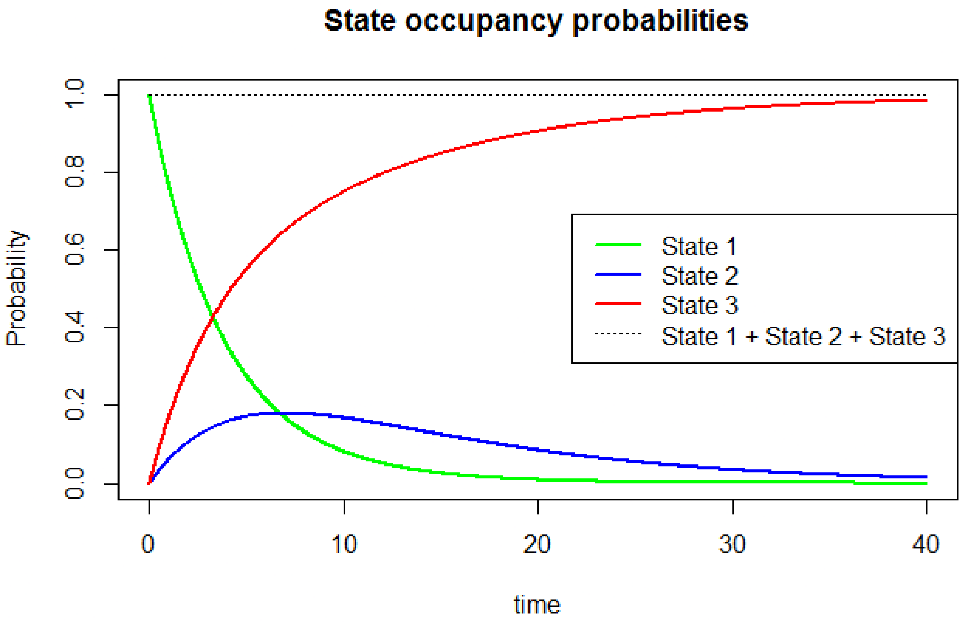
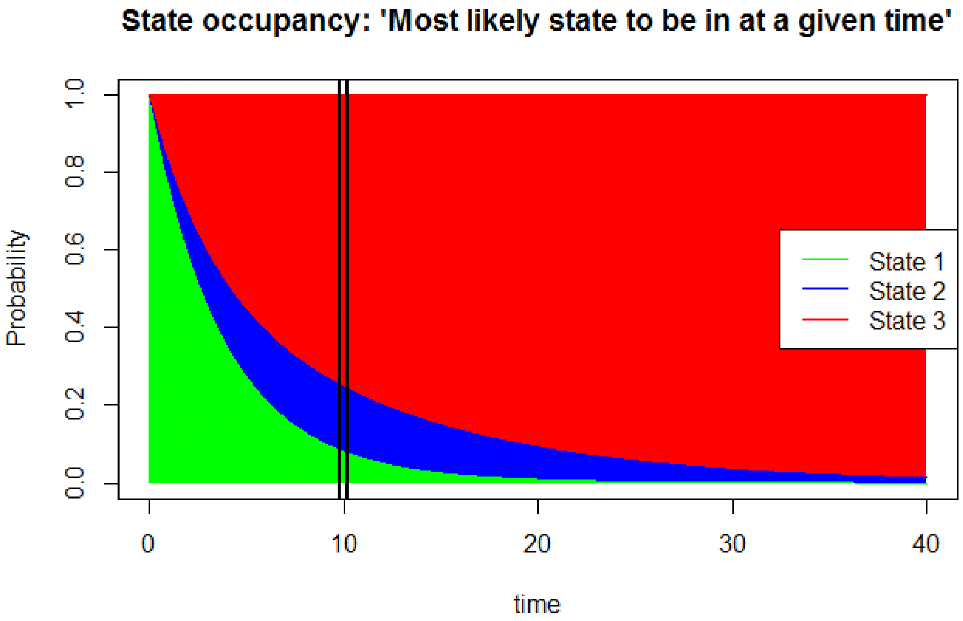


Figure 4.2: Illness-death model used to derive state occupancy probabilities.



(a) The figure shows state occupancy probabilities for states 1 (green), 2 (blue) and 3 (red) respectively.



(b) The boundary between the green and blue areas is $P_1(t) = S_1(t)$ and the boundary between the blue and red areas is $P_1(t) + P_2(t)$. Hence, the size of each coloured area when considering any vertical strip on the figure represents the relative chance that individuals are likely to be in that state. When considering the interval $(9.8, 10.2)$ (with interval endpoints given by the vertical black lines), the majority of individuals are going to be in state 3.

Figure 4.3: Visualising state occupancy probabilities for the illness-death model

4.4 Proposed hypothesis tests

We propose formal parametric hypothesis tests to quantify and evaluate potential patient benefit between two groups of interest *e.g.* a group on active treatment versus a control group. The first test, denoted **Test A**, is based on differences of the conditional expected total sojourn times given passage through a particular state before reaching other states of interest (as per Section 4.3.1), while the second test, denoted **Test B**, is based on the differences between the survival functions of the holding time between treatment arms, for specific time intervals of interest.

The former test might be useful as a more general "global" test checking for overall benefit, while the latter test could be used to deep-dive into specific states of interest. The details are in Section 4.4.1 and 4.4.2 respectively. The underlying distributions of the test statistics involve using the asymptotic Gaussian distribution of the maximum likelihood estimator and the delta method as described in Section 2.2.2.

For simplicity, we consider a single covariate $Z \in \{0, 1\}$ denoting the treatment arm (where $Z = 1$ denotes active treatment), but these ideas can potentially be generalised and extended to the situation with a general covariate vector. For example, we could have Cox-type proportional hazards as per Section 2.6, or some other appropriate setup. See Section 7.2 for further discussion.

4.4.1 Test based on average total sojourn times

First, we propose a test that seeks to establish whether there are significant differences in the (conditional) average total sojourn time between treatment groups. This test is based on the expectation defined in equation (4.10) associated with the distribution of total sojourn times after a particular state i is reached, before ending up in one of the states $j_1, j_2, \ldots, j_q$ of interest after passing through state i. We call this test "**Test A**".

The null hypothesis would be that the expectations of the conditional random variable η (as defined in Section 4.3.1) between treatment groups are identical, and therefore there is no (average) difference in benefit to the patient. Note that since every state in the multi-state model is defined to be detrimental to patients, significant differences in the expected values of $\eta|Z = z$ ($z \in \{0, 1\}$) imply that the group with the higher average sojourn time is better off since they are (on average) resisting transitions to other undesirable states for longer compared to the other group.

To formalise this, define $\mu_z = \mathrm{E}(\eta|Z = z)$ for each of $z = 0, 1$. Then,

$$g_A(\boldsymbol{\theta}) = \mu_1 - \mu_0 \tag{4.18}$$

is the function of $\boldsymbol{\theta}$ required to construct our test statistic $T_A(\hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator of the model parameters. The expression for $T_A(\hat{\boldsymbol{\theta}})$ is shown in equation (4.19) below.

Suppose

$$\mathrm{H}_0 : \mu_1 - \mu_0 = 0 \text{ vs } \mathrm{H}_1 : \mu_1 - \mu_0 > 0$$

to test for the presence of benefit to the patient. It would also be possible to specify a two-tailed alternative hypothesis, if desired. We make use of our knowledge of the asymptotic distribution associated with the maximum likelihood estimator, as well as the delta method, described in Section 2.2.2 to ascertain the asymptotic distribution of function $g_A(\hat{\boldsymbol{\theta}})$ under H_0 and therefore construct the test statistic, $T_A(\hat{\boldsymbol{\theta}})$. Using the delta method, the test statistic is

$$T_A(\hat{\boldsymbol{\theta}}) = \frac{g_A(\hat{\boldsymbol{\theta}})}{V_A} \tag{4.19}$$

under H_0, where $V_A = \sqrt{\{\nabla g_A(\hat{\boldsymbol{\theta}})\}^\top \mathbf{I}^{-1}(\hat{\boldsymbol{\theta}})\{\nabla g_A(\hat{\boldsymbol{\theta}})\}/m}$. The test statistic has an approximate standard Gaussian distribution under H_0.

Large positive values of the test statistic lead us to reject H_0, and suggest that patients in active treatment are benefiting on average, while large negative values suggest these patients are suffering more than the control arm on average.

In practice, it might make the most sense to define i as the starting state and $\{j_1, j_2, \ldots, j_q\} = \mathcal{S} \setminus \mathcal{S}_+$ *i.e.* the set of all absorbing states. In other words, the test is for differences in average total sojourn time before absorption. If there are specific reasons to choose other states for i and $j_1, j_2, \ldots, j_q$, then it would be possible to do so. This test result alone, however, may not be sufficiently informative as it merely gives a general overview without exploring what might be causing differences in patient benefit (if any). For this reason, in the next Section 4.4.2 we propose a test that considers the survival function of the holding time in specific states of interest.

4.4.2 Test based on differences in survival function of holding time

As mentioned in the previous section, it may be necessary to ascertain which states are contributing to benefit (or disbenefit) for patients. Suppose, then, it is of interest to test for a difference in holding time distributions between treatment arms in state $i \in \mathcal{S}_+$. We call this "**Test B**".

The null hypothesis would be that the survival functions of the holding time in state i between both arms are "identical", and therefore there is no difference in patient benefit. We propose a general function of the form

$$g_B(\boldsymbol{\theta}) = \int_E \Big(S_i(\cdot|Z=1) - S_i(\cdot|Z=0)\Big)\mathrm{d}\mu \tag{4.20}$$

to derive the test statistic. Here μ is an appropriately chosen measure and E is a subinterval of $\mathbb{R}_{\geq 0}$. Hence, the null and alternative hypotheses are

$$\begin{aligned}
\mathrm{H}_0 &: \int_E \Big(S_i(\cdot|Z=1) - S_i(\cdot|Z=0)\Big)\mathrm{d}\mu = 0 \;\text{ vs} \\
\mathrm{H}_1 &: \int_E \Big(S_i(\cdot|Z=1) - S_i(\cdot|Z=0)\Big)\mathrm{d}\mu > 0.
\end{aligned}$$

Once again, the alternative hypothesis could be two-sided, if desired.

For example, if we wish to compare the survival functions of the holding times at a specific point $t_0 > 0$ then we can choose μ to be the Dirac measure taking value 1 at $t_0 \in E = \mathbb{R}_{\geq 0}$. If, more generally, we wish to consider some notion of "average" benefit in the time interval (a, b) then we can take μ to be the Lebesgue measure on (a, b) with $E = (a, b)$. Note that in the latter scenario, setting $a = 0$ and $b = \infty$ gives equation (4.20) as the difference of the expected holding times in state i between the active and control treatment groups.

The test statistic is obtained analogously to that in equation (4.19):

$$T_B(\hat{\boldsymbol{\theta}}) = \frac{g_B(\hat{\boldsymbol{\theta}})}{V_B} \tag{4.21}$$

where $V_B = \sqrt{\{\nabla g_B(\hat{\boldsymbol{\theta}})\}^{\top}\mathbf{I}^{-1}(\hat{\boldsymbol{\theta}})\{\nabla g_B(\hat{\boldsymbol{\theta}})\}/m}$. Once again, $T_B(\hat{\boldsymbol{\theta}})$ has an approximate standard Gaussian distribution under H_0.

The intuition behind this test statistic is that there should be no difference in the time spent in state i (on "average") if both treatments are equally beneficial, and so we reject the null hypothesis if the test statistic is significantly different from zero. If the test statistic is significantly large in magnitude and positive (resp. negative), then it suggests that patients in active treatment are benefiting (resp. disbenefiting) in state i.

It might also be possible to transform the function g_B for computational savings, especially when using the intensity transition functions approach. For instance, performing the test at time t_0 using the Dirac measure and substituting equation (2.33) into equation (4.20) yields

$$g_B(\boldsymbol{\theta}) = \exp\left(-\sum_{i \neq k} \tilde{H}_{ik}(t_0|Z=1)\right) - \exp\left(-\sum_{i \neq k} \tilde{H}_{ik}(t_0|Z=0)\right) \tag{4.22}$$

where $\tilde{H}_{ij}$ is the cumulative intensity transition function associated with transition

$i \to j$. From this, we can deduce that H_0 is true if and only if

$$\sum_{i \neq k} \tilde{H}_{ik}(t_0 | Z = 1) = \sum_{i \neq k} \tilde{H}_{ik}(t_0 | Z = 0).$$

We can then use $g_B(\boldsymbol{\theta}) = \sum_{i \neq k} \tilde{H}_{ik}(t_0 | Z = 1) - \sum_{i \neq k} \tilde{H}_{ik}(t_0 | Z = 0)$ in the test statistic for **Test B** (equation (4.21)) instead of equation (4.22).

It is worth pointing out that **Test B** differs from the class of quadratic empirical distribution function (EDF) test statistics (see, for example, [Stephens, 1974]) such as the Anderson-Darling test ([Anderson and Darling, 1952]) in that the integrand is not quadratic nor restricted to being non-parametric. Furthermore, the proposed test is not a goodness-of-fit test since we are not so concerned about whether the holding time distributions are actually different, but whether there is an "overall" difference in the time taken to leave the current state i in a given time interval. We could very well fail to reject the null hypothesis when, in fact, the distributions of the holding times between each treatment arm are very different (but the integral of the difference in survival functions is approximately zero in a given interval (a, b)).

In practice, it might make the most sense to perform the above test with the Lebesgue measure on some time interval (a, b) and with $E = (a, b)$, and to perform this test in several states of interest. We can choose a and b sensibly based on *e.g.* previously established pharmacological evidence or other criteria. Note, however, that rejecting the null hypothesis does not give us much information about what is causing the potential difference in benefit between the treatment arms. To determine which of the transitions out of state i might be causing the benefit (or disbenefit), one would need to further investigate all the transitions out of state i. Furthermore, it might be helpful to consider other quantities, such as state occupancy probabilities, alongside the test results to give a broader picture of patient benefit. The quantities discussed in Section 4.3 might be most useful for this, and examples of such applications can be found in Section 5.6.

Chapter 5

Results: Simulation study assessing proposed tests

The goal of this section is to demonstrate some of the properties of the proposed hypothesis tests as per Section 4.4, by using simulated data.

Three different model setups are considered, with the empirical Type I error rates and statistical power assessed in each case. All the true models involve Weibull sojourn time hazard functions, but misspecified models (exponential and gamma distributions) are also fitted to the data to see the effect of misspecification. This is of particular interest since some of Weibull shape parameters are chosen to have values relatively close to unity, making the sojourn time distributions "almost" exponential-distributed.

Section 5.1 describes the common elements of each model and the data simulated from them, with exact details in the next three sections. Specifically, Section 5.2 discusses the baseline setup, with Section 5.3 repeating the analysis after adding significant amounts of right-censoring. Section 5.4 repeats the analysis after reducing the amount of detectable benefit when there is also right-censoring. Results related to some of the quantities of interest as per Section 4.3 are shown in Section 5.6. A summary of results associated with various hypothesis tests can be found in figures and tables, especially in each of Section 5.2.2, Section 5.3.2, and Section 5.4.2. The estimated power functions shown in these sections are for tests of size $\alpha = 0.05$, with power functions for other values of α found in Section B.2, Section B.3, and Section B.4.

The summary tables of results in each of these three sections are also presented again in Section 5.5, for convenience. Finally, a short discussion of the results is given in Section 5.7.

5.1 General setup and description of tests

The purpose of this section is to facilitate easy navigation of this chapter, by describing the common characteristics of all the models considered in the simulation study setups. Details of the hypothesis tests and the results are also given.

5.1.1 General setup

The graph structure of the model chosen for the simulation study is a 5-state model similar to the one depicted in Figure 4.1, reproduced in Figure 5.1 below:

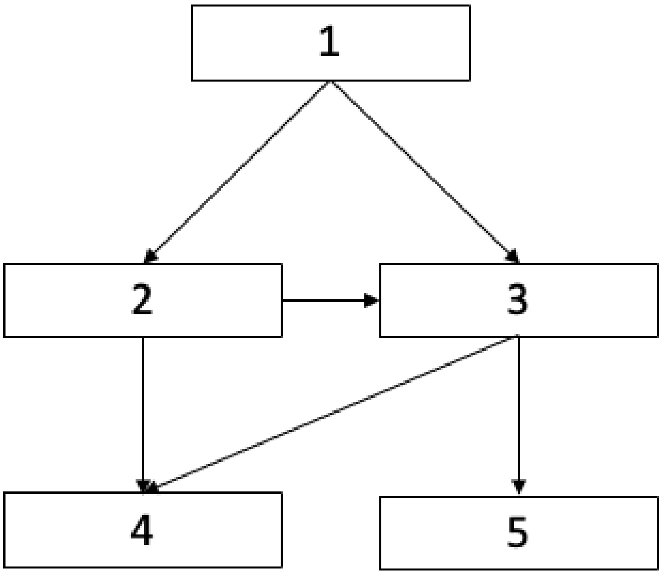


Figure 5.1: Underlying graph structure of 5 state model used for simulation study.

The underlying graph structure is kept consistent for all three setups under consideration. Section 5.2.1 describes the baseline setup, with Section 5.3.1 describing the baseline setup with significant amounts of right-censoring added. Section 5.4.1 describes the setup where there is a reduced amount of detectable benefit while also having right-censoring.

There is a single covariate, Z, taking values 1 and 0 for active treatment and control treatment respectively. In each case, there is an equal proportion of patients in each treatment arm. Active treatment is associated with benefit to patients. For each setup, the parameters are chosen such that there is negligible difference between treatment arms in states 1 and 2, with the bulk of significant benefit being associated with state 3 for patients in active treatment. The exact model parameters and other characteristics of each model setup are detailed in each of the aforementioned sections.

Furthermore, the survival functions of the holding time in state 3 for each treatment arm intersect at a specific interior time point (t_0, say) in every scenario, with $S_3(t|Z = 1) < S_3(t|Z = 0)$ for all $t < t_0$ and $S_3(t|Z = 1) > S_3(t|Z = 0)$ for all $t > t_0$. This becomes relevant when considering **Test B**. More details can be found in Section 5.1.3.

As mentioned in Section 4.2, any model with this graph structure is somewhat realistic in the sense that a real clinical trial could have one or two absorbing states (*e.g.* "death" and "lost to follow up") as well as one or two non-starting transient states (*e.g.* "disease progression" and "premature discontinuation of treatment").

Two different sample sizes are chosen for each setup: $m = 10000$ to ascertain large sample properties, and $m = 1000$ to ascertain real-world performance. A sample size of 1000 is realistic and, in fact, not that large in the context of clinical oncology – see [Miller et al., 2020], [Fehrenbacher et al., 2020], and [Mamounas et al., 2019] for examples, of which the latter two have $m > 3000$.

Finally, $M = 1000$ datasets are simulated for each scenario and sample size. For a given significance level α, probability of type I error as well as power of the proposed tests in the different scenarios are chosen as performance measures.

Hence, similarly to Section 3.2.1, the MCSE for each model and scenario can be estimated by

$$\text{MCSE} = \sqrt{\frac{\hat{p}(1-\hat{p})}{M}} \tag{5.1}$$

where $\hat{p}$ is either the estimated probability of type I error or estimated power of the hypothesis test, as appropriate. For $M = 1000$, the MCSE is just under 1.6%. As reliable results were desired without overly costly computation, this MCSE value was deemed to be sufficiently low.

5.1.2 Model estimation

Each model is estimated via the method of maximum likelihood, as described in Section 2.4.5. The mixture approach is used to specify each model.

For each transition $i \to j$, all true models involve Weibull sojourn time hazard functions of the form

$$h_{ij}^{weib}(t) = a_{ij} b_{ij} (a_{ij} t)^{b_{ij}-1},$$

where a_{ij} is the rate parameter and b_{ij} is the shape parameter.

To assess the effects of model misspecification, exponential and gamma models are also fitted to the data. For each transition $i \to j$, the exponential sojourn time hazard function is

$$h_{ij}^{exp} = a_{ij}$$

where a_{ij} is the rate parameter. The gamma sojourn time hazard function is

$$h_{ij}^{gamma} = \frac{a_{ij}^{d_{ij}} t^{d_{ij}-1} \exp\left(-a_{ij} t\right)}{\mathit{\Gamma}(d_{ij}) - \gamma(d_{ij}, a_{ij} t)},$$

where a_{ij} is the rate parameter, d_{ij} is the shape parameter, $\mathit{\Gamma}(s) =$

$\int_0^\infty u^{s-1} \exp(-u) \, \mathrm{d}u$ is the gamma function, and $\gamma(s, x) = \int_0^x u^{s-1} \exp(-u) \, \mathrm{d}u$ is the lower incomplete gamma function.

As before, all models parameters are estimated using `Rsolnp` ([Ghalanos and Theussl, 2015]) in R, with reported numerical Hessian matrices used for the estimation of all observed Fisher information matrices.

5.1.3 Description of tests and how results are obtained

Test A: Testing for differences in average total sojourn time before absorption

This test is based on that described in Section 4.4.1. The quantity of interest as per equation (4.18), and we define η as the average total sojourn time before absorption, given that the starting state is state 1. Thus, we can compare the average total sojourn time in the different states before absorption for each treatment arm. This gives a general measure of whether there is potential patient benefit associated with the treatment of interest ($Z = 1$). Let us then consider,

$$\mathrm{H}_0 : \mathrm{E}(\eta|Z = 1) = \mathrm{E}(\eta|Z = 0) \text{ vs } \mathrm{H}_1 : \mathrm{E}(\eta|Z = 1) > \mathrm{E}(\eta|Z = 0).$$

The test statistic is as defined in equation (4.19), with approximate standard Gaussian distribution under H_0.

To check the Type I error, the procedure is as follows. First, for a given study, we estimate $\mathrm{E}(\eta|Z = 0)$ using the estimated parameters in the usual way. However, $\mathrm{E}(\eta|Z = 1)$ is computed with parameters estimated using data from another simulation replication which is associated with $Z = 0$. In this way, H_0 is true since the test statistic is associated with $g_A(\boldsymbol{\theta}) = 0$ and so $g_A(\hat{\boldsymbol{\theta}}) \approx 0$ as a result. Since H_0 is true, we evaluate the proportion of times in the $M = 1000$ simulations where we erroneously reject H_0 *i.e.* whenever the test statistic $T_A(\hat{\boldsymbol{\theta}}) = \frac{g_A(\hat{\boldsymbol{\theta}})}{V_A} > q_{1-\alpha}$, where $q_{1-\alpha}$ denotes the $(1 - \alpha)$ quantile of the standard Gaussian distribution. We let ε denote the estimated value of the Type I error, α, for common values 0.01, 0.05 and 0.10. The left column of the respective tables in each of Section 5.2.2, Section 5.3.2, and Section 5.4.2 show the estimated values of ε for each α and sample size (either $m = 10000$ or $m = 1000$). This is done for all three model fits.

To check the power of the test, we repeat the procedure as per the previous paragraph except we correctly estimate $g_A(\boldsymbol{\theta}) = \mathrm{E}(\eta|Z = 1) - \mathrm{E}(\eta|Z = 0)$, which should be significantly more than zero. We then check the proportion of times we (correctly) reject H_0. The proportion of studies in which we correctly reject H_0 is denoted ρ. The right column of the respective tables in each of Section 5.2.2, Section 5.3.2, and Section 5.4.2 show the estimated values of ρ for each α and sample size (either $m = 10000$ or $m = 1000$). This is done for all three model fits.

Test B: Testing for differences between treatment arms with two variations of the test comparing holding times in state 3

The following two tests are variations of that described in Section 4.4.2. **Test B1** compares the survival functions of the holding time associated with each treatment arm at a specific time point, while **Test B2** compares the survival functions in a given time interval instead.

Test B1

As mentioned at the beginning of Section 5.1, the bulk of significant benefit is associated with state 3 for patients in active treatment ($Z = 1$). Furthermore, $S_3(t|Z = 1) < S_3(t|Z = 0)$ for all $t < t_0$ and $S_3(t|Z = 1) > S_3(t|Z = 0)$ for all $t > t_0$ where t_0 is the (only) interior intersection point of both survival functions. Suppose then we use a Dirac measure to ascertain the difference in survival functions of the holding time in state 3 between treatment arms, with reference to t_0. Then, consider the null and alternative hypotheses

$$\mathrm{H}_0 : S_3(t_0|Z = 1) = S_3(t_0|Z = 0) \;\text{ vs }\; \mathrm{H}_1 : S_3(t_0|Z = 1) > S_3(t_0|Z = 0).$$

Hypothesis H_0 is clearly true under these conditions. The test statistic is defined as per equation (4.21), with the numerator as $g_{B1}(\hat{\boldsymbol{\theta}}) = \Big(\hat{S}_3(t_0|Z=1) - \hat{S}_3(t_0|Z=0)\Big)$. We use equation (2.30) for $S_3(\cdot)$, and the hat denotes that the survival function is estimated with the maximum likelihood estimate.

Since H_0 is true, we evaluate the proportion of times in the $M = 1000$ simulations where we erroneously reject H_0 *i.e.* whenever $T_{B1}(\hat{\boldsymbol{\theta}}) = \frac{g_{B1}(\hat{\boldsymbol{\theta}})}{V_{B1}} > q_{1-\alpha}$. We let ε denote the estimated value of the Type I error, α, for common values 0.01, 0.05 and 0.10. The left column of the respective tables in each of Section 5.2.2, Section 5.3.2, and Section 5.4.2 show the estimated values of ε for each α and sample size (either $m = 10000$ or $m = 1000$). This is done for all three model fits.

We then consider the power of the test *i.e.* the probability of rejecting H_0 when H_0 is not true. We investigate by calculating the proportion of times we reject H_0 when we perform the test at t_0, and then increase time in increments of 0.01 to see how quickly the power function increases. Each of Section 5.2.2, Section 5.3.2, and Section 5.4.2 show figures which plot the estimated power as a function of time. This is done for all three model fits. Additionally, the right column of the relevant tables in each of these sections show the values of t which give us at least 80% power, denoted $t_{80\%}$. If the test is correctly rejecting H_0, then the test can be considered relatively powerful if the quantity $t_{80\%} - t_0$ (which must be positive) is relatively small.

Test B2

Consider now the test with null and alternative hypotheses

$$\begin{aligned} \mathrm{H}_0 &: \int_{0.2}^{t_1} S_3(u|Z=1)\mathrm{d}u = \int_{0.2}^{t_1} S_3(u|Z=0)\mathrm{d}u \\ \text{vs }\ \mathrm{H}_1 &: \int_{0.2}^{t_1} S_3(u|Z=1)\mathrm{d}u > \int_{0.2}^{t_1} S_3(u|Z=0)\mathrm{d}u \end{aligned}$$

where the lower limit of each integral, 0.2, is chosen arbitrarily. This test is concerned with detecting benefit in the interval $(0.2, t_1)$. For fixed t_1, the test statistic and its distribution under H_0 is obtained similarly to that described for **Test B1**. This time, the numerator of the test statistic is $g_{B2}(\hat{\boldsymbol{\theta}}) = \int_{0.2}^{t_1} \Big(\hat{S}_3(u|Z=1) - \hat{S}_3(u|Z=0)\Big)\,\mathrm{d}u$. Due to the fact that both survival functions of holding time in state 3 intersect only once (at t_0) such that $S_3(t|Z=1) < S_3(t|Z=0)$ for all $t < t_0$ and $S_3(t|Z=1) > S_3(t|Z=0)$ for all $t > t_0$, we find that $g_{B2}(\boldsymbol{\theta}) = \int_{0.2}^{t_1} \Big(S_3(u|Z=1) - S_3(u|Z=0)\Big)\mathrm{d}u < 0$ for $0.2 < t_1 < t^*$, where t^* is the time point such that $g_{B2}(\boldsymbol{\theta}) = 0$. This signifies negative "overall benefit" in $0.2 < t_1 < t^*$, and we start to have positive benefit once we allow $t_1 > t^*$.

We thus use this as a starting point for checking the estimated Type I error, similarly to the method used for **Test B1**, by carrying out the test at $t_1 = t^*$ (where H_0 is true). The left column of the respective tables in each of Section 5.2.2, Section 5.3.2, and Section 5.4.2 show the estimated values of ε for each α and sample size (either $m = 10000$ or $m = 1000$). This is done for all three model fits.

To estimate the power of this test, we now vary t_1 from 0.21 to 6 in increments of 0.01 and, for each t_1, calculate the proportion of times H_0 is rejected when it should be rejected. We pay particular attention to $t_1 > t^*$, where there is positive "overall benefit". Each of Section 5.2.2, Section 5.3.2, and Section 5.4.2 show figures which plot the estimated power as a function of time. This is done for all three model fits. Additionally, the right column of relevant tables in each of these sections show the values of t which give us at least 80% power, denoted $t_{80\%}$. If the test is correctly rejecting H_0, then the test can be considered relatively powerful if the quantity $t_{80\%} - t^*$ (which must be positive) is relatively small.

5.2 Baseline model setup

Section 5.2.1 describes the baseline model in detail, while Section 5.2.2 discusses tables and figures summarising the findings associated with Type I errors and statistical power.

5.2.1 Description of baseline model

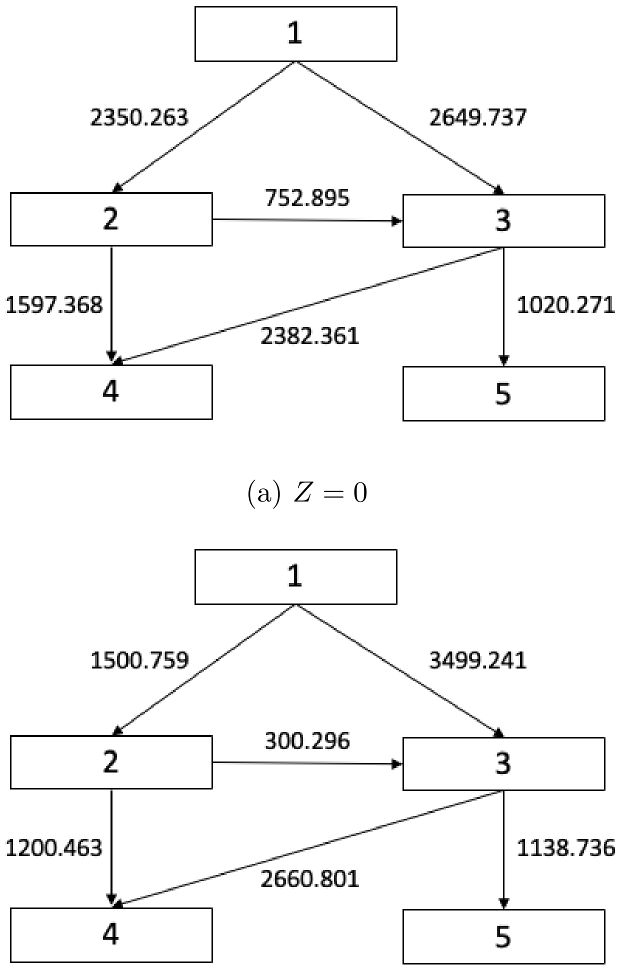
There are $M = 1000$ datasets simulated from a 5 state model based on the mixture approach. We have $\mathcal{S} = \{1, 2, 3, 4, 5\}$, $\mathcal{S}_+ = \{1, 2, 3\}$ and $\mathcal{V} = \{(1, 2), (1, 3), (2, 3), (2, 4), (3, 4), (3, 5)\}$. Figure 5.2 depicts the multi-state model. There are 5000 patients in each treatment arm, and the numbers alongside the arrows depict the average number of observed transitions.

For each possible transition (i, j) in $\mathcal{V}$, we have Weibull sojourn time hazard functions (as described in Section 5.1.2). We choose sample size $m = 10000$ and $m = 1000$ for each dataset, the former to assess large sample properties and the latter to assess real-world performance. There are an equal proportion of individuals in each of active and control treatment ($Z = 1$ and $Z = 0$ respectively), with parameters chosen differently based on treatment arm (more details are below). The parameter vector has length 30 with values as per Table 5.1 below.

$Z = 0$	p_{12} 0.47	a_{12} 8	a_{13} 20	b_{12} 0.791	b_{13} 0.899
	p_{23} 0.32	a_{23} 4.2	a_{24} 17.5	b_{23} 0.903	b_{24} 0.939
	p_{34} 0.7	a_{34} 5.1	a_{35} 2.1	b_{34} 0.957	b_{35} 0.903
$Z = 1$	p_{12} 0.3	a_{12} 5	a_{13} 41	b_{12} 0.709	b_{13} 0.905
	p_{23} 0.2	a_{23} 2.3	a_{24} 35	b_{23} 0.784	b_{24} 0.887
	p_{34} 0.7	a_{34} 13.9	a_{35} 1.1	b_{34} 0.256	b_{35} 0.475

Table 5.1: (Baseline model) Chosen parameter values for the simulated data with Weibull sojourn time hazard functions.

In almost every state, the shape parameters are specifically chosen to be close to unity, so that event times associated with each transition are “close” to exponential. The exception is in state 3 for the active treatment group ($Z = 1$), as this state is a main focus for analysis. The shape parameters in state 3 associated with the active treatment arm are significantly below unity, which leads to significant differences between the survival functions of the holding time in both treatment arms (see



(a) $Z = 0$

(b) $Z = 1$

Figure 5.2: (Baseline) 5-state model with six possible transitions depicting the baseline model. There are 5000 patients in each treatment arm, and the numbers alongside the arrows depict the average number of observed transitions.

Figure 5.3 for an illustration). The results will show that there is a large amount of patient benefit to be found by patients in active treatment ($Z = 1$) who reach state 3. The case where the state 3 shape parameters are closer to unity (so that there is relatively less benefit for patients in active treatment) is discussed in Section 5.4.

Note that even though the transition probability for 3 → 4 is identical for each treatment arm, the different hazard rates for 3 → 4 and 3 → 5 for each treatment arm lead to different survival functions of holding time in state 3. The survival functions of holding times in states 1 and 2 do not have significant differences between treatment arms and will not be discussed. The respective figures of the

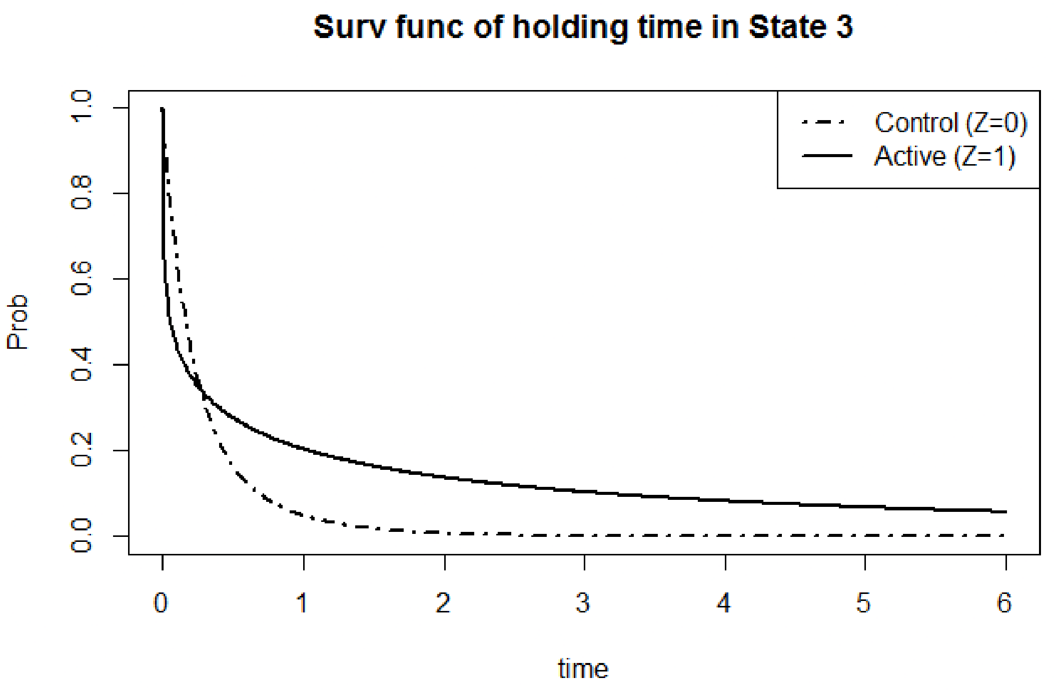
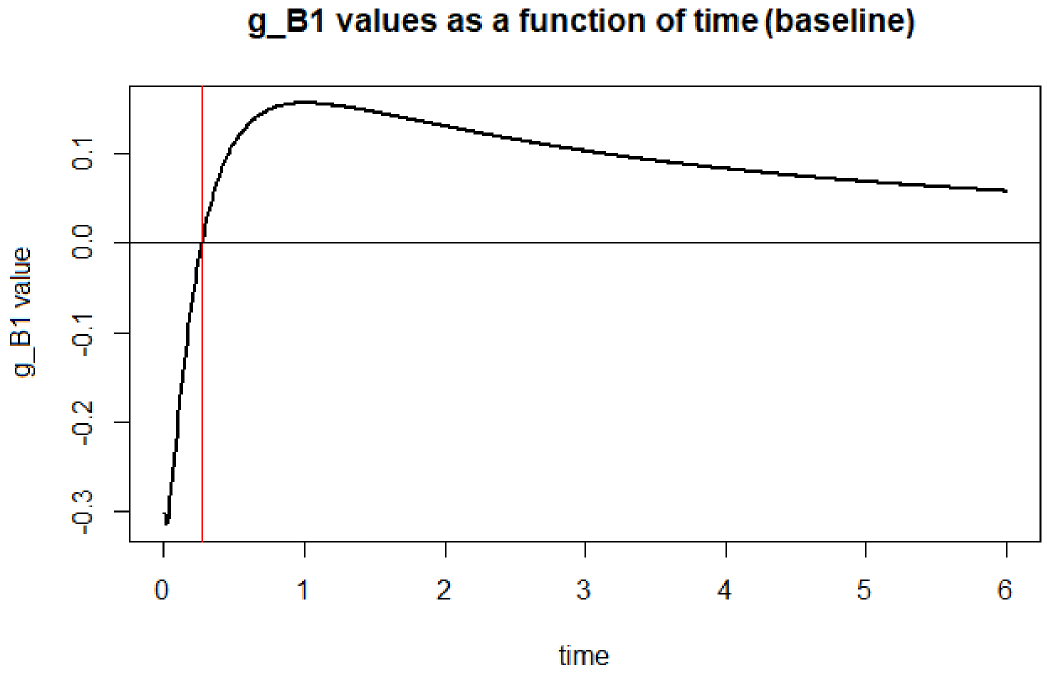


Figure 5.3: True survival functions of holding times for each treatment arm in state 3 for baseline model.

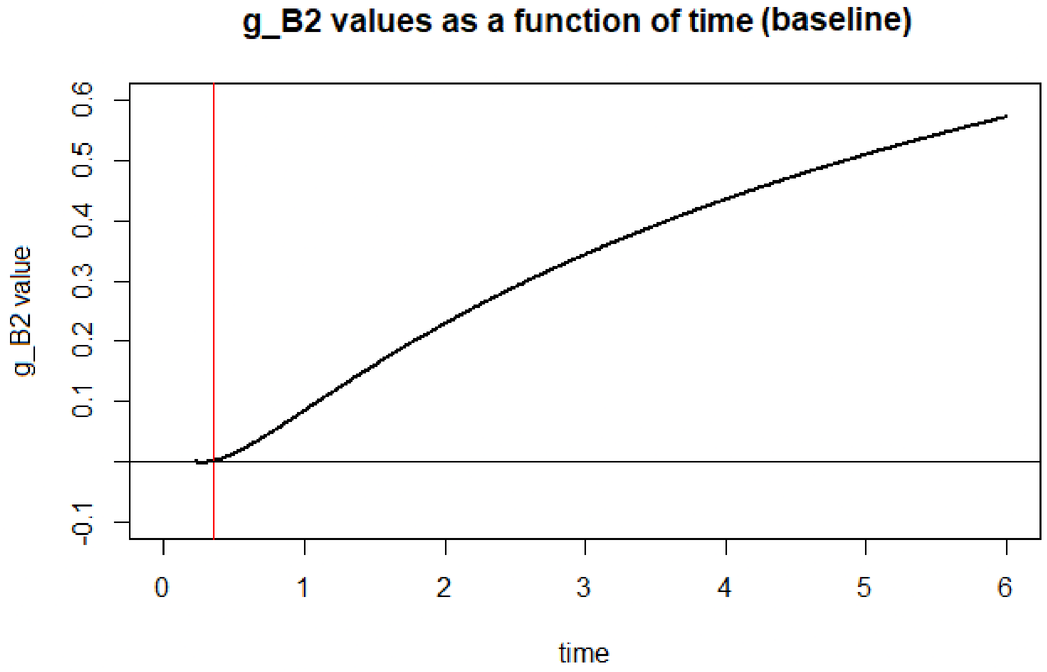
survival functions of holding time in states 1 and 2 can be found in Section B.1.

To summarise, the key features of the data simulated from this model are that the active and control treatment arms have no significant differences in how quickly they leave any state except for state 3, with easily detectable differences in the rate at which they leave it. The transitions all have Weibull hazard functions for given transitions $i \to j$, with shape parameters b_{ij} close to unity (mostly slightly less than 1), except in state 3 where the shape parameters are significantly lower.

As discussed in Section 5.1.3, for **Test B1** we are concerned with t_0, the interior point of intersection of the two survival functions of holding time associated with state 3 such that $g_{B1}(\boldsymbol{\theta}) = 0$. For **Test B2** we are concerned with t^*, which is the value of t_1 such that $g_{B2}(\boldsymbol{\theta}) = 0$. In this setup, $g_{B1}(\boldsymbol{\theta}) \approx 0$ when $t_0 \approx 0.272$. Figure 5.4a shows this visually by means of a plot of the values of $g_{B1}(\boldsymbol{\theta})$ as a function of time. Furthermore, we have that $g_{B2}(\boldsymbol{\theta}) \approx 0$ when $t_1 = t^* \approx 0.353$. Figure 5.4b illustrates this by showing a plot of $g_{B2}(\boldsymbol{\theta})$ as a function of time. We note that the function g_{B1}, as a function of time, has a global maximum. This



(a) (Baseline) Plot of $g_{B1}(\boldsymbol{\theta})$ as a function of time. The vertical red line is at $t_0 \approx 0.272$ which leads to $g_{B1}(\boldsymbol{\theta}) \approx 0$.



(b) (Baseline) Plot of $g_{B2}(\boldsymbol{\theta})$ as a function of time. The vertical red line is at $t_0 \approx 0.353$ which leads to $g_{B2}(\boldsymbol{\theta}) \approx 0$.

Figure 5.4: (Baseline) Both $g_{B1}(\boldsymbol{\theta})$ and $g_{B2}(\boldsymbol{\theta})$ as functions of time. The function values are calculated using the true parameter values of $\boldsymbol{\theta}$.

means that the power of **Test B1** may not be monotonic (non-decreasing). On the other hand, the function g_{B2} is non-decreasing and so we should expect the power of **Test B2** to be monotonic (non-decreasing).

5.2.2 Results

Test A

Table 5.2 summarises the results for **Test A** associated with the baseline setup. We can see from the first column of Table 5.2 that, regardless of sample size, the empirical type I error rates look sensible for the most part. The empirical type I error rates seem a little low relative to the respective nominal levels in some cases when fitting the correctly-specified model to the data, but this is most likely due to sampling variability. This issue can most likely be mitigated by choosing a larger value for the number of simulated datasets (M) and therefore reducing the Monte Carlo standard error of the simulation. We note that datasets of different sample sizes were simulated to verify the value of the empirical type I error rate in this case and we did not find this anomaly to be a consistent issue.

Figure 5.5 shows the distributions of $g_A(\hat{\boldsymbol{\theta}})$ for each model fit and sample size. The thick black line is the theoretical asymptotic Gaussian distribution of $g_A(\hat{\boldsymbol{\theta}})$, while the dotted blue, brown, and orange lines are that for the Weibull, exponential, and gamma fits respectively. The variance of the asymptotic Gaussian distribution of $g_A(\hat{\boldsymbol{\theta}})$ is approximated by taking the inverse of the negative average of $M = 1000$ numerical Hessian matrices as reported by `Rsolnp` ([Ghalanos and Theussl, 2015]). The vertical red line denotes the true value of the difference in average total sojourn time, $g_A(\boldsymbol{\theta})$.

One observation is that we have almost identical distributions of $g_A(\hat{\boldsymbol{\theta}})$ when the exponential and gamma model fits are concerned. It is noted that these misspecified models add bias to our estimates – this is more apparent when $m = 1000$. For $m = 1000$, even the correct Weibull model fit appears to add some positive skew to the shape of the distribution. This suggests that a sample of size $m = 1000$ may

Type I error ($m = 10000$)				Power ($m = 10000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.007	0.037	0.083	ρ (Weibull)	1	1	1
ε (Exponential)	0.009	0.049	0.098	ρ (Exponential)	0.989	0.996	0.997
ε (Gamma)	0.007	0.046	0.093	ρ (Gamma)	0.998	0.998	0.998
Type I error ($m = 1000$)				Power ($m = 1000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.007	0.033	0.080	ρ (Weibull)	1	1	1
ε (Exponential)	0.009	0.049	0.098	ρ (Exponential)	0.988	0.993	0.994
ε (Gamma)	0.008	0.039	0.078	ρ (Gamma)	1	1	1

Table 5.2: (Baseline, **Test A**) Table summarising results of empirical type I error rates and power. The first column shows results for empirical type I error rates, ε, under H_0. The two different treatment arms have the same model parameters. The second column shows the estimated power of the test when both treatment arms are different, with patients in active treatment potentially benefiting in state 3. Each of the tests are performed using parameters associated with: (*i*) the true Weibull model fit, (*ii*) an exponential model fit, and (*iii*) a gamma model fit. The sample size is either $m = 10000$ or $m = 1000$.

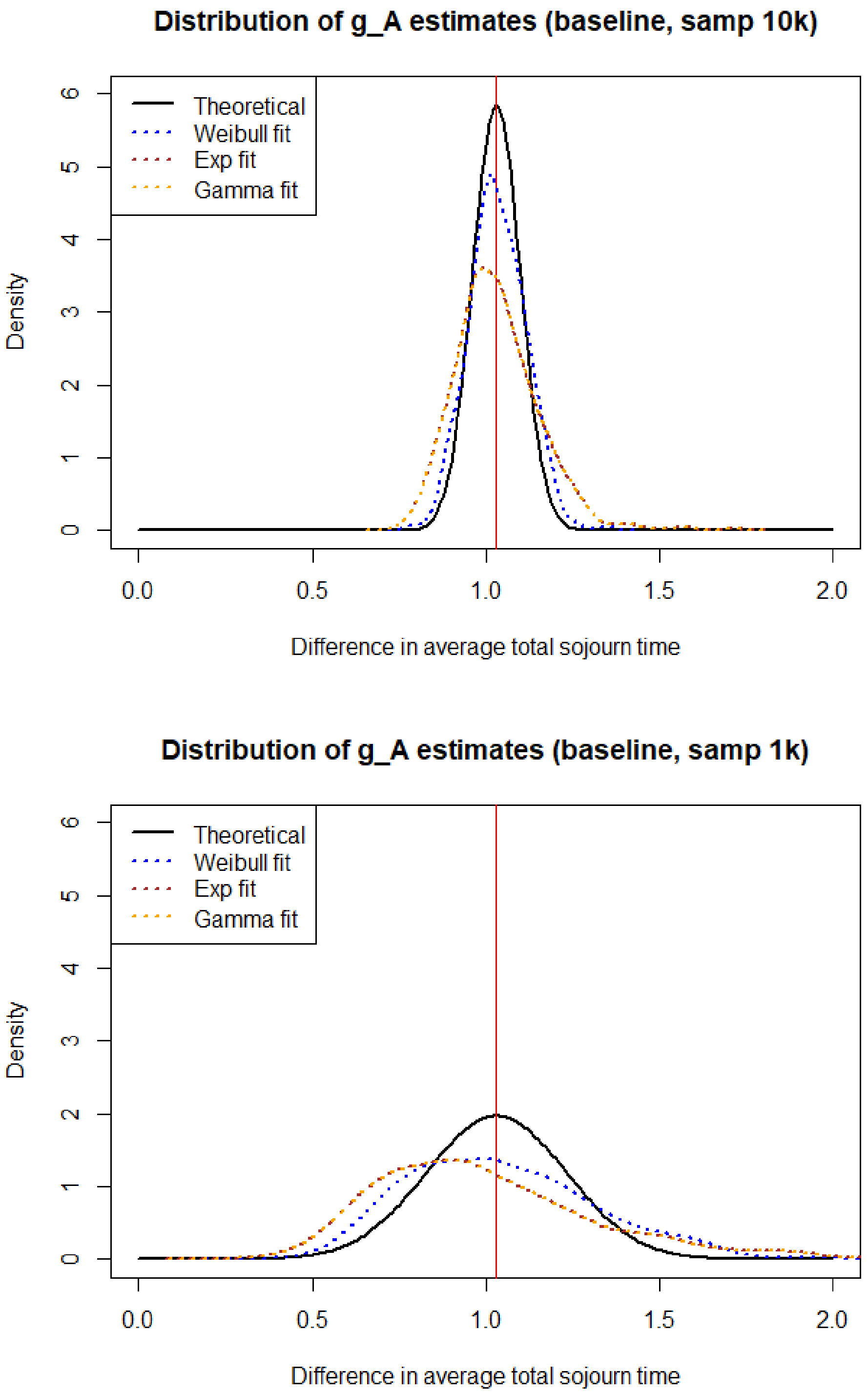


Figure 5.5: (Baseline) Distribution of $g_A(\hat{\boldsymbol{\theta}})$, for each of $m = 10000$ and $m = 1000$. The thick black lines are the respective theoretical asymptotic Gaussian distributions with variances approximated by taking the inverse of the negative average of $M = 1000$ numerical Hessian matrices while the dotted blue, brown, and orange lines are estimated densities for the Weibull, exponential, and gamma fits respectively. The vertical red lines denote the true values of the average total sojourn time, $g_A(\boldsymbol{\theta})$, in each case.

be insufficient in this scenario to assume that $g_A(\hat{\boldsymbol{\theta}})$ is approximately Gaussian-distributed. The non-Gaussian shape of the distribution of the test statistic is most likely caused by some of the rarer transitions relative to the total sample size.

Despite these observations, the second column of Table 5.2 shows that we are still able to correctly reject H_0 when there is a clear difference in the average total sojourn time between treatment arms. This is regardless of the sample size. Overall it would appear that model misspecification is not a major detriment in this scenario, when **Test A** is concerned. This is because the estimates seem to be robust enough to capture the fact there is significant positive benefit for patients in active treatment.

Test B

Table 5.3 and Table 5.4 summarise the results for **Test B1** and **Test B2**, respectively. It is evident from the first column of each table that Weibull model fit results in sensible values for empirical type I error rates, but the exponential and gamma model fits lead to values which are very high. In fact, the empirical type I error rates are close to one in most cases, regardless of sample size. Figure 5.6 illustrates the reason for this. It shows the true survival function of the holding time in state 3 for each treatment arm (thick black lines) while the blue, brown, and orange lines depict the average of the $M = 1000$ parametric fits associated with the Weibull, exponential and gamma models respectively.

It can be observed that all the models fit the survival function of the holding time for the control treatment arm ($Z = 0$) well, but the exponential and gamma models are very poor fits for the survival function associated with the active treatment arm ($Z = 1$). Neither model is able to fully capture the fact that the rate of leaving state 3 is decreasing with time when the active treatment arm is concerned. Thus, at $t_0 = 0.272$, both $g_{B1}(\hat{\boldsymbol{\theta}})$ and $g_{B2}(\hat{\boldsymbol{\theta}})$ are unambiguously significantly positive under the exponential and gamma model fits. This is what leads to the high rejection rates seen in the tables.

Type I error ($m = 10000$)				Power ($m = 10000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.005	0.050	0.090	$t_{80\%}$ (Weibull)	0.32	0.31	0.30
ε (Exponential)	0.997	0.998	1	$t_{80\%}$ (Exponential)	-	-	-
ε (Gamma)	0.998	0.999	0.999	$t_{80\%}$ (Gamma)	-	-	-

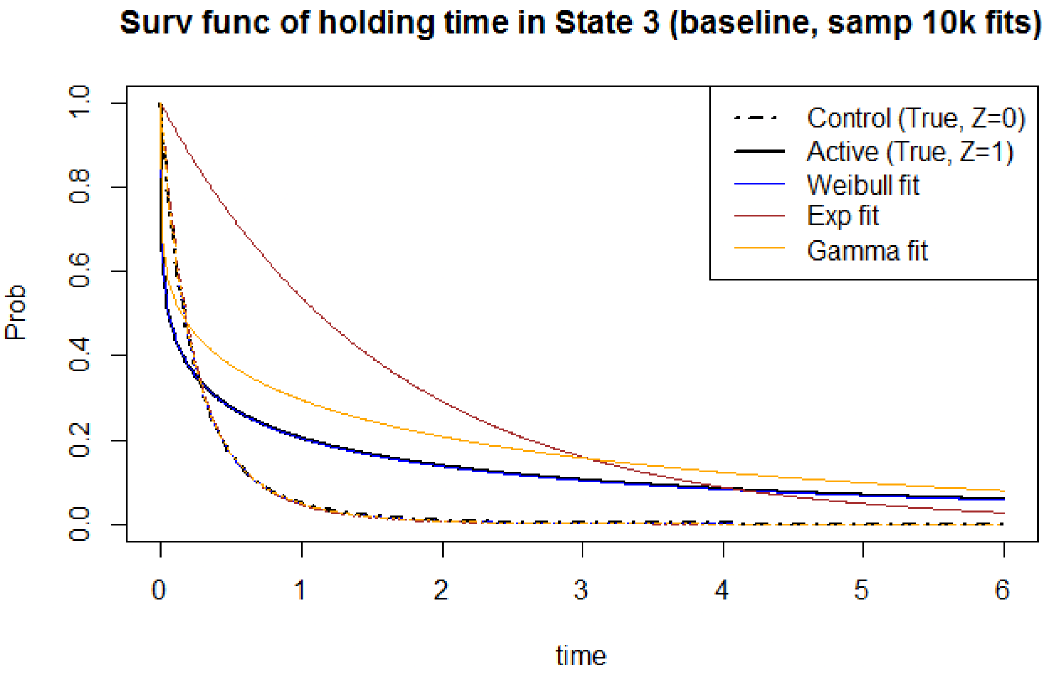
Type I error ($m = 1000$)				Power ($m = 1000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.006	0.043	0.099	$t_{80\%}$ (Weibull)	0.42	0.38	0.37
ε (Exponential)	0.997	0.997	0.999	$t_{80\%}$ (Exponential)	-	-	-
ε (Gamma)	0.828	0.943	0.978	$t_{80\%}$ (Gamma)	-	-	-

Table 5.3: (Baseline, **Test B1**) Table summarising results of empirical type I error rates and power. The first column shows results for empirical type I error rates, ε, under H_0. The test is carried out at $t_0 = 0.272$, which leads to $g_{B1}(\boldsymbol{\theta}) \approx 0$. The second column shows the value of $t_{80\%}$, which is defined as the smallest value of t which gives us estimated power of at least 0.8. The closer $t_{80\%}$ is to $t_0 = 0.272$, the more powerful the test is. Each of the tests are performed using parameters associated with: (*i*) the true Weibull model fit, (*ii*) an exponential model fit, and (*iii*) a gamma model fit. The sample size is either $m = 10000$ or $m = 1000$.

Type I error ($m = 10000$)				Power ($m = 10000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.005	0.050	0.091	$t_{80\%}$ (Weibull)	0.45	0.43	0.42
ε (Exponential)	0.997	0.998	1	$t_{80\%}$ (Exponential)	-	-	-
ε (Gamma)	0.998	0.999	0.999	$t_{80\%}$ (Gamma)	-	-	-

Type I error ($m = 1000$)				Power ($m = 1000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.006	0.043	0.101	$t_{80\%}$ (Weibull)	0.72	0.62	0.57
$t_{80\%}$ (Exponential)	0.997	0.997	0.999	$t_{80\%}$ (Exponential)	-	-	-
ε (Gamma)	0.832	0.945	0.977	$t_{80\%}$ (Gamma)	-	-	-

Table 5.4: (Baseline, **Test B2**) Table summarising results of empirical type I error rates and power. The first column shows results for empirical type I error rates, ε, under H_0. The test is carried out at $t^* = 0.353$, which leads to $g_{B2}(\boldsymbol{\theta}) \approx 0$. The second column shows the value of $t_{80\%}$, which is defined as the smallest value of t_1 which gives us estimated power of at least 0.8. A value of "-" indicates that it was not possible to obtain a sensible value of $t_{80\%}$ (which must be greater than t^*. The closer $t_{80\%}$ is to $t^* = 0.353$, the more powerful the test is. Each of the tests are performed using parameters associated with: (*i*) the true Weibull model fit, (*ii*) an exponential model fit, and (*iii*) a gamma model fit. The sample size is either $m = 10000$ or $m = 1000$.



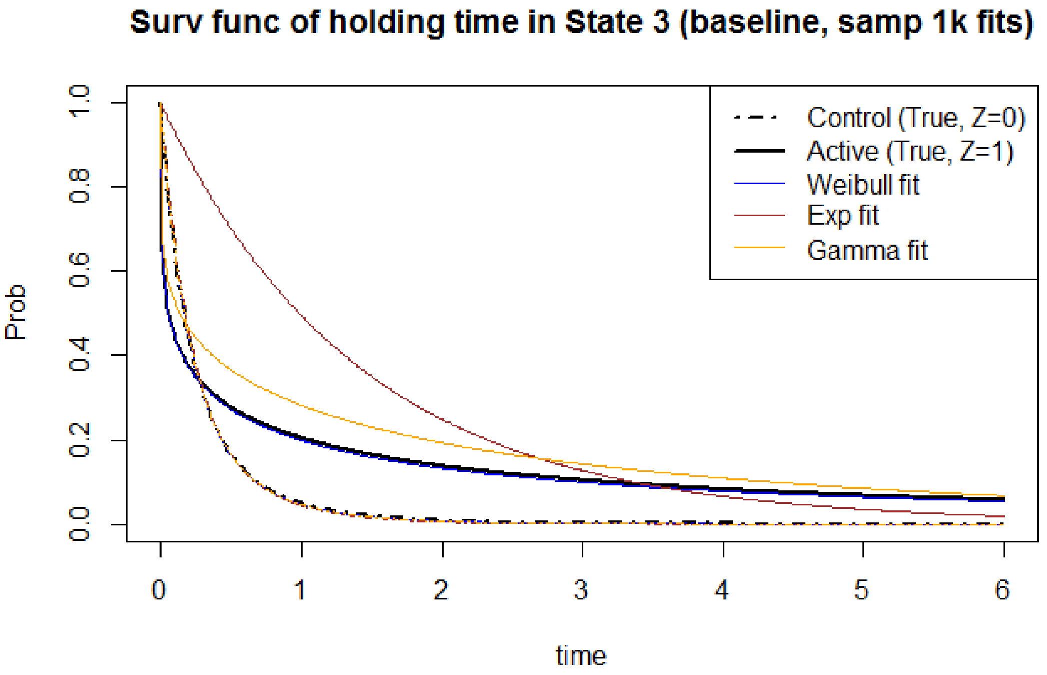


Figure 5.6: (Baseline) Parametric model fits: the survival function of holding time for each treatment arm in state 3, for each of $m = 10000$ and $m = 1000$. The thick black lines are the true survival functions while the blue, brown, and orange lines are that of the average of $M = 1000$ fits associated with the Weibull, exponential, and gamma models respectively.

It is for this reason that there are no sensible values of $t_{80\%}$ in the second column of each table whenever the exponential and gamma models are being considered, as there is almost 100% rejection of H_0 even at values of $t \ll 0.272$ (for **Test B1**) and $t_1 \ll 0.353$ (for **Test B2**) when patients in active treatment are actually worse off due to having lower probabilities of being event-free. This can be seen in Figures 5.7a and 5.7b which show the estimated power functions.

On the other hand, because the Weibull model fit of the survival function of the holding time in state 3 is good when $Z = 1$, we have power functions which look sensible since there does not seem to be significant rejection before each of $t = 0.272$ (**Test B1**) or $t_1 = 0.353$ (**Test B2**). The power of each test increases only when we start to have positive deviation away from each of these two values. Overall, it seems that **Test B** can be very sensitive to poor model fits and is not robust to model misspecification.

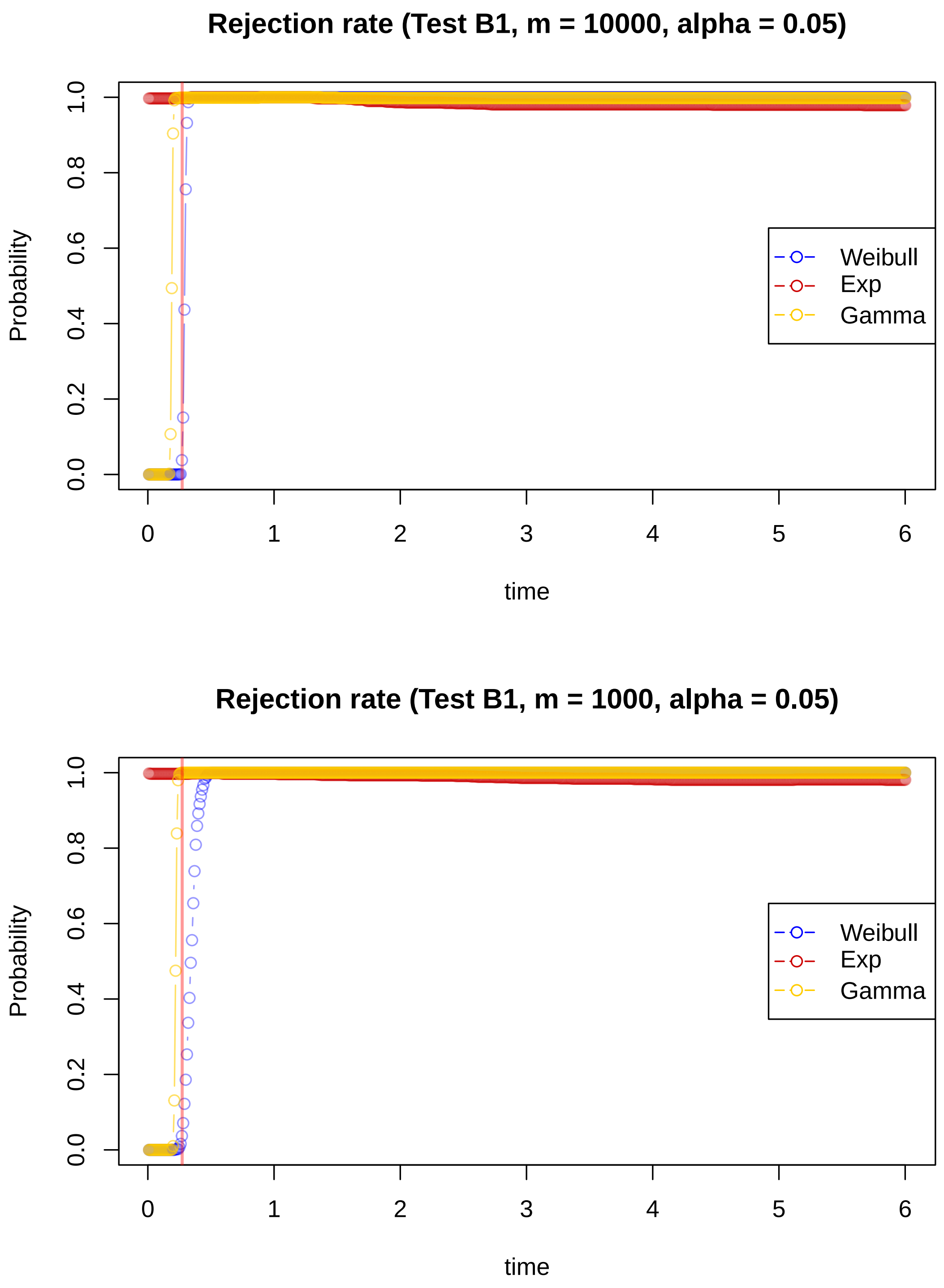
See Section 5.5 for copies of Tables 5.2 – 5.4, displayed alongside tables associated with the other model setups discussed this chapter.

5.3 Model with significant right-censoring

Section 5.3.1 describes the setup for a model similar to the baseline model, except with a significant amount of right-censoring incorporated. Section 5.3.2 discusses the findings and makes comparisons to the results in Section 5.2.2.

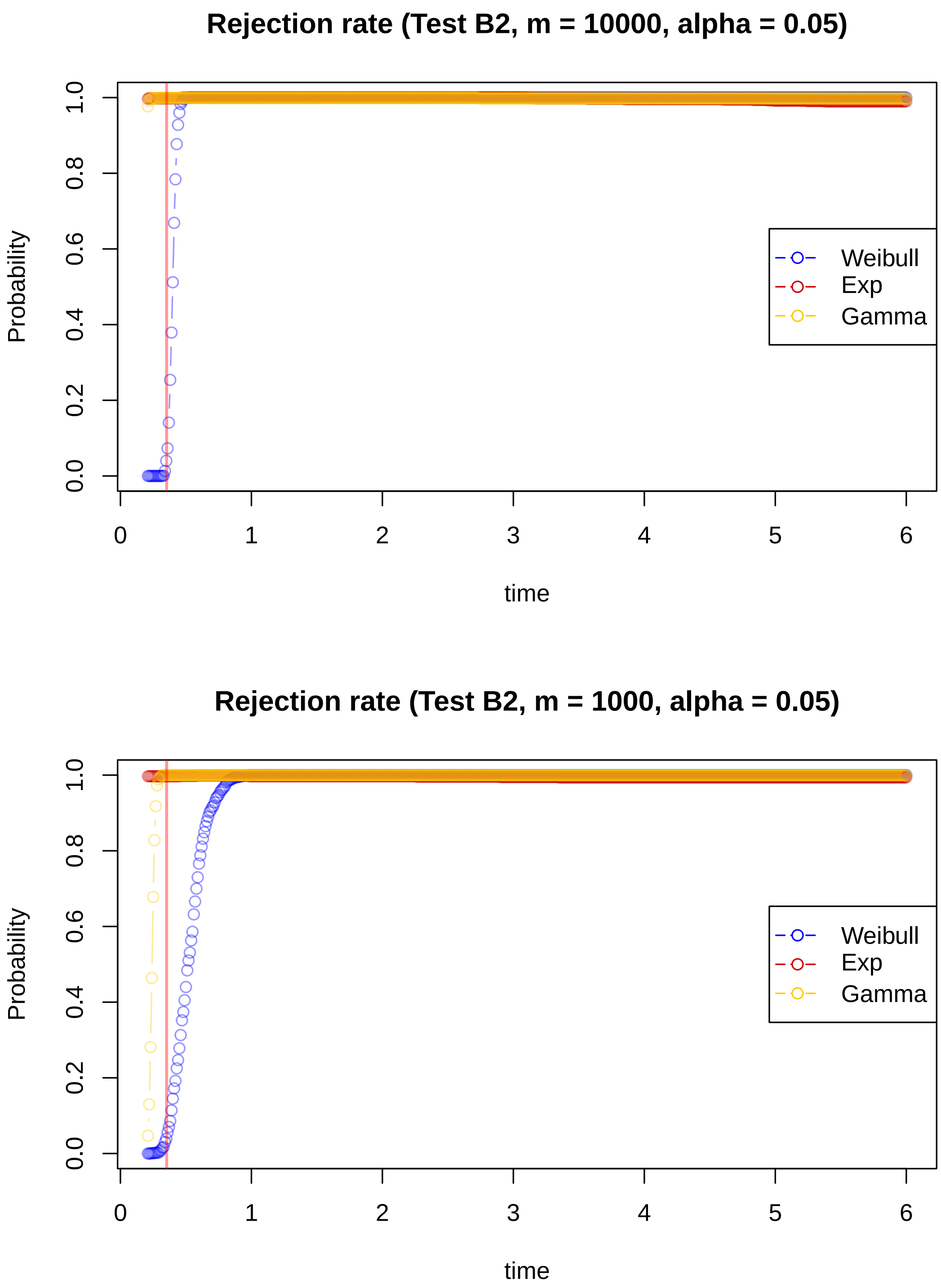
5.3.1 Description of model

The setup of this model is almost identical to that described in Section 5.2.1. The only difference is that right-censoring is incorporated. Specifically, the data are simulated using **Algorithm 1** where there is a single random censoring time simulated at the start of the semi-Markov process. Then, the SMP is allowed to evolve until either an absorbing state is reached, or the censoring time is exceeded before an absorbing state is reached. If the censoring time is exceeded before an



(a) Estimates of power of **Test B1** as a function of time.

Figure 5.7: (Baseline) Estimates of test power associated with **Test B** as described in Section 5.1.3.



Rejection rate (Test B2, m = 10000, alpha = 0.05)

Rejection rate (Test B2, m = 1000, alpha = 0.05)

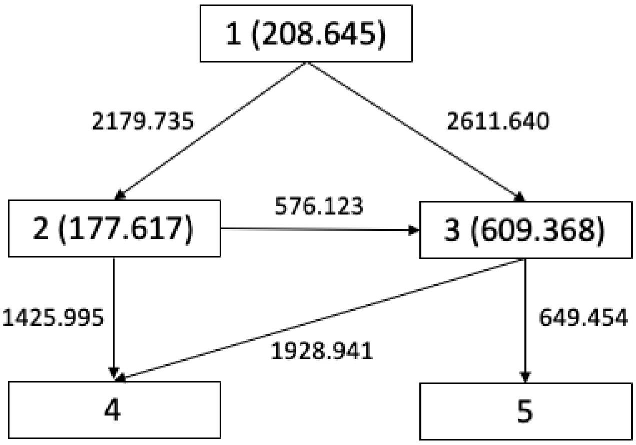
(b) Estimates of power **Test B2** as a function of time.

Figure 5.7: (Baseline) Estimates of test power associated with **Test B** as described in Section 5.1.3.

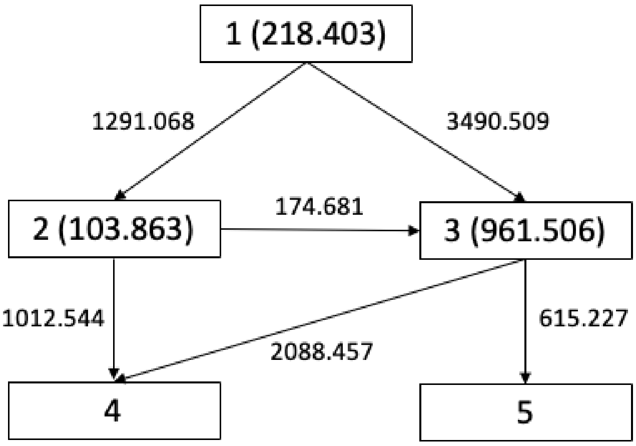
absorbing state is reached, the last (non-absorbing) observed state is noted and the time spent in that state is the right-censoring time in that state.

The distribution of the censoring time generated at the start of the SMP is chosen to be Uniform$(0.05, 1.5)$. The parameters of this distribution are chosen deliberately such that the proportion of patients censored in state 3 is relatively large. This setup would represent an example of a clinical trial which is too short. As before, there are 5000 patients in each treatment arm. Figure 5.8 shows the average number of transitions for each $i \to j$ as well as the average numbers censored in each non-absorbing state. The numbers alongside the arrows depict the average number of observed transitions, while the numbers in brackets in the boxes depict the average numbers right-censored in those states.

We can see that an average of roughly 6853 patients reached state 3, of which roughly an average of 1571 were censored there. This is a censoring rate of roughly 22.9% in state 3 given that patients reach state 3, and a censoring rate of roughly 15.7% in state 3 overall. Furthermore, by comparing with Figure 5.2, we can see that the number of observations associated with $3 \to 5$ is reduced drastically. This is because many of such observations have been right-censored due to the fact that the parameters associated with $3 \to 5$ are associated with a much lower rate of transition out of state 3. This means that patients associated with transition $3 \to 5$ are having much larger transition times compared to their peers who experience transition $3 \to 4$. It will be seen in Section 5.3.2 that, whenever there is model misspecification, this will lead to very different results as compared to those obtained in Section 5.1.3.



(a) $Z = 0$



(b) $Z = 1$

Figure 5.8: 5-state model with six possible transitions depicting the baseline model with censoring. Figure 5.2 depicts the multi-state model. There are 5000 patients in each treatment arm. The numbers alongside the arrows depict the average number of observed transitions, while the numbers in brackets in the boxes depict the average number right-censored in those states.

5.3.2 Results and comments

Test A

The results for **Test A** are summarised in Table 5.5. We observe that we are rejecting H_0 when we should be – in the case of $m = 10000$ we have 100% rejection rate regardless of fit and regardless of α, while for $m = 1000$ we have well above 80% rejection rate for $\alpha = 0.05$ and $\alpha = 0.10$.

Figure 5.9 shows the distributions of $g_A(\hat{\boldsymbol{\theta}})$ for each model fit, as before. The bold black line is associated with the theoretical Gaussian asymptotic variance of the estimator $g_A(\hat{\boldsymbol{\theta}})$ (which takes the censoring into account). However, this time the estimated distributions look different from before. Regardless of sample size, we notice a relatively large variance when the Weibull fit is concerned, and there is a larger amount of (negative) bias associated with the misspecified models. Another interesting observation is that the variance of $g_A(\hat{\boldsymbol{\theta}})$ associated with the exponential fit has much smaller variance than that of the other fits, though the bias is the largest.

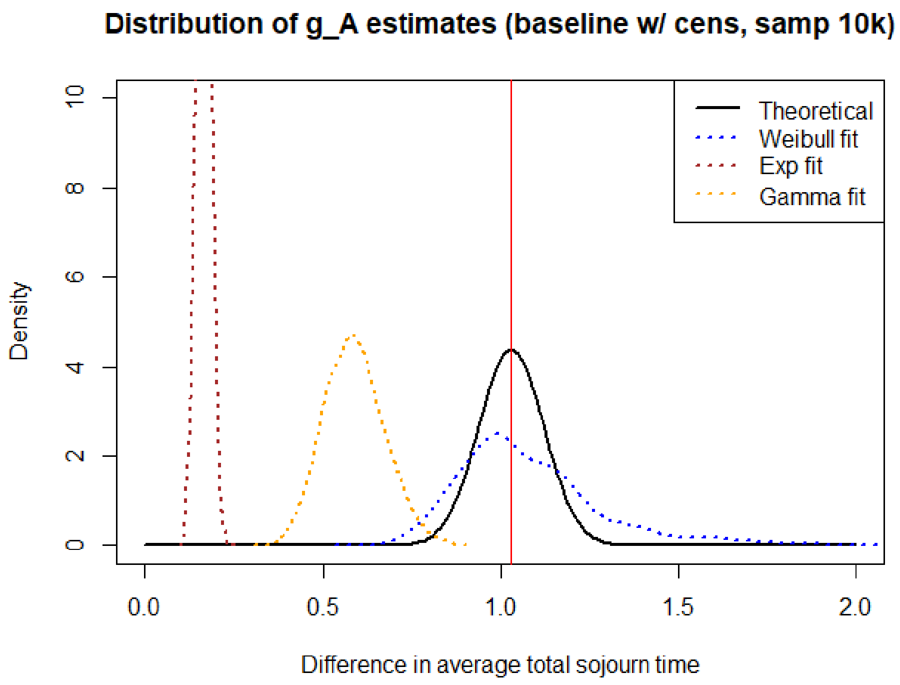
Further investigation reveals that the Weibull parameter estimates associated with $Z = 1$ and state 3 (where the right-censoring rate is the highest and $3 \to 5$ is a rare transition) have a high amount of variability when $m = 1000$, especially for the rate parameters. Furthermore, histograms of these rate parameters (Figure 5.10) depict positively-skewed distributions as opposed to approximate Gaussian distributions. Based on this, it is most likely that the variances associated with the respective asymptotic Gaussian distributions are not well-approximated by using the inverse of negative average of $M = 1000$ numerical Hessian matrices, and so the results in Table 5.5 for the Weibull fit are likely to be unreliable when $m = 1000$.

As previously mentioned, this particular scenario depicts a clinical trial which ends too prematurely. These findings emphasise the importance of ensuring clinical trials are not ended too prematurely. This point is further discussed in Section 7.1.

Type I error ($m = 10000$)				Power ($m = 10000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.008	0.044	0.093	ρ (Weibull)	1	1	1
ε (Exponential)	0.008	0.049	0.106	ρ (Exponential)	1	1	1
ε (Gamma)	0.006	0.051	0.086	ρ (Gamma)	1	1	1

Type I error ($m = 1000$)				Power ($m = 1000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.003	0.034	0.081	ρ (Weibull)	0.473	0.876	0.967
ε (Exponential)	0.006	0.048	0.091	ρ (Exponential)	0.818	0.952	0.982
ε (Gamma)	0.003	0.040	0.092	ρ (Gamma)	0.192	0.944	0.994

Table 5.5: (Baseline with censoring, **Test A**) Table summarising results of type I error estimates and power. The first column shows results for estimated type I errors, ε, under H_0. The two different treatment arms have the same model parameters. The second column shows the estimated power of the test when both treatment arms are different, with patients in active treatment potentially benefiting in state 3. Each of the tests are performed using parameters associated with: (*i*) the true Weibull model fit, (*ii*) an exponential model fit, and (*iii*) a gamma model fit. The sample size is either $m = 10000$ or $m = 1000$.



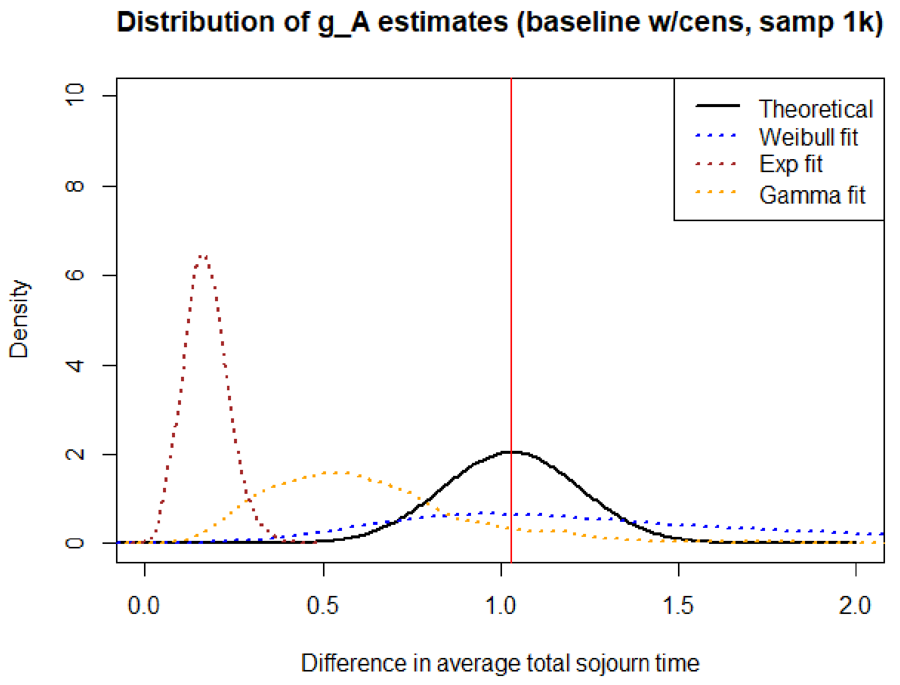
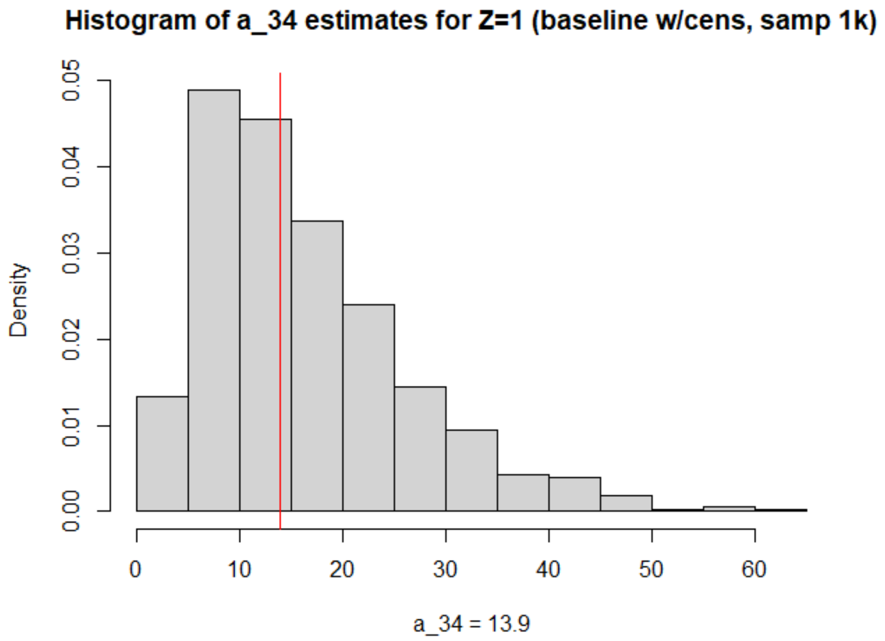


Figure 5.9: (Baseline with censoring) Distribution of $g_A(\hat{\boldsymbol{\theta}})$, for each of $m = 10000$ and $m = 1000$. The thick black lines are the respective theoretical asymptotic Gaussian distributions with variances approximated by taking the inverse of the negative average of $M = 1000$ numerical Hessian matrices while the dotted blue, brown, and orange lines are estimated densities for the Weibull, exponential, and gamma fits respectively. The vertical red lines denote the true value of the average total sojourn time, $g_A(\boldsymbol{\theta})$, in each case.



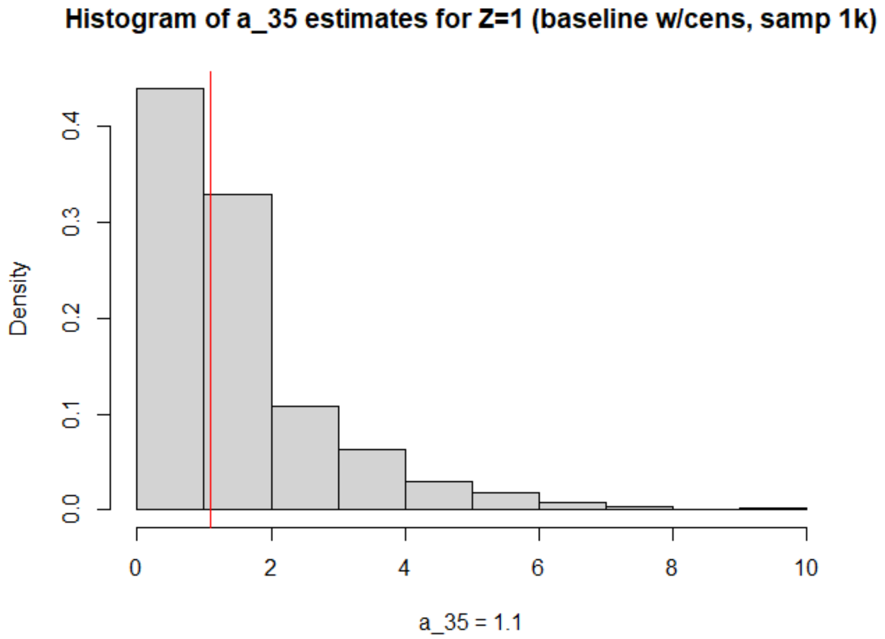


Figure 5.10: (Baseline with censoring) Distribution of $M = 1000$ Weibull rate parameter estimates associated with $Z = 1$ and state 3 for $m = 1000$. The vertical red lines denote the true parameter values. It can be seen that the maximum likelihood parameter estimates show a large amount of variability and do not look approximately Gaussian.

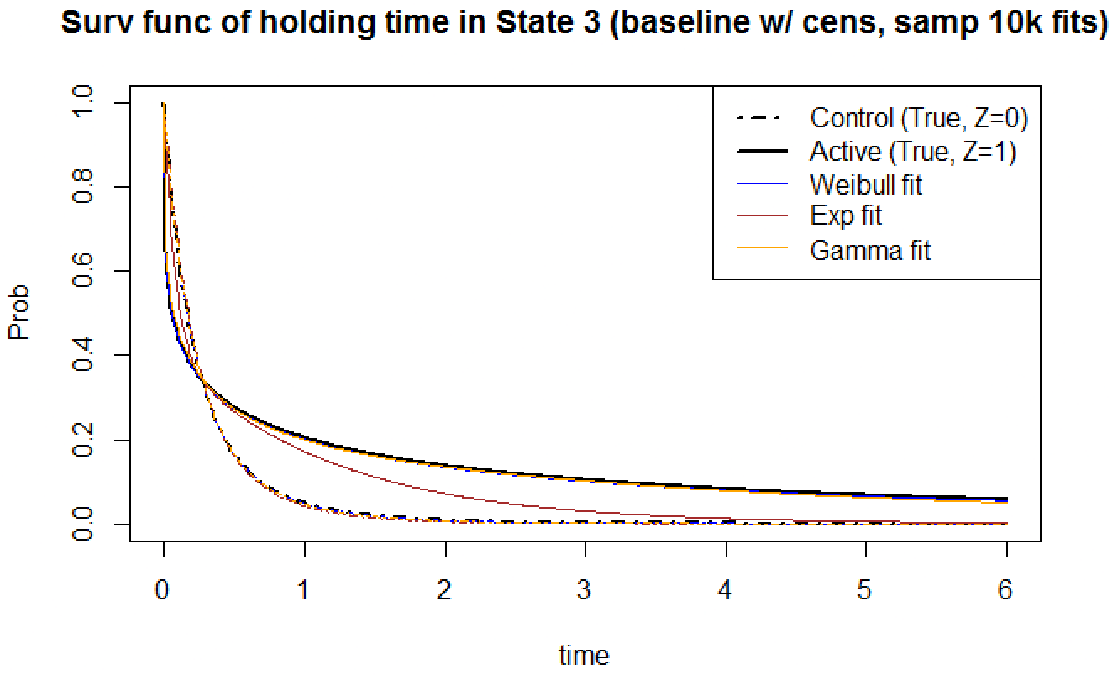
Test B

Table 5.6 and Table 5.7 summarise the results for **Test B1** and **Test B2** respectively.

Noting that the results for $m = 1000$ may be unreliable as per the findings associated with **Test A**, the first column of both tables 5.6 and 5.7 show sensible empirical type I error rates for the Weibull model fit. However, surprisingly, we no longer observe the unusually high values of ε associated with the misspecified models as seen in Table 5.3. In fact, they look reasonable save for a few larger values associated with $\alpha = 0.10$. Figure 5.11 might offer a reasonable explanation – it appears that the estimated survival functions of holding time in state 3 associated with misspecified models are closer to their true values as compared to the baseline model (compare this with Figure 5.6). The reason for this is most likely because of the observation in Section 5.3.1 that a larger proportion of censored observations are associated with transition 3 → 5 and these right-censored observations are associated with very large transition times. In fact, for these misspecified models the average of $M = 1000$ estimates of p_{34} for $Z = 1$ is consistently under-estimated (with average values close to 0.50) when the true value is 0.7. Equivalently, $p_{35} = 1 - p_{34}$ is consistently over-estimated for $Z = 1$. The fact that so many of the large transition times (associated with $3 \to 5$) are unobserved is leading to poor estimates of all the parameters associated with state 3. See Appendix A for further exploration and discussion of this result.

Despite this phenomenon, we find that we have sensible-looking power functions as per Figures 5.12a and 5.12b. However, it seems that the results cannot be trusted due to the aforementioned observations. It would appear that the detrimental effects of misspecification on **Test B** are exacerbated greatly by significant amounts of right-censoring. Once again, these results demonstrate that the effects of model misspecification might be greatly exacerbated in certain scenarios where clinicial trials end too prematurely. This point is discussed in Section 7.1.

See Section 5.5 for copies of Tables 5.5 – 5.7, displayed alongside tables associated with other model setups discussed in this chapter.



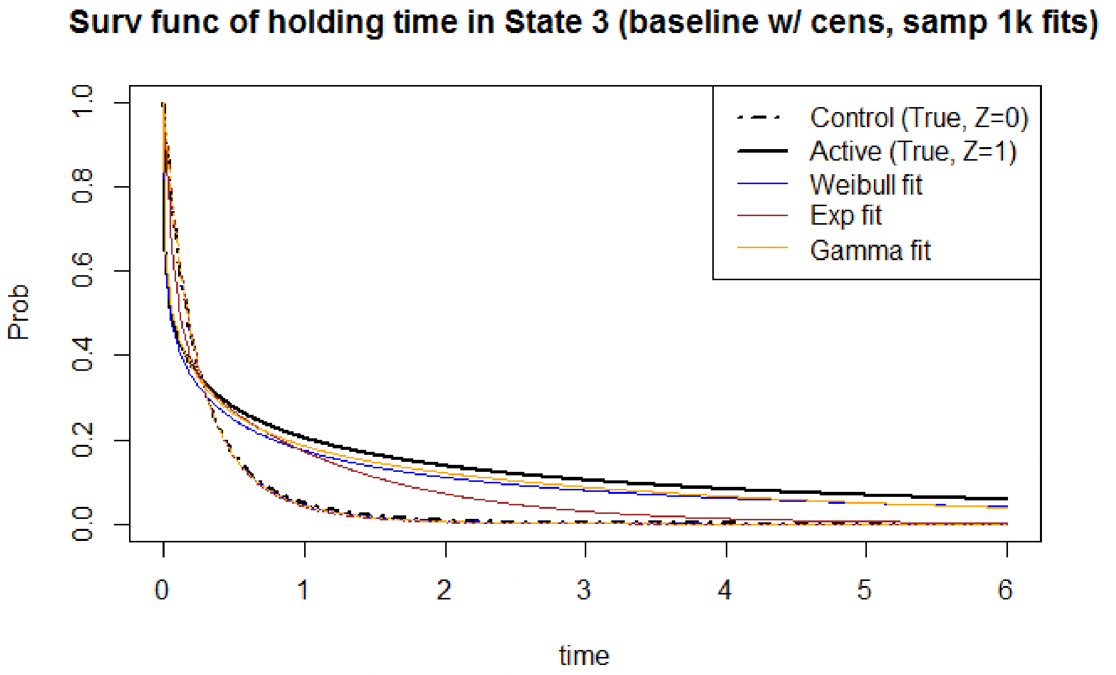


Figure 5.11: (Baseline with censoring) Parametric model fits: the survival function of holding time for each treatment arm in state 3, for each of $m = 10000$ and $m = 1000$. The thick black lines are the true survival functions while the blue, brown, and orange lines are that for the average of $M = 1000$ fits associated with the Weibull, exponential, and gamma models respectively.

Type I error ($m = 10000$)				Power ($m = 10000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.004	0.038	0.094	$t_{80\%}$ (Weibull)	0.32	0.31	0.31
ε (Exponential)	0.004	0.024	0.062	$t_{80\%}$ (Exponential)	0.33	0.32	0.31
ε (Gamma)	0.016	0.081	0.169	$t_{80\%}$ (Gamma)	0.32	0.31	0.30

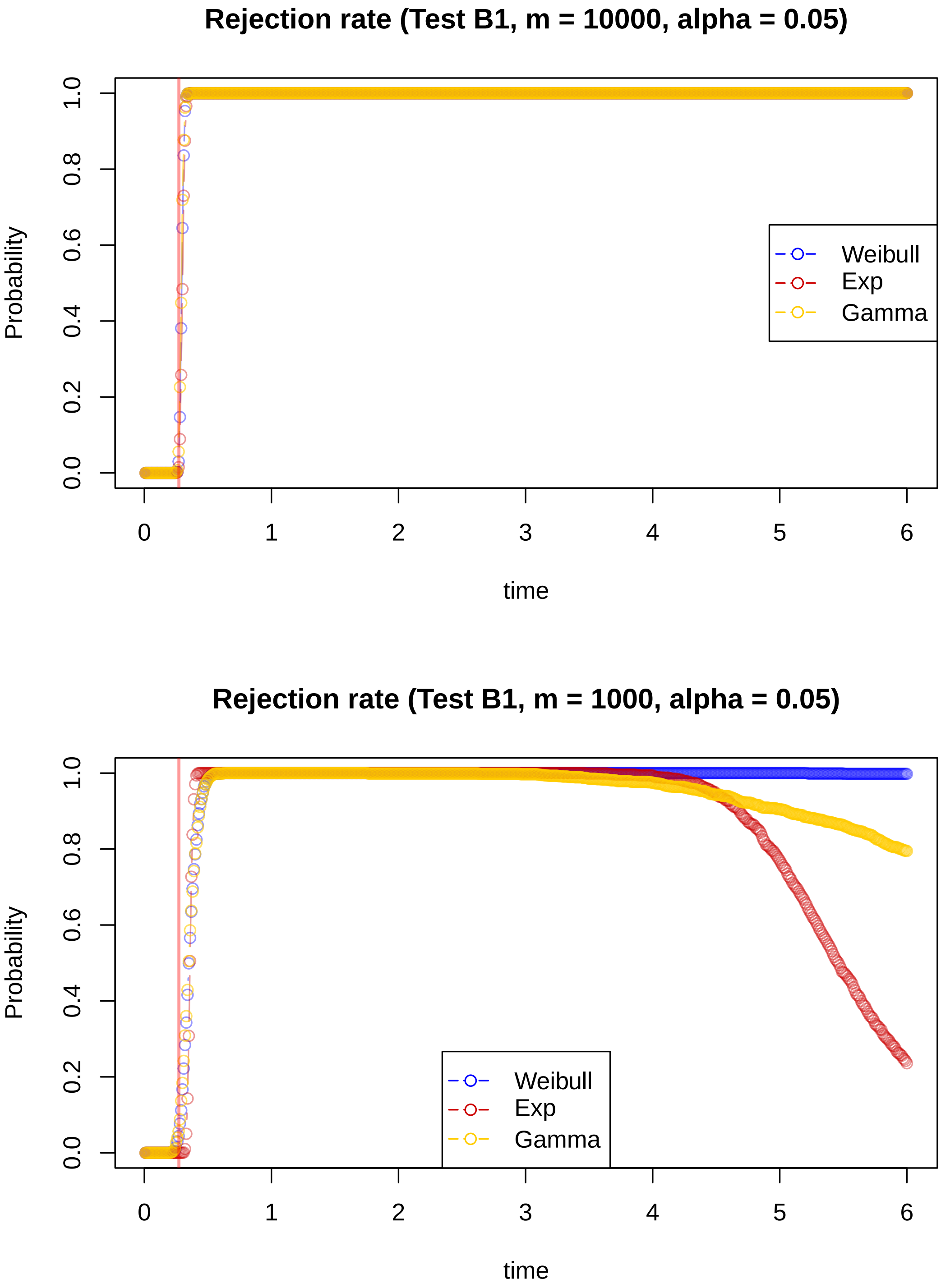
Type I error ($m = 1000$)				Power ($m = 1000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.012	0.047	0.101	$t_{80\%}$ (Weibull)	0.46	0.41	0.39
ε (Exponential)	0.002	0.055	0.100	$t_{80\%}$ (Exponential)	0.43	0.38	0.36
ε (Gamma)	0.021	0.064	0.123	$t_{80\%}$ (Gamma)	0.46	0.41	0.38

Table 5.6: (Baseline with censoring, **Test B1**) Table summarising results of empirical type I error rates and power. The first column shows results for estimated type I errors, ε, under H_0. The test is carried out at $t_0 = 0.272$, which leads to $g_{B1}(\boldsymbol{\theta}) \approx 0$. The second column shows the value of $t_{80\%}$, which is defined as the smallest value of t which gives us estimated power of at least 0.8. The closer $t_{80\%}$ is to $t_0 = 0.272$, the more powerful the test is. Each of the tests are performed using parameters associated with: (i) the true Weibull model fit, (ii) an exponential model fit, and (iii) a gamma model fit. The sample size is either $m = 10000$ or $m = 1000$.

Type I error ($m = 10000$)				Power ($m = 10000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.004	0.038	0.094	$t_{80\%}$ (Weibull)	0.46	0.44	0.43
ε (Exponential)	0.005	0.034	0.079	$t_{80\%}$ (Exponential)	0.48	0.45	0.44
ε (Gamma)	0.017	0.082	0.173	$t_{80\%}$ (Gamma)	0.42	0.43	0.46

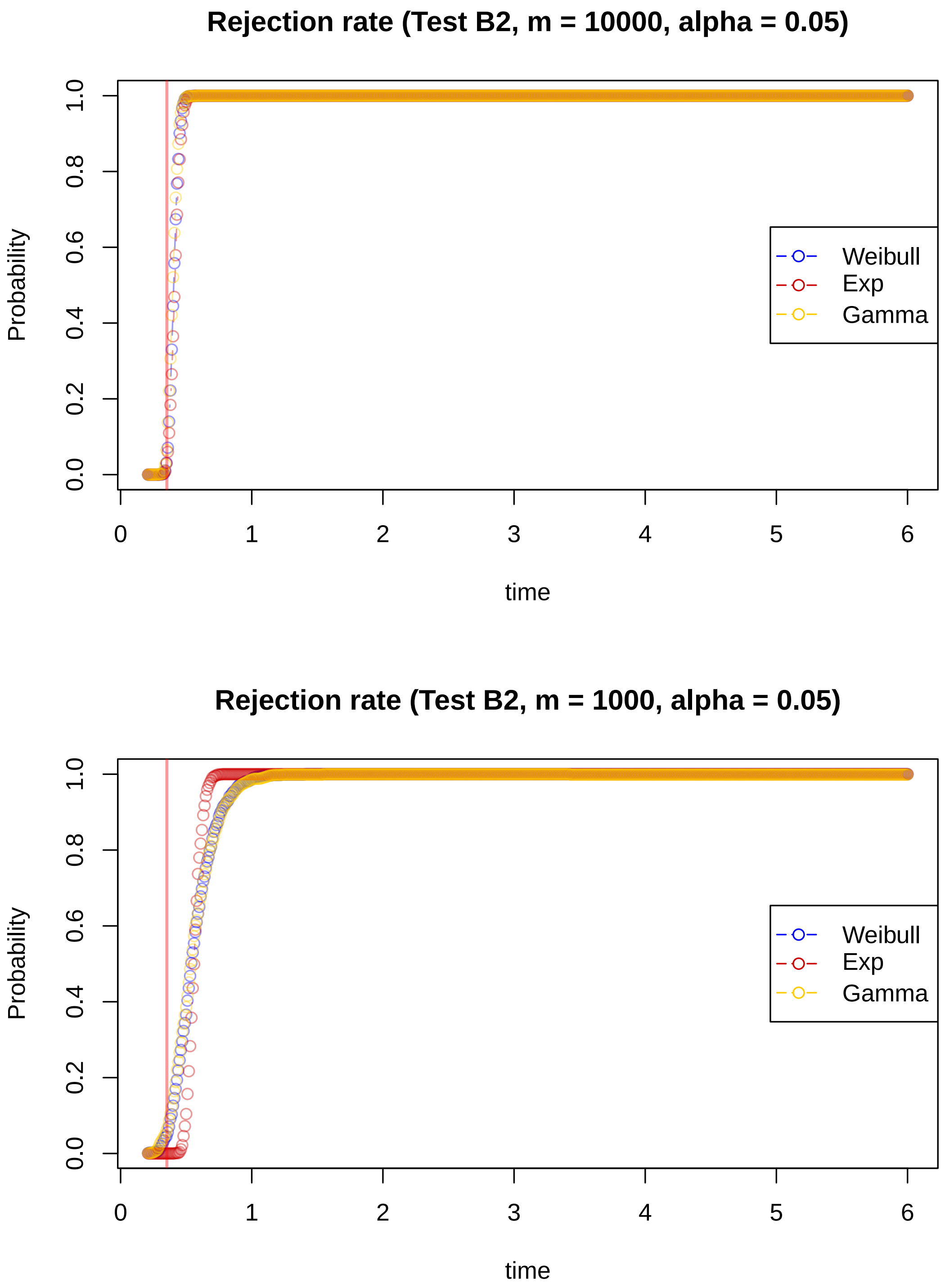
Type I error ($m = 1000$)				Power ($m = 1000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.012	0.046	0.094	$t_{80\%}$ (Weibull)	0.83	0.69	0.62
$t_{80\%}$ (Exponential)	0.021	0.059	0.106	$t_{80\%}$ (Exponential)	0.75	0.61	0.55
ε (Gamma)	0.022	0.063	0.126	$t_{80\%}$ (Gamma)	0.84	0.68	0.62

Table 5.7: (Baseline with censoring, **Test B2**) Table summarising results of empirical type I error rates and power. The first column shows results for estimated type I errors, ε, under H_0. The test is carried out at $t^* = 0.353$, which leads to $g_{B2}(\boldsymbol{\theta}) \approx 0$. The second column shows the value of $t_{80\%}$, which is defined as the smallest value of t_1 which gives us estimated power of at least 0.8. The closer $t_{80\%}$ is to $t^* = 0.353$, the more powerful the test is. Each of the tests are performed using parameters associated with: (i) the true Weibull model fit, (ii) an exponential model fit, and (iii) a gamma model fit. The sample size is either $m = 10000$ or $m = 1000$.



(a) Estimates of power of **Test B1** as a function of time.

Figure 5.12: Estimates of power of **Test B** as described in Section 5.1.3.



(b) Estimates of power of **Test B2** as a function of time.

Figure 5.12: Estimates of power of **Test B** as described in Section 5.1.3.

5.4 Model with significantly reduced patient benefit

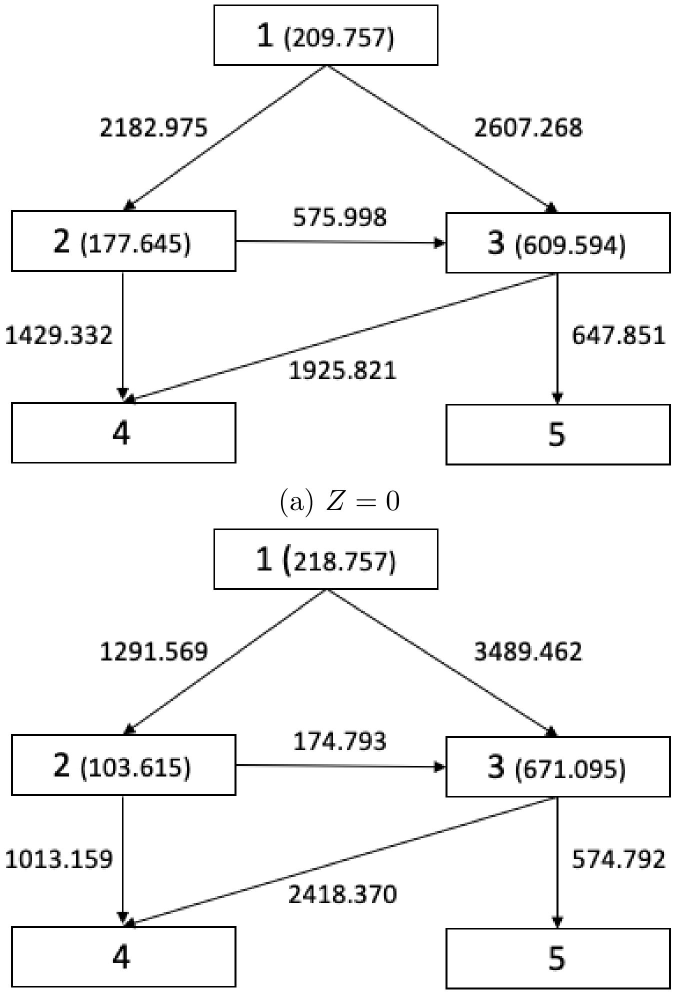
Section 5.4.1 describes the setup for a model similar to the baseline model, except with less detectable benefit in addition to right-censoring. Section 5.4.2 discusses the findings and makes comparisons to the results in the previous Sections 5.2.2 and 5.3.2.

5.4.1 Description of model

The setup of this model is almost identical to that described in Section 5.2.1, except now the amount of potential benefit to patients in the active treatment arm ($Z = 1$) is altered. This is achieved by increasing the values of shape parameters associated with state 3 so that patients transition out of state 3 more quickly as compared to the baseline model. For $Z = 1$, the new shape parameters are now $b_{34} = 0.630$ (increased from $b_{34} = 0.256$) and $b_{35} = 0.835$ and (increased from $b_{35} = 0.475$).

The same Uniform$(0.05, 1.5)$ distribution is chosen to simulate censoring times for each individual at the start of the SMP. We can see from Figure 5.13 that an average of roughly 6848 patients reached state 3, of which an average of roughly 1281 were censored there. This is a censoring rate of roughly 18.7% in state 3 given that patients reach state 3, and a censoring rate of roughly 12.8% in state 3 overall. This is less than that of the baseline model with censoring as per Section 5.3.2 (22.9% and 15.7% respectively). This is unsurprising, since we know that patients leave state 3 at a faster rate in this model compared to the baseline model.

Since the model parameters have changed, the survival functions of holding times in state 3 between both treatment arms have less significant differences. Consequently, values of $g_A(\boldsymbol{\theta})$, $g_{B1}(\boldsymbol{\theta})$, and $g_{B2}(\boldsymbol{\theta})$ have reduced overall. The new set of survival functions of holding time for each treatment arm can be seen in Figure 5.14. Figures 5.15a and 5.15b, respectively, show the new values of $g_{B1}(\boldsymbol{\theta})$ and $g_{B2}(\boldsymbol{\theta})$ as functions of time (solid lines). The dashed lines are the associated



(a) $Z = 0$

(b) $Z = 1$

Figure 5.13: 5-state model with six possible transitions depicting the model with less benefit and with censoring. There are 5000 patients in each treatment arm. The numbers alongside the arrows depict the average number of observed transitions, while the numbers in brackets in the boxes depict the average number right-censored in those states.

values of each quantity as per the baseline model. Now, we have $t_0 = 0.425$ which gives rise to $g_{B1}(\boldsymbol{\theta}) \approx 0$ and $t^* \approx 0.764$ which gives rise to $g_{B2}(\boldsymbol{\theta}) \approx 0$.

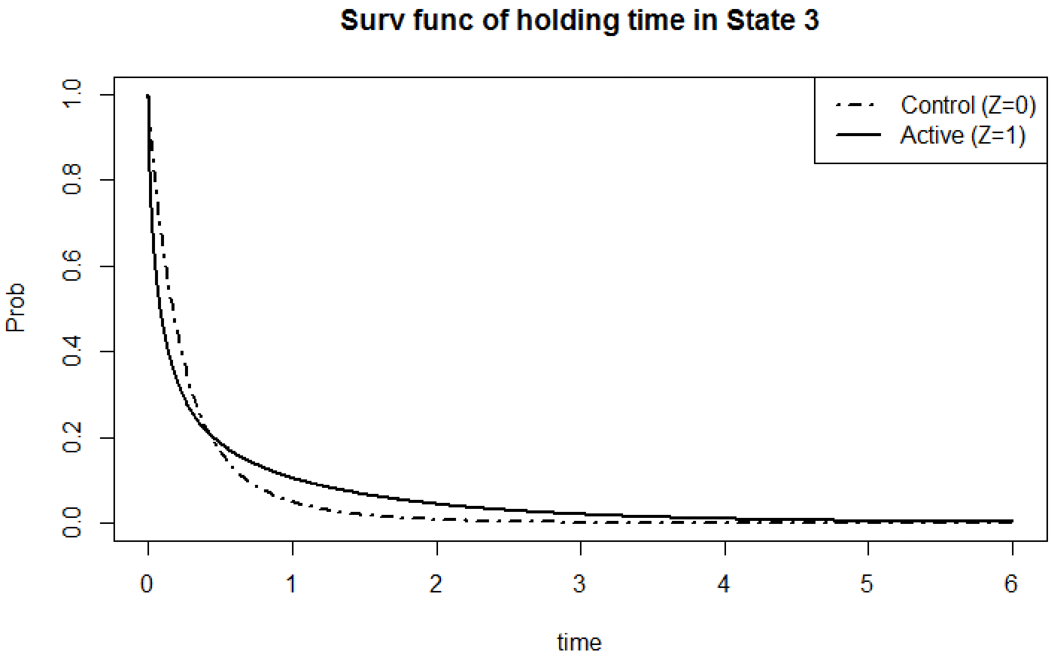
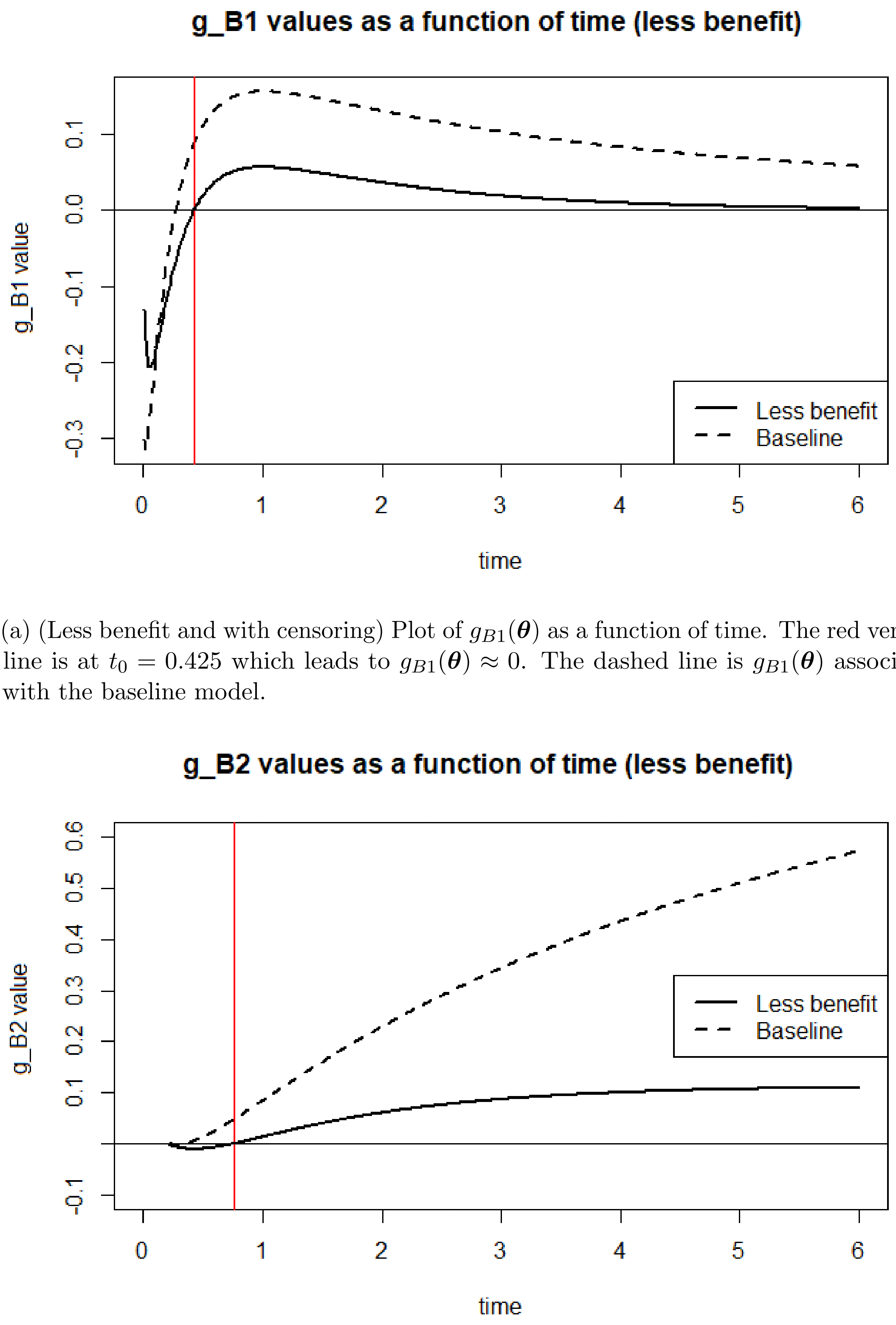


Figure 5.14: True survival functions of holding times in states 3 for model with less detectable benefit.

An interesting consequence of choosing the shape parameters as such is that now the conditional distributions associated with every possible $i \to j$ are all “close” to exponential. This is because now every shape parameter is not much smaller than unity. This will have an impact on our results in Section 5.4.2.



(a) (Less benefit and with censoring) Plot of $g_{B1}(\boldsymbol{\theta})$ as a function of time. The red vertical line is at $t_0 = 0.425$ which leads to $g_{B1}(\boldsymbol{\theta}) \approx 0$. The dashed line is $g_{B1}(\boldsymbol{\theta})$ associated with the baseline model.

(b) (Less benefit and with censoring) Plot of $g_{B2}(\boldsymbol{\theta})$ as a function of time as per the solid line. The vertical red line is at $t = 0.764$ which leads to $g_{B2}(\boldsymbol{\theta}) \approx 0$. The dashed line is $g_{B2}(\boldsymbol{\theta})$ associated with the baseline model.

Figure 5.15: (Less benefit and with censoring) Plots of $g_{B1}(\boldsymbol{\theta})$ and $g_{B2}(\boldsymbol{\theta})$ as functions of time. The function values are calculated using the true parameter values of $\boldsymbol{\theta}$.

5.4.2 Results and comments

Test A

Now that there is less detectable benefit together with censoring, we make a few interesting observations. First, by comparing Figure 5.9 and Figure 5.16, we observe that the estimates of $g_A(\boldsymbol{\theta})$ seem to have less variation than in the previous models. This is likely due to the fact that there are no obvious issues in the parameter estimates associated with state 3 that were observed in Section 5.3.2. Reducing the values of the shape parameters associated with $Z = 1$ and state 3 reduced the rate of right-censoring, which likely led to less uncertainty in the parameter estimates. Also, since the hazard rates associated with each given $i \to j$ are "close" to exponential as previously highlighted, the effects of model misspecification are also likely to be less significant. However, inspection of Figure 5.16 suggests it is still the case that negative bias is introduced for the misspecified models.

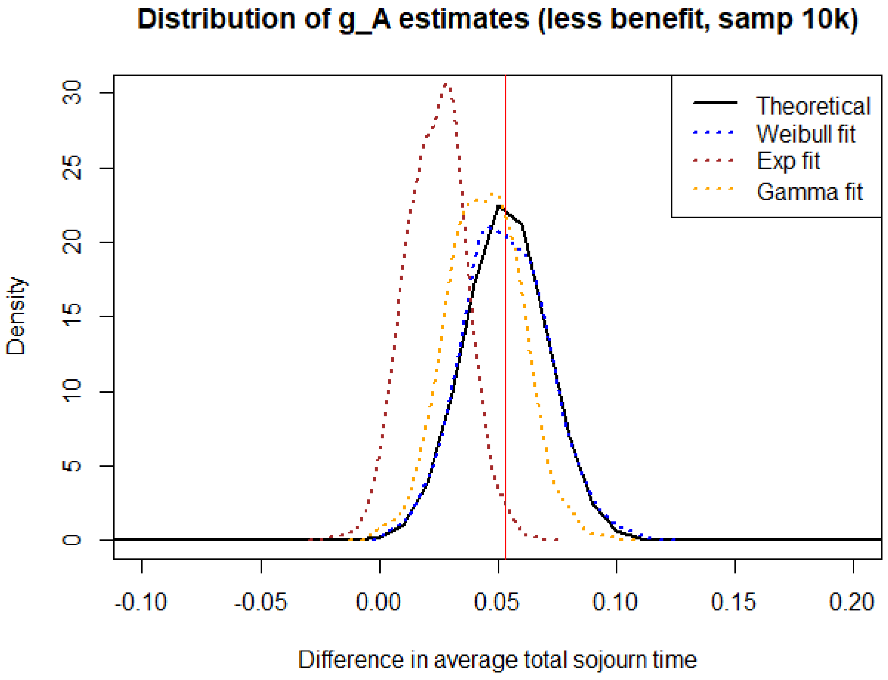
Another observation is that, because the amount of detectable benefit is relatively small, the power of the test reduces. The fact that the amount of detectable benefit is small can be observed from Figure 5.16, where the true value of $g_A(\boldsymbol{\theta})$ is relatively close to zero.

We see from the second column of Table 5.8 that we fail to attain 80% rejection rate for any α associated with the exponential model fit when $m = 10000$. A similar observation is made for every model fit when $m = 1000$.

Type I error ($m = 10000$)				Power ($m = 10000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.018	0.045	0.094	ρ (Weibull)	0.789	0.942	0.969
ε (Exponential)	0.013	0.051	0.100	ρ (Exponential)	0.374	0.627	0.745
ε (Gamma)	0.013	0.045	0.107	ρ (Gamma)	0.689	0.893	0.954

Type I error ($m = 1000$)				Power ($m = 1000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.002	0.044	0.090	ρ (Weibull)	0.01	0.165	0.333
ε (Exponential)	0.005	0.055	0.109	ρ (Exponential)	0.018	0.136	0.239
ε (Gamma)	0.003	0.045	0.091	ρ (Gamma)	0.010	0.157	0.302

Table 5.8: (Less benefit with censoring, **Test A**) Table summarising results of empirical type I error rates and power. The first column shows results for estimated type I errors, ε, under H_0. The two different treatment arms have the same model parameters. The second column shows the estimated power of the test when both treatment arms are different, with patients in active treatment potentially benefiting in state 3. Each of the tests are performed using parameters associated with: (*i*) the true Weibull model fit, (*ii*) an exponential model fit, and (*iii*) a gamma model fit. The sample size is either $m = 10000$ or $m = 1000$.



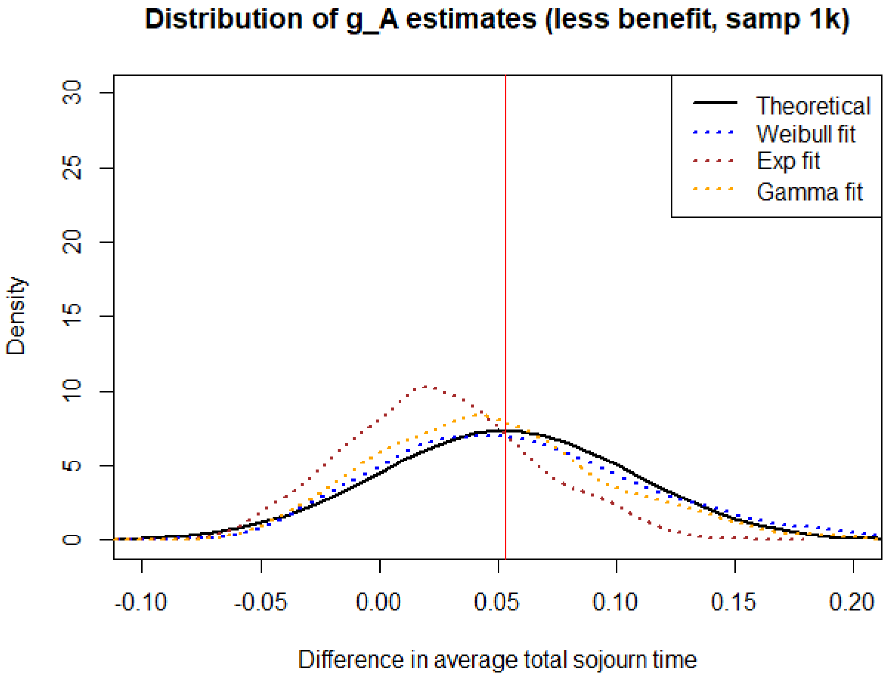


Figure 5.16: (Less benefit and with censoring) Distribution of $g_A(\hat{\boldsymbol{\theta}})$, for each of $m = 10000$ and $m = 1000$. The thick black lines are the respective theoretical asymptotic Gaussian distributions with variances approximated by taking the inverse of the negative average of $M = 1000$ numerical Hessian matrices while the dotted blue, brown, and orange lines are estimated densities for the Weibull, exponential, and gamma fits respectively. The vertical red lines denote the true value of the average total sojourn time, $g_A(\boldsymbol{\theta})$, in each case.

Test B

The results for **Test B** are similar, if not worse, than seen in the results for **Test A**. One observation from the left columns of tables 5.9 and 5.10 is that empirical type I error rates are far from the true value of α whenever the model is misspecified. We also don't have as much power as the previous models, even when $m = 10000$. This can be seen from the relative values of $t_{80\%} - t_0$ for **Test B1** (top row, right column of Table 5.9) and $t_{80\%} - t^*$ for **Test B2** (top row, right column of Table 5.10) respectively. This can also be observed visually in the top part of Figure 5.18a for **Test B1** and especially in the top part of Figure 5.18b for **Test B2**.

Looking at the bottom row, right column of Table 5.10: the situation is worse still when $m = 1000$ since we cannot attain at least 80% rejection rate for any model when **Test B2** is concerned. The bottom part of Figure 5.18b shows that the power remains constant at around 60% as t gets larger. Considering that Figure 5.15b suggests that $g_{B2}(\boldsymbol{\theta})$ is monotonic non-decreasing and concave (the integral cannot be decreasing after $t = 0.764$ for the model with less benefit), it makes sense that if there is insufficient benefit to detect then waiting for longer will not change anything with respect to test power.

On the other hand, for **Test B1** ($m = 1000$), we observe from the bottom row, right column of Table 5.9 that there are a few cases where we attain a valid value of $t_{80\%}$, but inspecting Figure 5.18a shows that there is at least 80% rejection only in specific time intervals. This result makes sense since we can see from Figure 5.15a that $g_{B1}(\boldsymbol{\theta})$ (as a function of time) has a global maximum for some t close to 1. Hence, maximum benefit is detected only close to $t = 1$ in this case.

Despite these observations, we note that we are estimating the survival functions of the holding time in state 3 correctly, on average. This can be seen from Figure 5.17. The reasons for this are as discussed in the beginning of this section, during the discussion of the results associated with **Test A**.

See Section 5.5 for copies of Tables 5.8 – 5.10, displayed alongside tables associated with the other model setups discussed in this chapter.

Type I error ($m = 10000$)				Power ($m = 10000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.013	0.047	0.108	$t_{80\%}$ (Weibull)	0.54	0.52	0.50
ε (Exponential)	0.001	0.005	0.015	$t_{80\%}$ (Exponential)	0.55	0.53	0.51
ε (Gamma)	0.002	0.015	0.037	$t_{80\%}$ (Gamma)	0.56	0.53	0.52

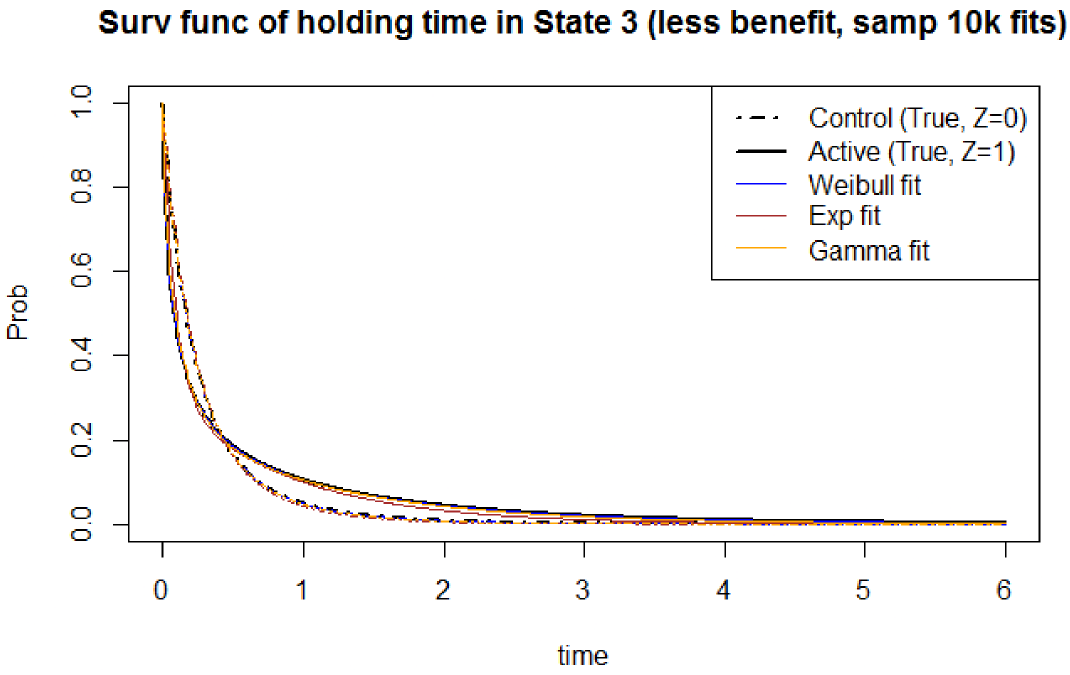
Type I error ($m = 1000$)				Power ($m = 1000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.012	0.051	0.109	$t_{80\%}$ (Weibull)	NA	1.09	0.78
ε (Exponential)	0.009	0.031	0.062	$t_{80\%}$ (Exponential)	0.89	0.74	0.68
ε (Gamma)	0.006	0.034	0.087	ε (Gamma)	NA	0.99	0.80

Table 5.9: (Less benefit with censoring, **Test B1**) Table summarising results of empirical type I error rates and power. The first column shows results for estimated type I errors, ε, under H_0. The test is carried out at $t_0 = 0.425$, which leads to $g_{B1}(\boldsymbol{\theta}) \approx 0$. The second column shows the value of $t_{80\%}$, which is defined as the smallest value of t which gives us estimated power of at least 0.8. The closer $t_{80\%}$ is to $t_0 = 0.425$, the more powerful the test is. A value of "NA" indicates that it was not possible to obtain at least 80% rejection of H_0 in the 1000 studies. Each of the tests are performed using parameters associated with: (*i*) the true Weibull model fit, (*ii*) an exponential model fit, and (*iii*) a gamma model fit. The sample size is either $m = 10000$ or $m = 1000$.

Type I error ($m = 10000$)				Power ($m = 10000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.013	0.047	0.102	$t_{80\%}$ (Weibull)	1.22	1.08	1.03
ε (Exponential)	0.001	0.008	0.018	$t_{80\%}$ (Exponential)	1.31	1.17	1.11
ε (Gamma)	0.002	0.020	0.044	$t_{80\%}$ (Gamma)	1.26	1.12	1.06

Type I error ($m = 1000$)				Power ($m = 1000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.014	0.049	0.106	$t_{80\%}$ (Weibull)	NA	NA	NA
$t_{80\%}$ (Exponential)	0.008	0.029	0.067	$t_{80\%}$ (Exponential)	NA	NA	NA
ε (Gamma)	0.007	0.037	0.087	ε (Gamma)	NA	NA	NA

Table 5.10: (Less benefit with censoring, **Test B2**) Table summarising results of empirical type I error rates and power. The first column shows results for estimated type I errors, ε, under H_0. The test is carried out at $t^* = 0.764$, which leads to $g_{B2}(\boldsymbol{\theta}) \approx 0$. The second column shows the value of $t_{80\%}$, which is defined as the smallest value of t_1 which gives us estimated power of at least 0.8. The closer $t_{80\%}$ is to $t^* = 0.764$, the more powerful the test is. A value of "NA" indicates that it was not possible to obtain at least 80% rejection of H_0 in the 1000 studies. Each of the tests are performed using parameters associated with: (*i*) the true Weibull model fit, (*ii*) an exponential model fit, and (*iii*) a gamma model fit. The sample size is either $m = 10000$ or $m = 1000$.



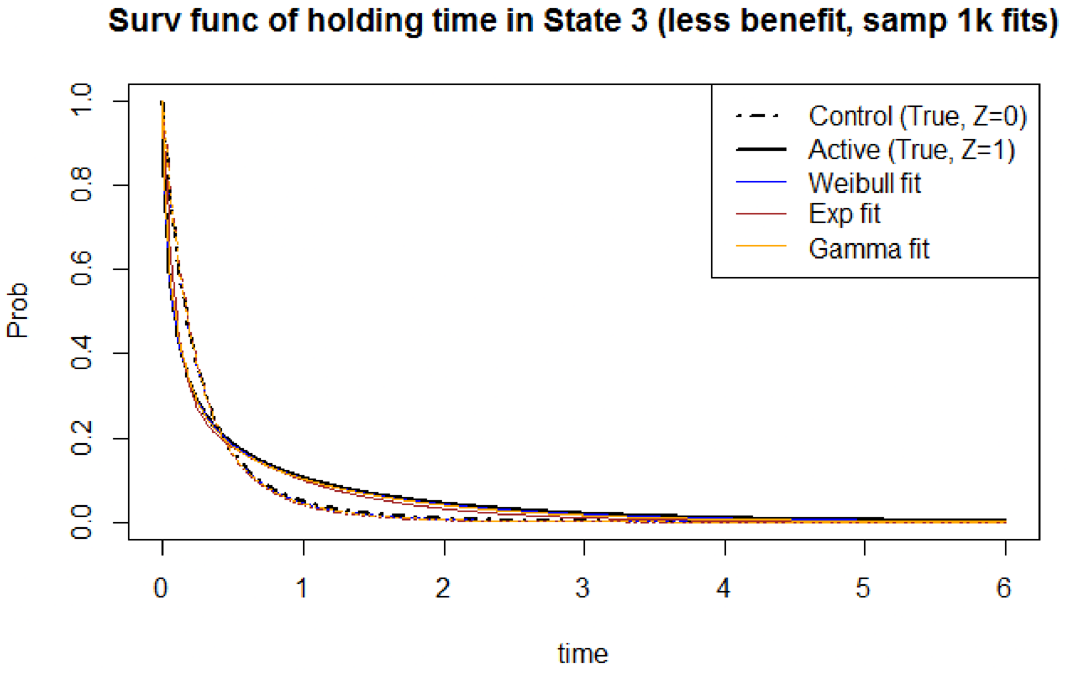
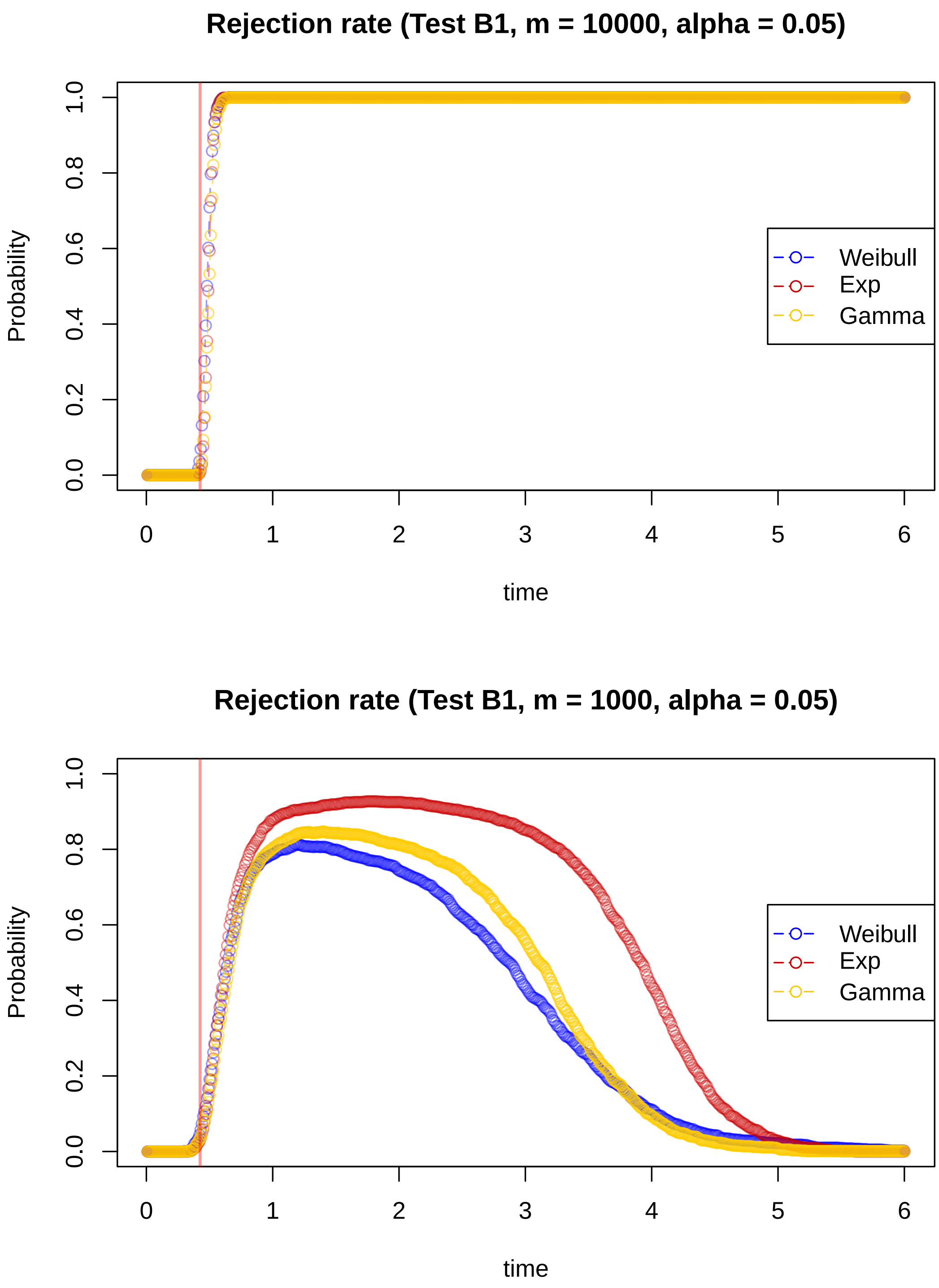
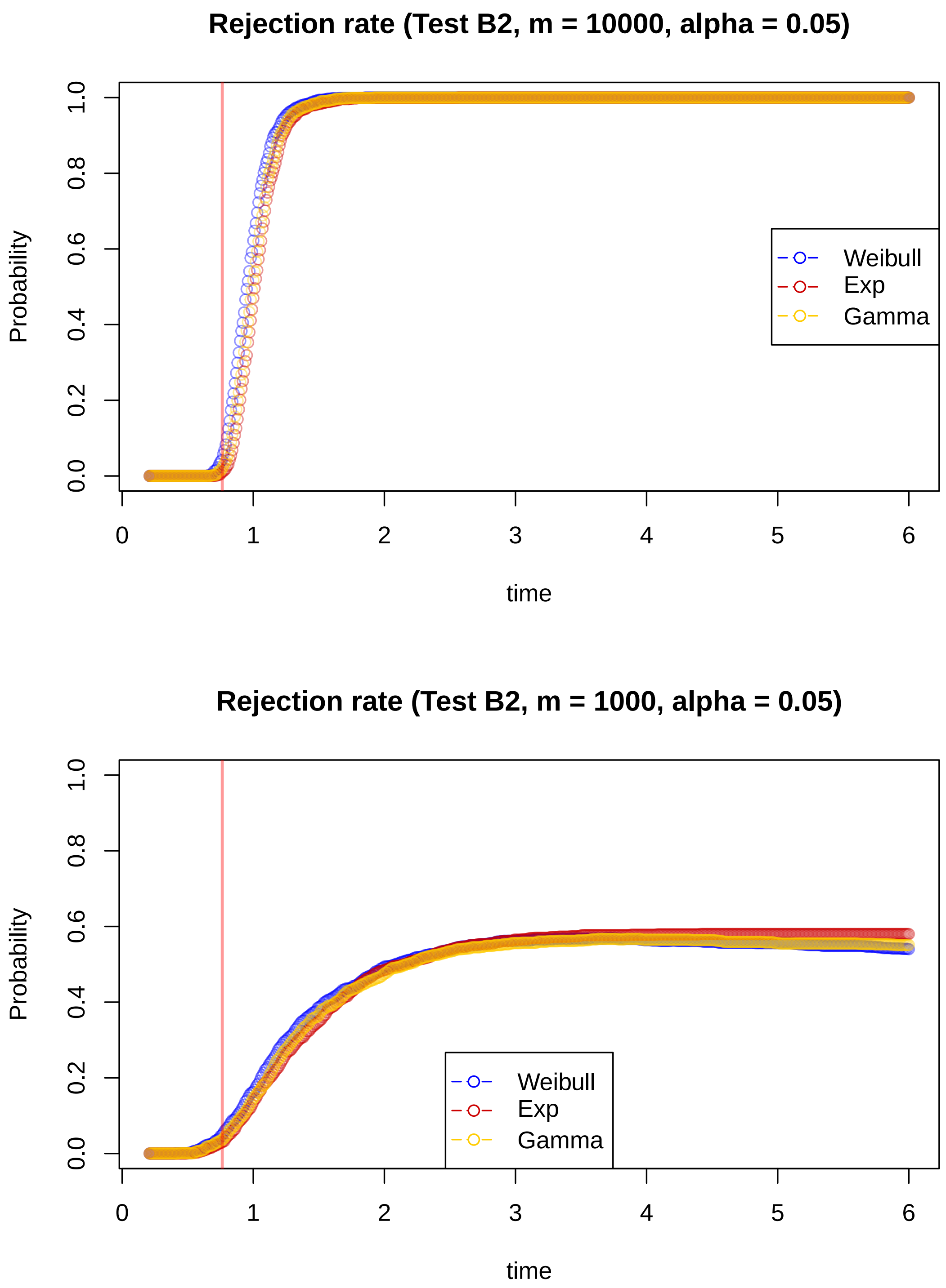


Figure 5.17: (Less benefit with censoring) Parametric model fits: the survival function of holding time for each treatment arm in state 3, for each of $m = 10000$ and $m = 1000$. The thick black lines are the true survival functions while the blue, brown, and orange lines are that for the average of $M = 1000$ fits associated with the Weibull, exponential, and gamma models respectively.



(a) Estimates of power of **Test B1** as a function of time.

Figure 5.18: (Less benefit and with censoring) Estimates of power of **Test B** as described in Section 5.1.3



(b) Estimates of power of **Test B2** as a function of time.

Figure 5.18: (Less benefit and with censoring) Estimates of power of **Test B** as described in Section 5.1.3

5.5 Copies of Test A and Test B summary tables

Type I error ($m = 10000$)				Power ($m = 10000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.007	0.037	0.083	ρ (Weibull)	1	1	1
ε (Exp.)	0.009	0.049	0.098	ρ (Exp.)	0.989	0.996	0.997
ε (Gamma)	0.007	0.046	0.093	ρ (Gamma)	0.998	0.998	0.998

Type I error ($m = 1000$)				Power ($m = 1000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.007	0.033	0.080	ρ (Weibull)	1	1	1
ε (Exp.)	0.009	0.049	0.098	ρ (Exp.)	0.988	0.993	0.994
ε (Gamma)	0.008	0.039	0.078	ρ (Gamma)	1	1	1

Table 5.11: (Baseline, **Test A**) Copy of Table 5.2

Type I error ($m = 10000$)				Power ($m = 10000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.008	0.044	0.093	ρ (Weibull)	1	1	1
ε (Exp.)	0.008	0.049	0.106	ρ (Exp.)	1	1	1
ε (Gamma)	0.006	0.051	0.086	ρ (Gamma)	1	1	1

Type I error ($m = 1000$)				Power ($m = 1000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.003	0.034	0.081	ρ (Weibull)	0.473	0.876	0.967
ε (Exp.)	0.006	0.048	0.091	ρ (Exp.)	0.818	0.952	0.982
ε (Gamma)	0.003	0.040	0.092	ρ (Gamma)	0.192	0.944	0.994

Table 5.12: (Baseline with censoring, **Test A**) Copy of Table 5.5

Type I error ($m = 10000$)				Power ($m = 10000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.018	0.045	0.094	ρ (Weibull)	0.789	0.942	0.969
ε (Exp.)	0.013	0.051	0.100	ρ (Exp.)	0.374	0.627	0.745
ε (Gamma)	0.013	0.045	0.107	ρ (Gamma)	0.689	0.893	0.954

Type I error ($m = 1000$)				Power ($m = 1000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.002	0.044	0.090	ρ (Weibull)	0.01	0.165	0.333
ε (Exp.)	0.005	0.055	0.109	ρ (Exp.)	0.018	0.136	0.239
ε (Gamma)	0.003	0.045	0.091	ρ (Gamma)	0.010	0.157	0.302

Table 5.13: (Less benefit with censoring, **Test A**) Copy of Table 5.8

Type I error ($m = 10000$)				Power ($m = 10000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.005	0.050	0.090	$t_{80\%}$ (Weibull)	0.32	0.31	0.30
ε (Exp.)	0.997	0.998	1	$t_{80\%}$ (Exp.)	-	-	-
ε (Gamma)	0.998	0.999	0.999	$t_{80\%}$ (Gamma)	-	-	-

Type I error ($m = 1000$)				Power ($m = 1000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.006	0.043	0.099	$t_{80\%}$ (Weibull)	0.42	0.38	0.37
ε (Exp.)	0.997	0.997	0.999	$t_{80\%}$ (Exp.)	-	-	-
ε (Gamma)	0.828	0.943	0.978	$t_{80\%}$ (Gamma)	-	-	-

Table 5.14: (Baseline, **Test B1**) Copy of Table 5.3

Type I error ($m = 10000$)				Power ($m = 10000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.004	0.038	0.094	$t_{80\%}$ (Weibull)	0.32	0.31	0.31
ε (Exp.)	0.004	0.024	0.062	$t_{80\%}$ (Exp.)	0.33	0.32	0.31
ε (Gamma)	0.016	0.081	0.169	$t_{80\%}$ (Gamma)	0.32	0.31	0.30

Type I error ($m = 1000$)				Power ($m = 1000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.012	0.047	0.101	$t_{80\%}$ (Weibull)	0.46	0.41	0.39
ε (Exp.)	0.002	0.055	0.100	$t_{80\%}$ (Exp.)	0.43	0.38	0.36
ε (Gamma)	0.021	0.064	0.123	$t_{80\%}$ (Gamma)	0.46	0.41	0.38

Table 5.15: (Baseline with censoring, **Test B1**) Copy of Table 5.6

Type I error ($m = 10000$)				Power ($m = 10000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.004	0.038	0.094	$t_{80\%}$ (Weibull)	0.32	0.31	0.31
ε (Exp.)	0.004	0.024	0.062	$t_{80\%}$ (Exp.)	0.33	0.32	0.31
ε (Gamma)	0.016	0.081	0.169	$t_{80\%}$ (Gamma)	0.32	0.31	0.30

Type I error ($m = 1000$)				Power ($m = 1000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.012	0.047	0.101	$t_{80\%}$ (Weibull)	0.46	0.41	0.39
ε (Exp.)	0.002	0.055	0.100	$t_{80\%}$ (Exp.)	0.43	0.38	0.36
ε (Gamma)	0.021	0.064	0.123	$t_{80\%}$ (Gamma)	0.46	0.41	0.38

Table 5.16: (Less benefit with censoring, **Test B1**) Copy of Table 5.9

Type I error ($m = 10000$)				Power ($m = 10000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.005	0.050	0.091	$t_{80\%}$ (Weibull)	0.45	0.43	0.42
ε (Exp.)	0.997	0.998	1	$t_{80\%}$ (Exp.)	-	-	-
ε (Gamma)	0.998	0.999	0.999	$t_{80\%}$ (Gamma)	-	-	-

Type I error ($m = 1000$)				Power ($m = 1000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.006	0.043	0.101	$t_{80\%}$ (Weibull)	0.72	0.62	0.57
$t_{80\%}$ (Exp.)	0.997	0.997	0.999	$t_{80\%}$ (Exp.)	-	-	-
ε (Gamma)	0.832	0.945	0.977	$t_{80\%}$ (Gamma)	-	-	-

Table 5.17: (Baseline, **Test B2**) Copy of Table 5.4

Type I error ($m = 10000$)				Power ($m = 10000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.004	0.038	0.094	$t_{80\%}$ (Weibull)	0.46	0.44	0.43
ε (Exp.)	0.005	0.034	0.079	$t_{80\%}$ (Exp.)	0.48	0.45	0.44
ε (Gamma)	0.017	0.082	0.173	$t_{80\%}$ (Gamma)	0.42	0.43	0.46

Type I error ($m = 1000$)				Power ($m = 1000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.012	0.046	0.094	$t_{80\%}$ (Weibull)	0.83	0.69	0.62
$t_{80\%}$ (Exp.)	0.021	0.059	0.106	$t_{80\%}$ (Exp.)	0.75	0.61	0.55
ε (Gamma)	0.022	0.063	0.126	$t_{80\%}$ (Gamma)	0.84	0.68	0.62

Table 5.18: (Baseline with censoring, **Test B2**) Copy of Table 5.7

Type I error ($m = 10000$)				Power ($m = 10000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.013	0.047	0.102	$t_{80\%}$ (Weibull)	1.22	1.08	1.03
ε (Exp.)	0.001	0.008	0.018	$t_{80\%}$ (Exponential)	1.31	1.17	1.11
ε (Gamma)	0.002	0.020	0.044	$t_{80\%}$ (Gamma)	1.26	1.12	1.06

Type I error ($m = 1000$)				Power ($m = 1000$)			
α	0.01	0.05	0.10	α	0.01	0.05	0.10
ε (Weibull)	0.014	0.049	0.106	$t_{80\%}$ (Weibull)	NA	NA	NA
$t_{80\%}$ (Exp.)	0.008	0.029	0.067	$t_{80\%}$ (Exp.)	NA	NA	NA
ε (Gamma)	0.007	0.037	0.087	ε (Gamma)	NA	NA	NA

Table 5.19: (Less benefit with censoring, **Test B2**) Copy of Table 5.10

5.6 Related quantities of interest

As mentioned in Section 4.3, it is helpful to consider other quantities of interest in assessing patient benefit, as average total sojourn times and survival functions of the holding time in particular states do not give a complete picture. We present results associated with the simulated data and show some potential effects of model misspecification and insufficient sample size. It is worth noting once again that the utility of the methods in this section (or any other method involving such functions of $\boldsymbol{\theta}$) is that one can quantify uncertainty by constructing confidence intervals derived using the delta method.

We might be interested in another state besides state 3 and want to investigate, for example, the expected value of η, where η is defined as the sojourn time given passage through state 2 before being absorbed. We can use equation (4.10) for this and compute the expected values $\mathrm{E}(\eta|Z=0)$ and $\mathrm{E}(\eta|Z=1)$ relatively easily. We are able to, if desired, investigate the quantities separately instead of studying the difference in means (through $g_A(\hat{\boldsymbol{\theta}})$). Table 5.20 shows the results. The expected values estimated from the $M = 1000$ sets of parameter estimates associated with Weibull, gamma, and exponential model fits can be compared with the true values. The estimates reported are the sample means of the $M = 1000$ estimates of $\mathrm{E}(\eta|Z=z)$, denoted $\bar{\eta}|Z=z$ for $z \in \{0,1\}$. The numbers in parentheses are 95% (parametric) bootstrap confidence intervals for the respective expected values, given by the 0.025 and 0.975 sample quantiles (respectively) of the bootstrap sample. The `quantile` function in R is used for this purpose. See [Efron, 1979] and [Efron, 1982] for details about the bootstrap method, and [Efron, 1985] for details about bootstrap confidence intervals.

Table 5.20 shows $\bar{\eta}|Z=z$, for both sample sizes $m = 10000$ and $m = 1000$ for all three models. We observe that the expectation of η is underestimated in the cases where the model is misspecified, with the exponential model performing worst. Furthermore, there is relatively larger bias and variance in the estimates associated with $\mathrm{E}(\eta|Z=1)$. These results are consistent with the results seen in the previous

Sections 5.1.3, 5.3.2, and 5.4.2 when discussing **Test A**. Figure 5.19 below shows the distributions associated with $M = 1000$ estimates of the expected sojourn time given passage through state 2, where one can see a visual representation of the results reported in Table 5.20. The vertical red lines are the respective true values while the blue, orange, and brown lines are estimated densities under the Weibull, gamma, and exponential model fits respectively.

We are also able to visualise the state occupancy probabilities of the baseline model similarly to that shown in the example in Section 4.3.3. Figure 5.20 below shows, for each of $Z = 1$ and $Z = 0$, the visualisations associated with the true state occupancy probabilities. The state occupancy probabilities for both states 4 and 5 are combined for convenience, since they are the only absorbing states. It would be possible to analyse the state occupancy probabilities of each of these states separately, especially if one of the states can be considered more undesirable than the other. See Section 7.1 for a discussion of composite outcomes in clinical trials.

We observe that individuals in the active treatment arm ($Z = 1$) have a relatively lower chance of ending up in any absorbing state by a given time as compared to their peers in control treatment ($Z = 0$). This is consistent with the setup and how the hazard rate of exiting state 3 is reducing with time. We observe that at around $t = 3$, there is still a small but significant chance of being in state 3 for individuals with $Z = 1$ as compared to their peers with $Z = 0$.

$\mathrm{E}(\eta \mid Z=0)$	0.213	$\mathrm{E}(\eta \mid Z=1)$	0.198
$m = 10000$		$m = 10000$	
$\bar{\eta} \mid Z = 0$ (Weibull fit)	0.213 (0.1998, 0.2263)	$\bar{\eta} \mid Z = 1$ (Weibull fit)	0.199 (0.1708, 0.2326)
$\bar{\eta} \mid Z = 0$ (Gamma fit)	0.212 (0.1987, 0.2248)	$\bar{\eta} \mid Z = 1$ (Gamma fit)	0.192 (0.1659, 0.2211)
$\bar{\eta} \mid Z = 0$ (Exponential fit)	0.209 (0.1967, 0.2218)	$\bar{\eta} \mid Z = 1$ (Exponential fit)	0.178 (0.1564, 0.2050)
$m = 1000$		$m = 1000$	
$\bar{\eta} \mid Z = 0$ (Weibull fit)	0.214 (0.1725, 0.2594)	$\bar{\eta} \mid Z = 1$ (Weibull fit)	0.206 (0.1200, 0.3394)
$\bar{\eta} \mid Z = 0$ (Gamma fit)	0.212 (0.1714, 0.2572)	$\bar{\eta} \mid Z = 1$ (Gamma fit)	0.195 (0.1180, 0.2992)
$\bar{\eta} \mid Z = 0$ (Exponential fit)	0.209 (0.1699, 0.2532)	$\bar{\eta} \mid Z = 1$ (Exponential fit)	0.180 (0.1114, 0.2575

Table 5.20: (Baseline) Estimated mean sojourn times after leaving state 2 (given passage through state 2), before absorption, based on the simulated datasets. $\bar{\eta} \mid Z = z$ ($z \in \{0, 1\}$) is a sample mean derived from $M = 1000$ estimates of $\mathrm{E}(\eta \mid Z = z)$ and the numbers in parentheses are 95% bootstrap confidence intervals obtained by taking the 0.025 and 0.975 sample quantiles of the 1000 estimated values of $\mathrm{E}(\eta \mid Z = z)$.

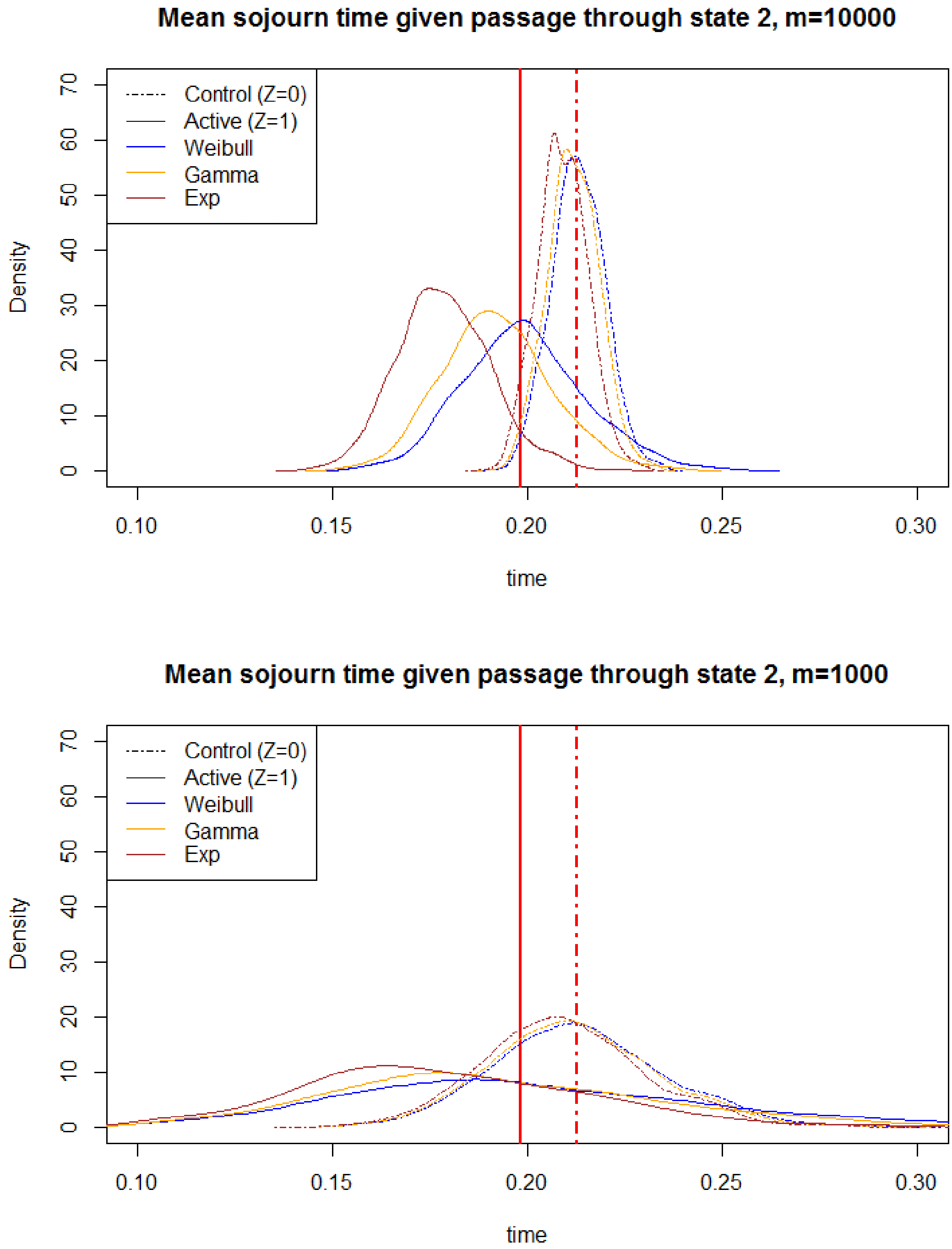


Figure 5.19: (Baseline) The distribution associated with each estimated expected sojourn time given passage through state 2, derived from $M = 1000$ estimates of $E(\eta|Z = z)$. The red vertical lines represent $\mathrm{E}(\eta|Z = 0)$ (dotted line) and $\mathrm{E}(\eta|Z = 1)$ (solid line). The estimated density functions under the true Weibull model fit, gamma model fit, and exponential model fit are in blue, orange, and red respectively.

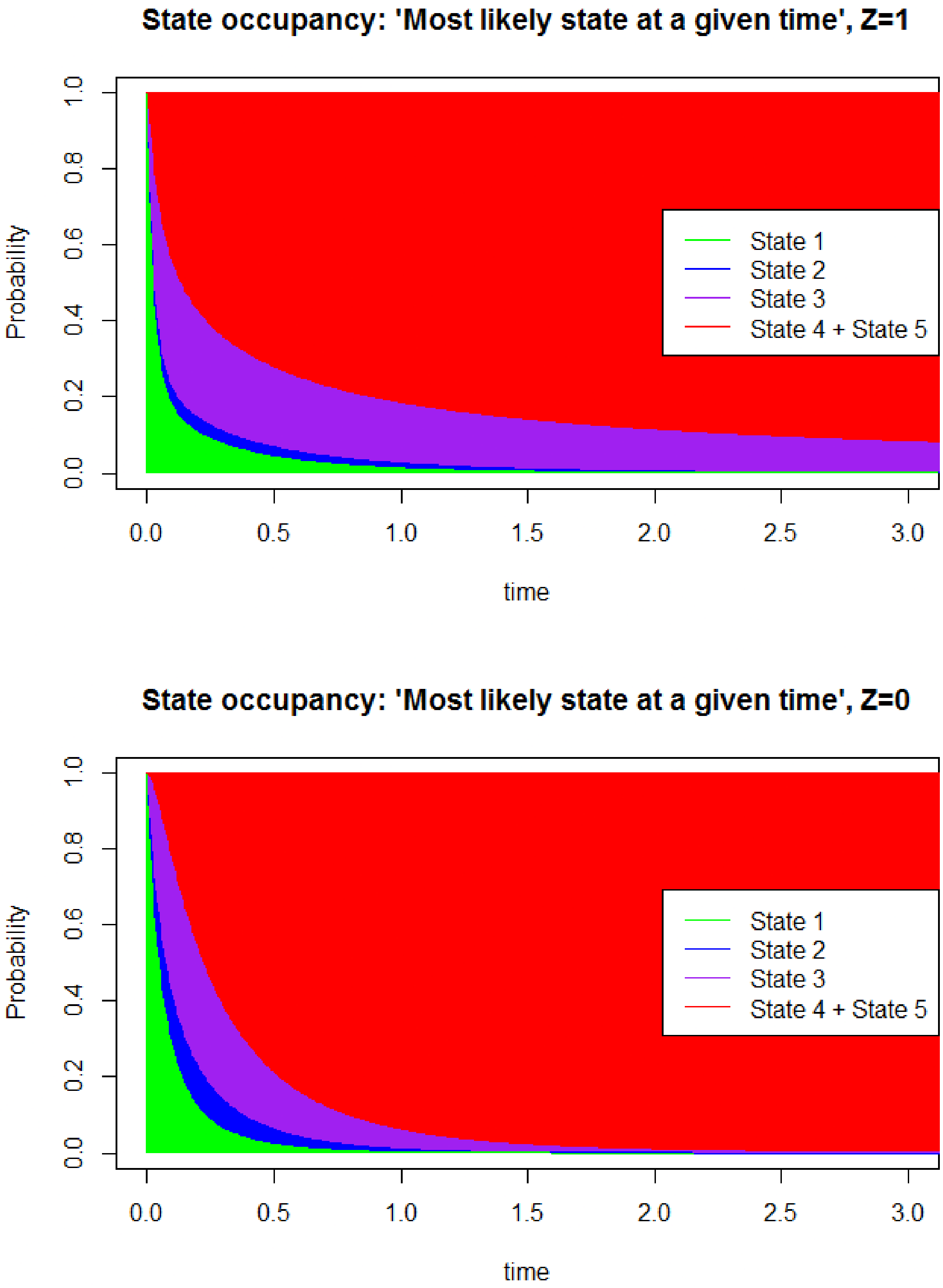


Figure 5.20: (Baseline) Visualising the state occupancy probabilities associated with the true model.

5.7 Further discussion

In this section, we summarise the results of the simulation study in this chapter. Note that discussions of statistical power in this section are in the context of a given value of α (either 0.01, 0.05, or 0.10). The type I error is reasonably controlled for since the empirical probability of type I errors are reasonably close α when the true (Weibull) model is fitted.

Overall, it seems that both **Test A** and **Test B** have different uses, and different advantages and drawbacks. For example, we observe that **Test A** is largely robust to model misspecification, despite the misspecification causing bias and potentially increasing the variance of $g_A(\hat{\boldsymbol{\theta}})$. Whenever there is benefit to detect (in the form of a significant difference in total average sojourn times), the test works reasonably well. Aside from the baseline model with censoring where the model fits for $m = 1000$ were suspect, the only scenario where **Test A** started to show major issues was in the model with less benefit and censoring, where the test had difficulty detecting the relatively small amount of benefit therein.

On the other hand, **Test B** was more problematic with respect to model misspecification. It was not robust in many cases, and did not seem to give reliable results except when the model was correctly specified. Once again, the test could not easily discern whether there was benefit in the model with less detectable benefit, especially when $m = 1000$. **Test B1**, while likely less useful in practice, was relatively better in detecting differences in the survival function at specific time points. **Test B2**, which is more useful due to the ability to choose a time interval of interest rather than a specific time point, was less sensitive to differences in "overall" patient benefit in the interval of interest.

The related quantities of interest as per Section 5.6 also allow us to focus on specific states of interest and to have a clearer picture of patient benefit. For example, we can see from the visualisation of the state occupancy probabilities (Figure 5.20) that patients in active treatment ($Z = 1$) take significantly longer to reach an absorbing state. In an actual clinical trial, this would be evidence of clear benefit to patients since they might be able to remain in active treatment while avoiding undesirable absorbing states (such as "death" or "loss to follow-up") for as long as possible.

As noted in Section 5.3.2, a significant rate of censoring impacts the results very much, even when the sample size is large ($m = 10000$). When $m = 1000$, even the model fit for the correctly-specified model suffered. This is partly due to the fact that states which are more likely to be reached later on may have smaller relative sample size and more individuals right-censored there if the observation period of the clinical study is not long enough. From the results, we can see that proper model specification, large sample sizes, and sufficiently long observation periods all play a role in ensuring reliable results. It is worth stressing that there may be cases where data with sample sizes smaller than $m = 1000$ might show sufficient evidence of patient benefit, provided the benefit is relatively large. We do see some evidence of this in certain states associated with the real dataset discussed in Chapter 6. However, this is difficult to determine *a priori* when planning clinical studies. However, as mentioned at the beginning of this chapter, it is not unsual to find $m \gg 1000$ in oncological clinical trials. Further discussion about these issues can be found in Section 7.1.

Chapter 6

Results: Application to real data

This section shows the methods applied to a real clinical dataset originating in continental Europe. A data sharing agreement has been signed by the clinical owners of the data, but the identity of the clinical study data cannot be shared in this thesis.

We start with a brief description of the data in Section 6.1. Then, analysis of the data is in Section 6.2, where we discuss both the Cox proportional hazards model (as it might be implemented to ascertain relative differences in drug efficacy), and the Fine-Gray proportional hazards model (as it might be implemented to ascertain absolute differences in drug efficacy). We also consider a semi-Markov multi-state model to ascertain patient benefit, by first carrying out some exploratory analysis before performing variations of **Test A** and **Test B**. Section 6.3 discusses the findings.

6.1 Setup

The original purpose of the study associated with this data was to try and determine whether a modified treatment regime (denoted by $Z = 1$) leads to any advantages over a standard treatment regime (denoted by $Z = 0$). This dataset consists of $m = 366$ cancer patients in 2 treatment arms, with equal proportion. There are a total of 114 censored observations, of which 68 are associated with the standard treatment arm and 46 are associated with the modified treatment arm. All patients had solid tumours, which were radiologically verified. The first and last entries into the database were recorded on July 2010 and January 2015 respectively. The median follow-up time was recorded as 21.6 months. The study protocol was reviewed by an institutional review board, and all participants gave informed consent.

In Section 6.2, the Cox and Fine-Gray proportional hazard models are both fitted to the data to provide a reference point. Then, after fitting the semi-Markov multi-state model, we have taken a parametric bootstrap approach in assessing the proposed hypothesis tests and supplementary quantities of interest. In other words, we assume that the estimated model is a good fit and the simulate 1000 datasets from the (assumed) true model. See [Efron, 1979] and [Efron, 1982] for details about the bootstrap method

The aim is to use the estimated distributions of the corresponding test statistics to draw conclusions about patient benefit. Right-censoring is incorporated into the simulated data by first estimating (from the real data) the probability of being censored in a particular state (given the state is reached) and then carrying out a binomial experiment to decide if a given individual is censored after reaching that state. The censoring time is then decided by simulating a censoring time from a Weibull distribution fitted to the (true) censored event times in each state.

The description and results associated with each of **Test A** and **Test B** are in Section 6.2.3.

6.2 Data analysis

6.2.1 Cox proportional hazards model

First, we consider a simple traditional analysis similar to that undertaken in most clinical trials, where we consider if there are differences between the treatment arms with respect to progression-free survival (PFS) and overall survival (OS).

PFS is defined as the time taken until cancer progression, and "death" times are treated as right-censored times with respect to "progression" times as per standard competing risks methodology. OS is defined as the time taken until patient death. Note that there is no competing risk when considering OS since every patient will eventually die.

A Cox proportional hazards model of the form $h_{\mathrm{P}}(t|Z) = h_{\mathrm{P},0}(t)\exp(\beta_{\mathrm{P}} Z)$ is fitted for PFS. For OS, a similar model of the form $h_{\mathrm{O}}(t|Z) = h_{\mathrm{O},0}(t)\exp(\beta_{\mathrm{O}} Z)$ is fitted. The results are obtained using the `survival` package [Therneau, 2023] in R, and shown in Table 6.1 below. The results suggest that there is no difference between modified and standard treatment regimes when it comes to PFS, though it seems that modified treatment adversely affects overall survival. Figure 6.1 below shows Kaplan-Meier plots to visualise the differences between treatment arms.

	Estimate	Hazard ratio
β_{P}	0.1317 (0.349)	1.1407 (0.866, 1.503)
β_{O}	0.2860 (0.024)	1.3311 (1.038, 1.706)

Table 6.1: (Real data) Estimates and hazard ratios for the Cox PH models specified for the data. Subscripts P and O denote progression-free and overall survival respectively. The numbers in parentheses for the estimates are p-values while the numbers in parentheses for the hazard ratios are 95% asymptotic confidence intervals.

6.2.2 Fine-Gray proportional hazards model

We could also do a similar analysis using the Fine-Gray proportional hazards model. Now, we have $h^S_{\mathrm{P}}(t|Z) = h^S_{\mathrm{P},0}(t)\exp(\gamma_{\mathrm{P}} Z)$ and $h^S_{\mathrm{O}}(t|Z) = h^S_{\mathrm{O},0}(t)\exp(\gamma_{\mathrm{O}} Z)$ for PFS and OS, respectively. The superscript S denotes the subdistribution hazard function. The `cmprsk` package [Gray, 2022] is used in R to fit the models.

Table 6.2 shows the results. In this case, the overall conclusions are similar. The results suggest that there is no difference between modified and standard treatment when PFS is concerned, though it seems that modified treatment adversely affects overall survival. Note that the results for OS are almost identical whether we use the Cox model or Fine-Gray model. This is because there is only one possible transition (towards death) when considering overall survival, and therefore the Cox and Fine-Gray models are equivalent *i.e.* $\beta_O = \gamma_O$. The equivalence of the CS and SD hazard functions when there are no competing risks can also be seen by using equation (2.21) and setting $K = 1$.

	Estimate	Hazard ratio
γ_{P}	0.0611 (0.660)	1.0630 (0.809, 1.400)
γ_{O}	0.2860 (0.024)	1.3311 (1.040, 1.710)

Table 6.2: (Real data) Estimates and hazard ratios for the Fine-Gray PH models specified for the data. Subscripts P and O denote progression-free and overall survival respectively. The numbers in parentheses for the estimates are p-values while the numbers in parentheses for the hazard ratios are 95% asymptotic confidence intervals.

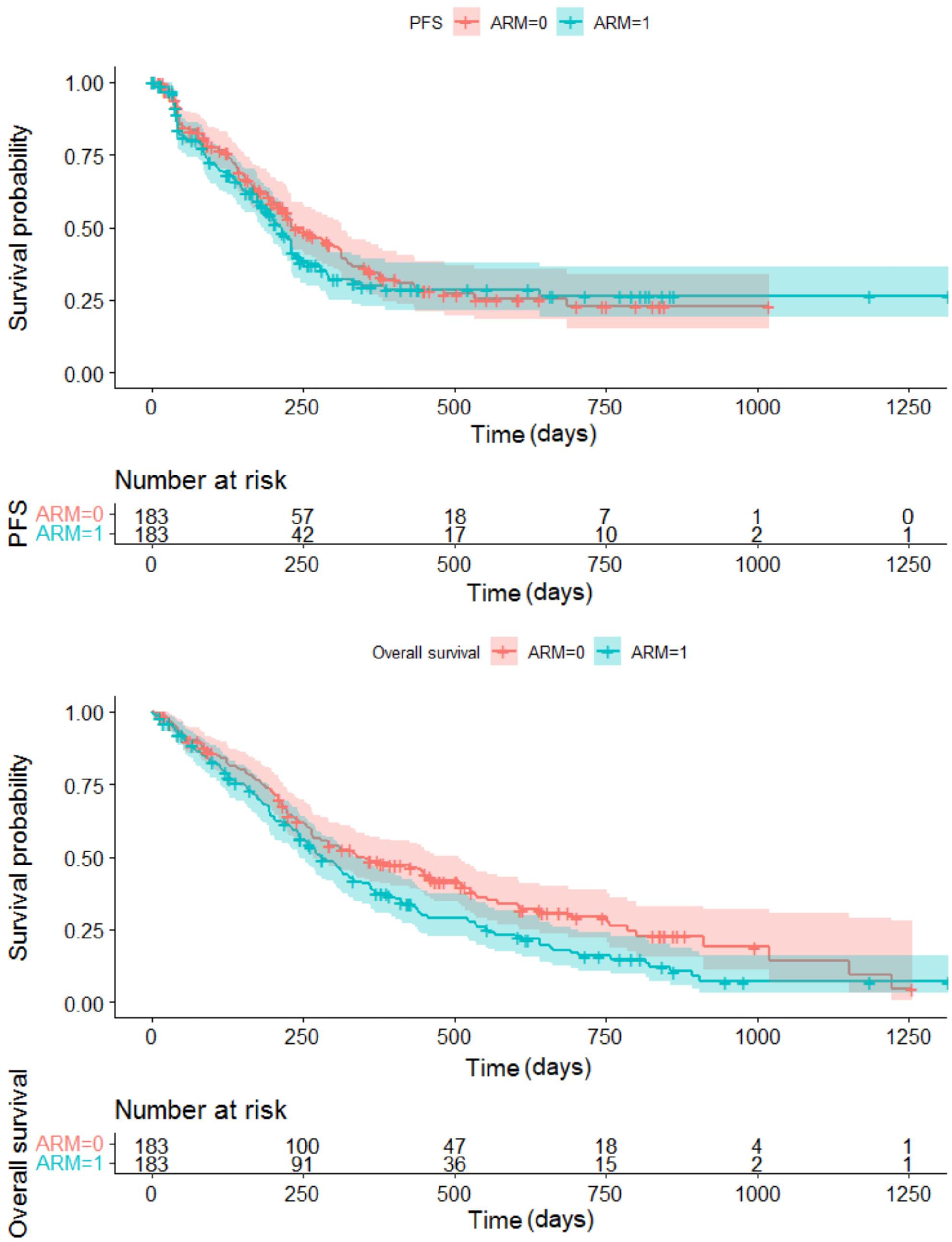
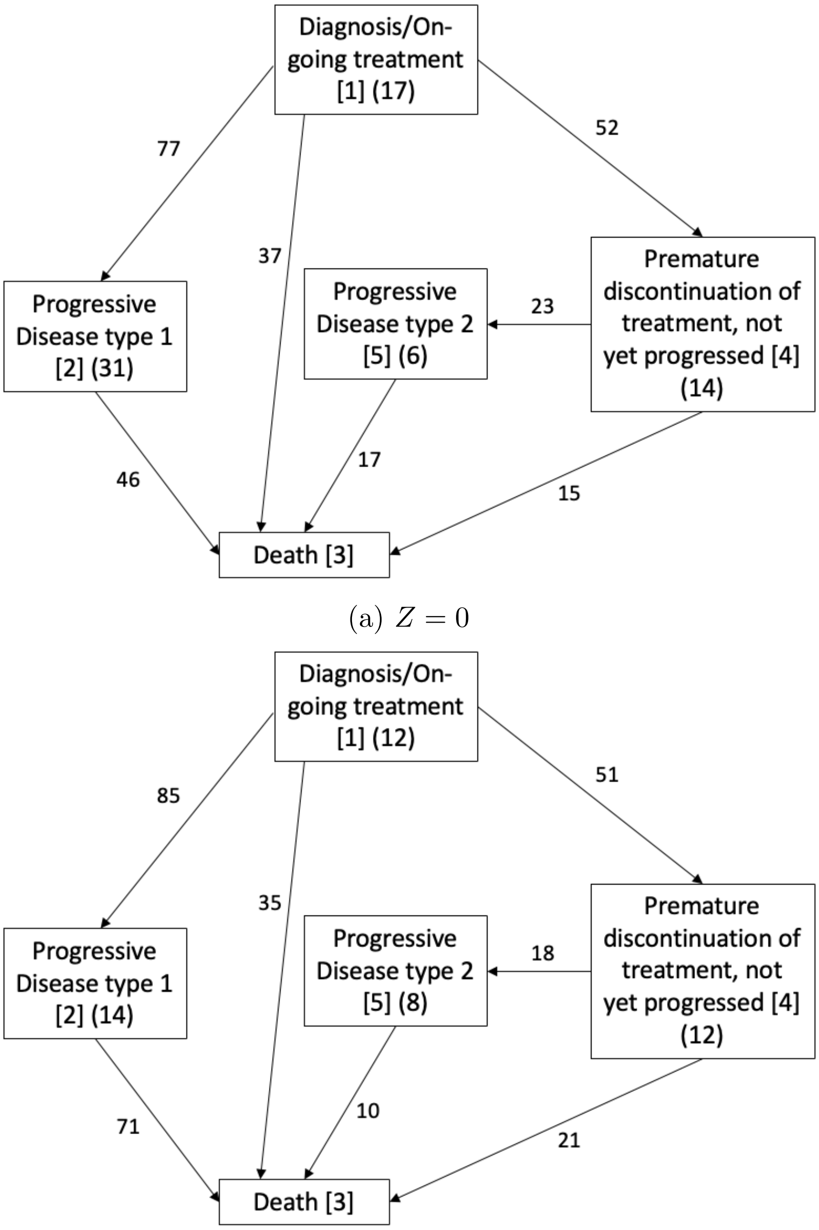


Figure 6.1: (Real data) Progression-free survival function and overall survival function for each treatment arm.

6.2.3 Semi-Markov multi-state model

On the other hand, we have sufficient information to build a semi-Markov multi-state model which consists of 5 states. We have $\mathcal{S} = \{1, 2, 3, 4, 5\}$, where each state in $\mathcal{S}$ respectively describes “diagnosis/ongoing treatment”, “progressive disease type 1”, “death”, “premature discontinuation of treatment”, and “progressive disease type 2”. Patients continue to be followed-up on despite discontinuing treatment prematurely (state 4) and thus it is not an absorbing state. Furthermore, there are 2 types of progression defined. “Progression 1” (state 2) denotes progressive disease that was detected while on active treatment (or within 4 weeks before or after treatment was stopped prematurely), while “Progression 2 (state 5)” denotes progressive disease that was detected at least 4 weeks after a premature discontinuation of treatment. We have $\mathcal{S}^+ = \{1, 2, 4, 5\}$ and $\mathcal{V} = \{(1, 2), (1, 3), (1, 4), (2, 3), (4, 3), (4, 5), (5, 3)\}$. Figure 6.2 depicts this model.

There are two important points to bring up at this juncture. First, we have specified two different types of progressive disease states. One reason for this is because we expect that there may be inherent biological differences between the patients who progress without prematurely discontinuing treatment and the patients who progress some time after having discontinued active treatment prematurely. Specifically, there is the possibility that patients die at different rates depending on which type of progression they experience. Since there is a chance that the underlying transition probabilities and/or transition rates of patients may change depending on their past history, we need to account for this carefully lest we violate the Markov assumption. For example, if we combined both types of progression and treated them as a single state, then we would be asserting (perhaps falsely) that patients who progress all die at the same rate regardless of whether they finished their full course of treatment or not. The second point to note is that, in an ideal scenario, there should be an extra state in the model to account for the patients who are lost to follow-up. The reason is that not accounting for such patients (and instead treating them as right-censored as is usually done in


Diagnosis/On-going treatment [1] (17)
77
37
52
Progressive Disease type 1 [2] (31)
Progressive Disease type 2 [5] (6)
23
Premature discontinuation of treatment, not yet progressed [4] (14)
46
17
15
Death [3]
Diagnosis/On-going treatment [1] (12)
85
35
51
Progressive Disease type 1 [2] (14)
Progressive Disease type 2 [5] (8)
18
Premature discontinuation of treatment, not yet progressed [4] (12)
71
10
21
Death [3]

(a) $Z = 0$

(b) $Z = 1$

Figure 6.2: (Real data) 5-state model with six possible transitions depicting the real dataset ($m = 366$). The numbers in square brackets are the states as per $\mathcal{S}$. The numbers alongside the arrows depict the number of observed transitions while the numbers in the round brackets are the number of right-censored observations.

such studies) would cause a violation of the non-informative censoring assumption which is important in such multi-state models. In this case, the dataset did not distinguish between the right-censored individuals from the ones who were lost to follow-up. The assumption made is that there is a relatively low number of patients who are lost to follow-up. See Section 7.3 and Section 7.4 respectively for more discussion on these points.

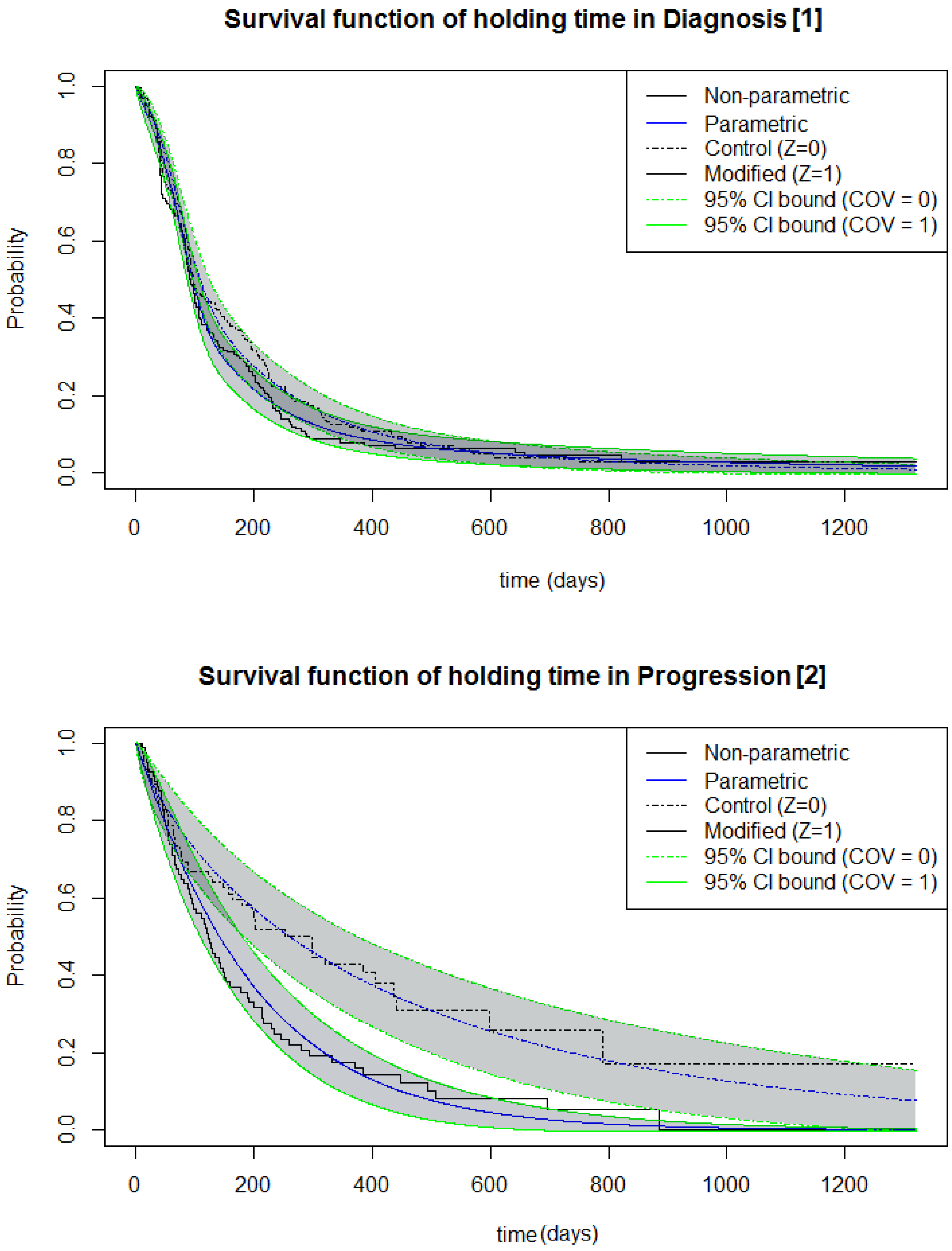
The data are stratified by treatment arm, and Weibull sojourn time hazard functions are chosen for each transition $i \to j$ in a given sub-sample. Specifically, we have

$$h_{ij}(t|Z = z) = b_{z,ij} a_{z,ij} \big(a_{z,ij} t\big)^{b_{z,ij}-1} \tag{6.1}$$

for each of $z = 0, 1$. Here, $a_{z,ij}$ and $b_{z,ij}$ respectively denote the rate and shape parameter for given z and transition $i \to j$.

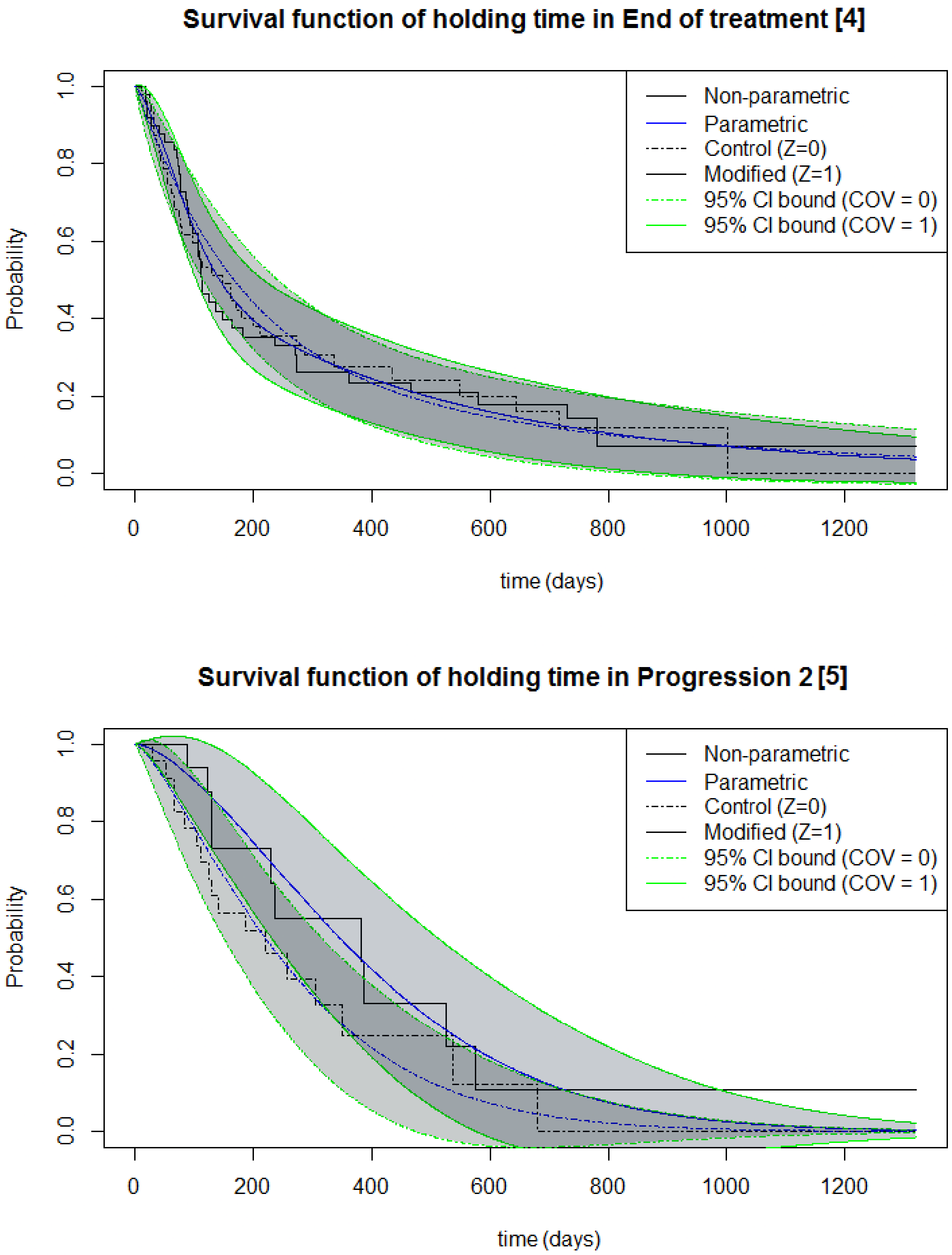
Exploratory analysis

Figure 6.3 depicts the estimated survival functions of holding times in each of the non-absorbing states. The dotted lines denote $Z = 0$ while the solid lines denote $Z = 1$. The blue lines are the estimated function values based on the parametric model fits, while the black stepped lines are estimated function values based on the Kaplan-Meier estimator. The light green lines are confidence intervals derived using the delta method.



(a) Survival functions of the holding times in each of the non-absorbing states 1 and 2.

Figure 6.3: Survival functions of the holding times in each of the non-absorbing states.



(b) Survival functions of the holding times in each of the non-absorbing states 4 and 5.

Figure 6.3: (Real data) Survival functions of the holding times in each of the non-absorbing states.

We note that there does not appear to be any major differences between the treatment arms when we consider the survival functions of the holding time in states 1 (ongoing treatment) or 4 (premature discontinuation). However, there seem to be differences between both treatment arms after patients arrive in either of the progression states (states 2 or 5). Looking at the patients who progressed without any premature discontinuation of treatment (state 2), it seems that patients undergoing modified treatment ($Z = 1$) left state 2 at a faster rate than patients in the standard treatment ($Z = 0$) arm. On the other hand, the situation is potentially reversed for patients who progressed some time after having prematurely discontinued treatment (state 5). Since there is only one exit out of each of the progression states (towards death (state 3)), we might claim the possibility of patients having benefit from the modified treatment regime only if they did not fully complete the treatment. However, the conclusion is less certain since there is some overlap in the confidence intervals associated with state 5.

It is unclear whether the treatment is beneficial for patients in state 1, since there do not seem to be appreciable differences between the survival functions in both treatment arms.

Without doing any formal analysis, we look more closely at patients in states 1 and 4, and we show cumulative incidence functions in Figure 6.4 below to determine what happens to patients who reach these states. With respect to progression, it seems that there might be some minor increase in rate of progression if one is in the modified treatment arm, but it is not clear just by inspecting the figure. On the other hand, after reaching state 4 (premature discontinuation), it seems that that there is a significant-looking decrease in the rate of progression but a significant-looking increase in the rate of death if one is in the modified treatment arm. These estimated probabilities are conditional on reaching state 4, so other quantities of interest may be more appropriate for making inference about patients in general.

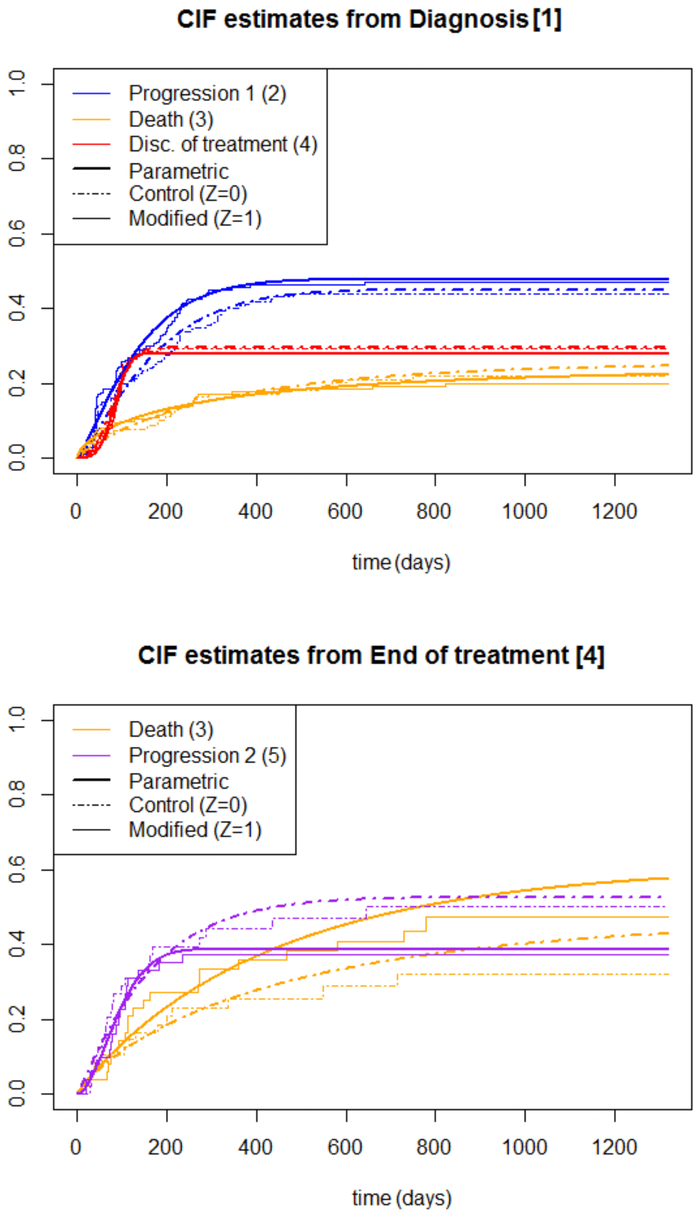


Figure 6.4: (Real data) CIFs associated with transitions out of state 1 (above) and state 4 (below)

Test A

Since we expect that there are differences between patients who experience progressive disease before fully completing treatment as compared to patients who progress some time after completing treatment, we carry out two variations of **Test A** to ascertain whether there might be differences in benefit for each type of patient.

First, define η_1 as the total sojourn time after leaving the "Progression 1 (state 2)" state, conditional on passage through the "Progression 1 (state 2)" state before reaching "Death (state 3)". Using the notation introduced in Section 4.3.1, we have $i = 2$ and $j = 3$. There is only one sub-path, $r(2,3) = \{(2,3)\}$ which satisfies the condition, and the conditional probability of this path is one. The associated total sojourn time is τ_{23}. Hence, the quantity of interest is

$$g_{A1}(\boldsymbol{\theta}) = \mathrm{E}(\tau_{23}|Z = 1) - \mathrm{E}(\tau_{23}|Z = 0) \tag{6.2}$$

with corresponding test statistic

$$T_{A1}(\hat{\boldsymbol{\theta}}) = \frac{g_{A1}(\hat{\boldsymbol{\theta}})}{V_{A1}}, \tag{6.3}$$

where $V_{A1} = \sqrt{\{\nabla g_{A1}(\hat{\boldsymbol{\theta}})\}^\top \mathbf{I}^{-1}(\hat{\boldsymbol{\theta}})\{\nabla g_{A1}\hat{\boldsymbol{\theta}})\}/m}$.

Similarly, define η_2 as the the total sojourn time after leaving the "Progression 2 (state 5)" state, conditional on passage through the "Progression 2 (state 5)" state before reaching "Death (state 3)". Now, $i = 5$ and $j = 3$. Once again, there is only one sub-path which satisfies the given condition, and so the conditional probability of the path is one. The required sub-path is $r(5,3) = \{(5,3)\}$ and the associated total sojourn time is τ_{53}. Hence, the quantity of interest is

$$g_{A2}(\boldsymbol{\theta}) = \mathrm{E}(\tau_{53}|Z = 1) - \mathrm{E}(\tau_{53}|Z = 0). \tag{6.4}$$

with corresponding test statistic

$$T_{A2}(\hat{\boldsymbol{\theta}}) = \frac{g_{A2}(\hat{\boldsymbol{\theta}})}{V_{A2}}, \tag{6.5}$$

where $V_{A2} = \sqrt{\{\nabla g_{A2}(\hat{\boldsymbol{\theta}})\}^{\top}\mathbf{I}^{-1}(\hat{\boldsymbol{\theta}})\{\nabla g_{A2}\hat{\boldsymbol{\theta}})\}/m}$.

Both of these test statistics have an approximate standard Gaussian distribution under H_0.

As described in Section 6.1, we have simulated $M = 1000$ datasets, each with bootstrap sample size $m = 366$. The estimated parameters of the Weibull model fitted on the real data is assumed to be that of the true model, and right-censoring is also incorporated. Table 6.3 shows the mean estimate of each quantity $g_{A1}(\boldsymbol{\theta})$ and $g_{A2}(\boldsymbol{\theta})$, as well as 95% bootstrap confidence intervals (see [Efron, 1985]) derived by computing the 0.025 and 0.975 sample quantile from each sample. The endpoints of the bootstrap confidence intervals are obtained using the `quantile` function in R.

Figure 6.5 shows estimated density functions associated with the distributions of $g_{A1}(\hat{\boldsymbol{\theta}})$ and $g_{A2}(\hat{\boldsymbol{\theta}})$, obtained using the estimates of $g_{A1}(\boldsymbol{\theta})$ and $g_{A2}(\boldsymbol{\theta})$ derived from $M = 1000$ bootstrap sample datasets.

	Mean estimate	Empirical 95% conf. interval
$g_{A1}(\boldsymbol{\theta})$	-280.0848	(-482.6259, -74.0890)
$g_{A2}(\boldsymbol{\theta})$	86.8428	(-155.0721, 311.8801)

Table 6.3: (Real data) Average of $M = 1000$ estimates of $g_{A1}(\boldsymbol{\theta})$ and $g_{A2}(\boldsymbol{\theta})$. The numbers in parentheses are the 95% bootstrap confidence intervals derived by computing the 0.025 and 0.975 sample quantile from each bootstrap sample. We observe that $g_{A1}(\boldsymbol{\theta})$ appears to be significantly less than zero, due to 0 not being in the bootstrap confidence interval.

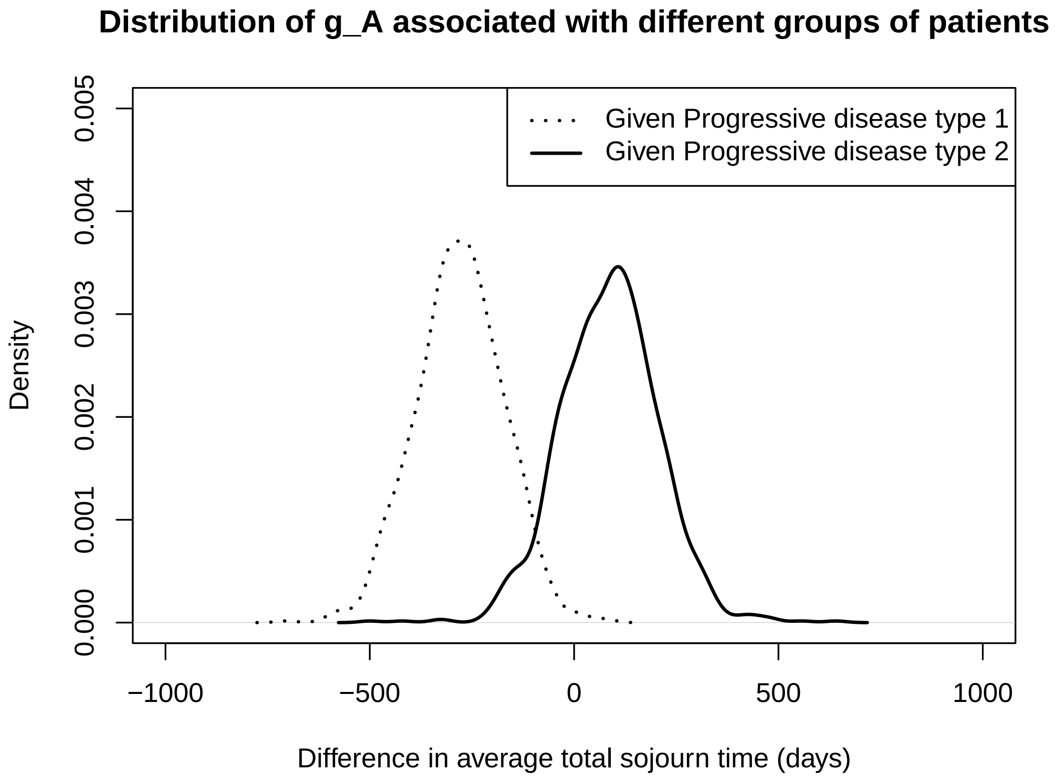


Figure 6.5: (Real data) Estimated density functions associated with distributions of $g_{A1}(\hat{\boldsymbol{\theta}})$ and $g_{A2}(\hat{\boldsymbol{\theta}})$, obtained using the estimates of $g_{A1}(\boldsymbol{\theta})$ and $g_{A2}(\boldsymbol{\theta})$ derived from $M = 1000$ bootstrap datasets. The dotted line is associated with $g_{A1}(\hat{\boldsymbol{\theta}})$ while the solid line is associated with $g_{A2}(\hat{\boldsymbol{\theta}})$.

The results suggest that there is no difference in average sojourn time between treatment arms given that patients progress after premature discontinuation of treatment (state 5) before dying. However, since the 95% bootstrap confidence interval does not contain zero, it seems that there is evidence that, given progression without prematurely discontinuing treatment (state 2) and then dying, patients in the modified treatment arm ($Z = 1$) are dying faster (on average) than patients in the standard treatment arm ($Z = 0$). These results are consistent with that observed during the exploratory analysis of the data. However, it is noted that the results are associated with a relatively large amount of uncertainty as the confidence intervals are relatively wide. This is most likely due to the relatively small sample size associated with the original data.

Test B

Estimated power functions, under the assumption that the estimated Weibull parameters are that of the true model, are derived by computing the proportion of times the null hypothesis is (correctly) rejected for every time point. Using all $M = 1000$ of the simulated datasets, we perform a two-tailed test based on **Test B** for state 2, and a corresponding right-tailed test for state 5. We use equation (4.20) and take $E = (30, 300) \subset \mathbb{R}_{\geq 0}$ and μ as the Lebesgue measure. That is,

$$g_B(\boldsymbol{\theta}) = \int\limits_{30}^{300} \Big(S_i(t|Z=1) - S_i(t|Z=0)\Big)\mathrm{d}t \tag{6.6}$$

for each of $i \in 2, 5$. The corresponding test statistic for each state is

$$T_B(\hat{\boldsymbol{\theta}}) = \frac{g_B(\hat{\boldsymbol{\theta}})}{V_B}, \tag{6.7}$$

where $V_B = \sqrt{\{\nabla g_B(\hat{\boldsymbol{\theta}})\}^\top \mathbf{I}^{-1}(\hat{\boldsymbol{\theta}})\{\nabla g_B \hat{\boldsymbol{\theta}})\}/m}$. Once again, the test statistic has an approximate standard Gaussian distribution under H_0.

In other words, we try to ascertain if there are differences in "average" benefit between treatment arms in the interval beginning 30 days after entering each of these states, until 300 days after entering the same states. The results (discussed below) suggest that, when $\alpha = 0.05$,

(i) there is evidence that benefit between treatment arms in state 2 is different, and

(ii) there is lack of evidence that patients undergoing modified treatment are benefitting more than patients undergoing standard treatment, if they discontinued treatment prematurely.

Figure 6.6 shows the estimated power functions. Similarly to that seen in Chapter 5, the power of the test in each state is estimated by evaluating the proportion of times H_0 is correctly rejected when the test is carried out for each of $M = 1000$ bootstrap datasets. This is done for every time point between 31 days and 300 days (inclusive). The vertical red line denotes 300 days.

We observe that, although the sample size is very small ($m = 366$), in state 2 we reject H_0 over 80% of the time (when it is true) even when the upper limit of the integral associated with the numerator of the test statistic is just below 200 days. On the other hand, we do not attain at least 80% rejection rate at any point when we consider state 5. However, this should not be surprising since the overlap in confidence intervals for the survival function of the holding time in state 5 (Figure 6.3) suggests that there may not be appreciable differences in the survival functions between treatment arms.

6.3 Discussion

The main takeaway is that semi-Markov multi-state models are more flexible and give greater insight as compared to a traditional Cox model analysis. The Cox model analysis suggests that patients undergoing modified treatment die faster, but when using the semi-Markov multi-state model it becomes apparent that the patients who prematurely discontinue treatment potentially die at a different rate from

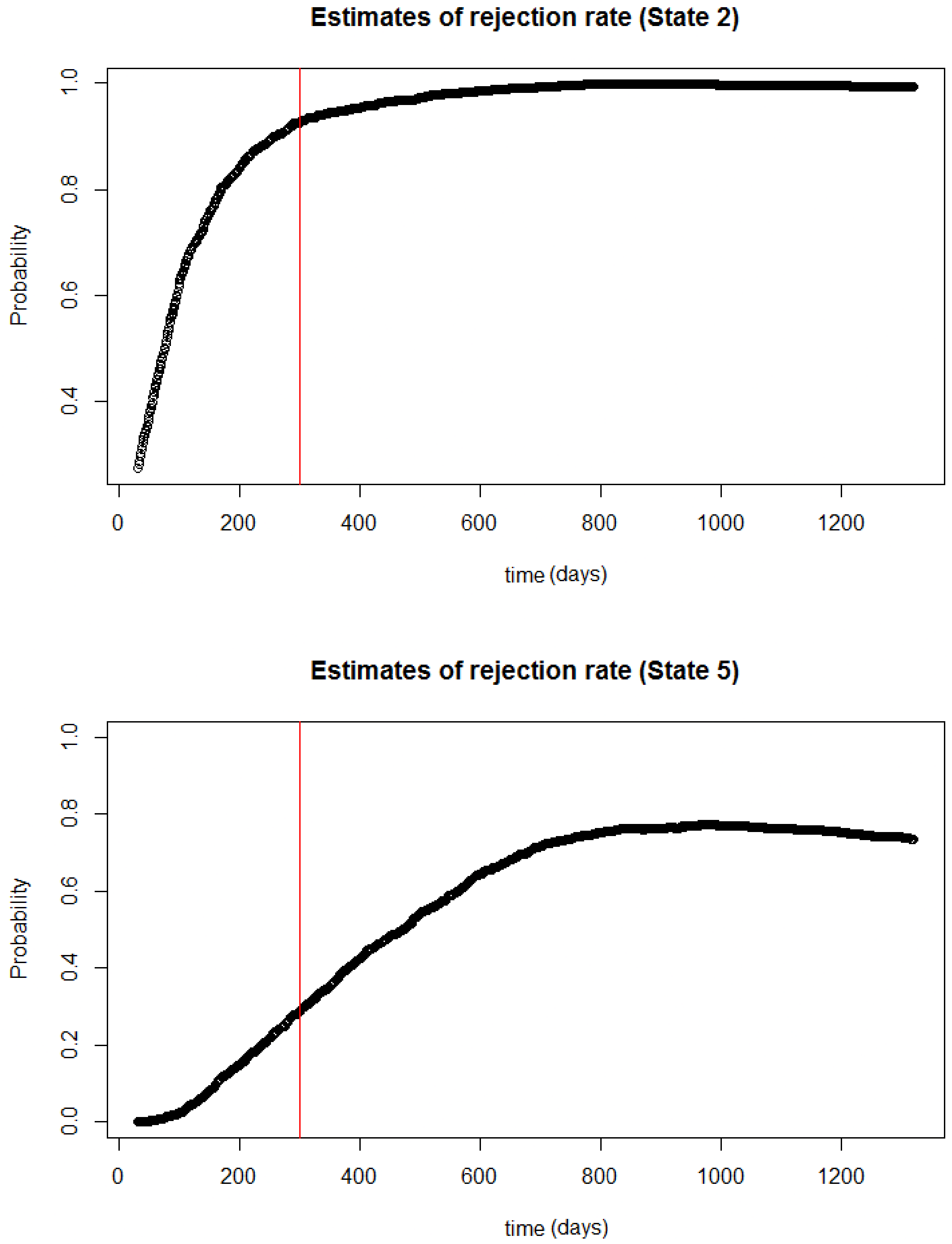


Figure 6.6: (Real data) **Test B**: Estimated power functions showing the rejection rates associated with a two-tailed test for state 2 and a right-tailed test for state 5, respectively. The power of the test in each state is estimated by evaluating the proportion of times H_0 is correctly rejected when the test is carried out for each of $M = 1000$ bootstrap datasets. The quantity of interest, $g_B(\boldsymbol{\theta})$, uses equation (6.6) in the derivation of the test statistic. The curve is plotted starting at 31 days, and the vertical red line denotes 300 days.

patients who do not prematurely discontinue treatment. It is worth acknowledging that we may have obtained similar general conclusions if we augmented the Cox model with information about whether or not a patient discontinued treatment prematurely. However, such an approach would likely have required introducing time-varying covariates to account for the fact that patients might discontinue treatment during the follow-up period. Taking such an approach may not have been obvious nor is it the usual consideration in clinical trials, since the main consideration is efficacy as opposed to overall patient benefit.

In Section 5.7, it was mentioned that smaller sample sizes could give sufficient power when it comes to the proposed hypothesis test as per Section 4.4.2. This is certainly the case when considering the results associated with state 2, since we have detected potential benefit for patients in active treatment despite the relatively small sample size. While we did not find an appreciable difference in patient benefit when considering patients undergoing active treatment in state 5, it is worth noting that the effective sample size in state 5 is significantly less than 366, since only 41 patients reach state 5, of which 14 remain right-censored there. This is compared to 162 patients who reach state 2, of which 45 are right-censored. Hence, we would need more information to draw firm conclusions about patient benefit in state 5.

This example clearly shows why it is important to consider the entire patient history when attempting to evaluate patient benefit, as there is useful information about what is working (or equally, not working) about the treatments.

Chapter 7

Discussion

Throughout this thesis, we have emphasised the importance of considering the entire patient history to ascertain benefit to the patient, as opposed to focusing on the treatment effect of a given treatment. We have proposed semi-Markov multi-state models, as well as statistical procedures, which seek to quantify patient benefit as we have defined it. That being said, our proposed methods should be considered alongside, and not in place of, current methods. This is because models such as the Cox proportional hazards model are still useful for ascertaining drug efficacy, all else being equal.

When considering both hypothesis tests proposed in this thesis, **Test A** seems to be quite robust to factors such as model misspecification and significant amounts of right-censoring (provided sample size is sufficiently large), while **Test B** seems to be less so. This emphasises the need for clinical studies which last sufficiently long (so that there is less right-censoring) and an appropriately specified model for each transition $i \to j$. Having sufficiently large sample size (discussed in more detail in Section 7.1) is also important, especially in cases where the detectable benefit is small. We have also shown in Chapter 6 that the methods can be used on real data, and that we may gain different insights by implementing the methods to explore potential patient benefit.

Unfortunately, multi-state models are often complex since we are interested in making inference about many events of interest (as opposed to just one event of interest as per the Cox model). We saw in Chapter 5 that the five-state model proposed has anywhere between 18 and 30 parameters for each model specification, which can be relatively computationally expensive when it comes to estimation and calculation of other quantities of interest. Furthermore, statistical power can be a concern if the sample size is too small relative to the number of parameters. Models could be simplified greatly by specifying specific forms of hazard functions *e.g.* parametric proportional hazards, or defining composite outcomes. However, these methods have their own associated drawbacks.

There are also other important points of consideration when fitting semi-Markov multi-state models to clinical trial data. We expand upon some of the aforementioned discussion points, and other discussion points, below.

7.1 Clinical trial design considerations

Based on the results shown in Chapter 5, it is apparent that a relatively large sample size is needed to make reliable inference. This is more so for models with less detectable benefit (Section 5.4). For all the models discussed in Chapter 5, we seem to be limited by the transitions in the model which have the least number of observations. For example, in our simulated data example, transition $3 \rightarrow 5$ is rare (relative to the total sample size). In particular, having states with many possible exits or states which can only be reached after experiencing several transitions are examples of complexity which may reduce the effective sample size associated with particular states or transitions.

A method proposed in the literature to preserve effective sample size is to define composite outcomes, which would involve combining one or more states in the multi-state model. The main advantage, as per [Cordoba et al., 2010], is that the use of composite outcomes "increases statistical efficiency because of higher event rates, which reduces sample size requirement, costs, and time;". [Ross, 2007] echoes this as a major advantage, stating that the approach provides an "overwhelming advantage of statistical efficiency, leading to more feasible trial size and increased probability of timely completion". However, the drawbacks of such an approach are also highlighted. For example, an assumption made when defining composite outcomes is that all the individual events used to define it are equally important. [Armstrong and Westerhout, 2017] state "It is clear that such an interpretation is neither realistic nor aligned with the perceptions of clinicians, patients, and other trial stakeholders.". Furthermore, it is difficult to reconcile whenever the individual events within the composite outcome consist of both desirable and undesirable outcomes for the patient. These issues are why we have opted to define all states in the multi-state model as undesirable to the patient, and then consider each state carefully on its own merits (or lack thereof).

A Bayesian approach may also mitigate some of the issues related to limited sample size. For example, [Muehlemann et al., 2023] state "Informative priors can be leveraged when there are relevant data external to the trial being planned. In such cases, Bayesian methods may result in a reduced sample size.". However, there is need for caution when using informative priors since, as per [Ursino and Stallard, 2021], "Choice of a prior distribution must therefore be done carefully, since the use of informative priors may be seen as introducing bias into posterior inferences and inflating type I error rates.". The authors also discuss the additional computational burden that potentially results from choosing a Bayesian approach.

Despite all this, excessively large sample sizes are not crucial if there already exist large differences between treatment arms. For example, we observed that inference was acceptable for the model described in Section 5.2 (baseline model) when $m = 1000$ but the sample size was not remotely large enough for the model as per Section 5.4 (model with less benefit and with censoring) to detect benefit which was present in the active treatment arm. A sample size of $m = 366$ also did reasonably well in at least one state where there were clear differences between treatment arms with respect to the fitted model in Chapter 6, associated with the real dataset.

Section 5.3 highlighted that high rates of censoring in addition to insufficient effective sample size can lead to unreliable model estimates, and that the uncertainty in the estimates propagate to the different quantities of interest used to assess patient benefit. It is for this reason that we have stressed the importance of sufficiently long clinical trials which are not prematurely cut short, in addition to trials with a sufficient number of patients. Note that we are referring to administrative censoring as per [Fine and Gray, 1999], discussed in Section 2.3.3 of this thesis. Discussion about censoring that arises from patients being lost to follow-up is discussed in Section 7.4.

Since effective sample size, model complexity, and rate of right-censoring are all factors which can limit reliable inference, an area for potential further research is to determine more generally simple criteria to determine when inference becomes unreliable in specific models or scenarios. In addition to the scenarios already discussed, we noted that in the example as per Section 5.4, we started running into obvious numerical and estimation issues in optimising the likelihood and obtaining the (numerical) observed Fisher information matrix for certain models when $m = 500$. Hence, bootstrap methods (either parametric or non-parametric) may prove useful to test the effect of different sample sizes on the sensitivity and stability of proposed models.

7.2 Incorporating multiple covariates

The methods presented in this thesis, especially those related to the proposed hypothesis tests, are a "proof of concept" and require extension when considering several covariates, especially if at least one of them is continuous. There are many available models which can allow for the incorporation of multiple covariates, such as the "Cox-like" proportional hazards model shown in Section 2.6. However, many of the commonly-used models make restrictive assumptions (such as the aforementioned proportional hazards assumption) which may not reflect real data. As mentioned earlier in this chapter, this is an issue because the proposed methods are not always robust to poorly specified models.

A proposed solution in [Younes and Lachin, 1997] involves modelling the conditional survival function given covariate $\mathbf{Z}$ as

$$g(S(t|\boldsymbol{Z})) = g(S_0(t)) + \boldsymbol{\beta}^\top \mathbf{Z} \tag{7.1}$$

where g is a link function and $S_0(t) = \exp\left(-\int_0^t h_0(u)\mathrm{d}u\right)$ is a baseline survival function assumed to be independent of the covariates. To avoid estimating the nuisance parameters associated with the baseline survival function, [Younes and Lachin, 1997] use the methods in [Rosenberg, 1995] to estimate the baseline hazard function $h_0(t)$ using *B-splines.* Further work using such "generalised link-based additive modelling" has been done since [Younes and Lachin, 1997], including [Dettoni et al., 2020] and [Eletti et al., 2023] among others.

An advantage of such generalised link-based additive modelling methods is the potential to fit models involving interval-censored data, which are very common in clinical studies. This is because events (such as cancer progression) are actually interval-censored, since the exact event times are not observed but are known to take place in a time interval $(l, r]$ where l and r are usually patient follow-up times. On the other hand, such generalised link-based additive modelling methods have traditionally been used in the ITFs framework and so further work is required to adapt the methods so that the mixture approach can be used instead. It is worth noting that [Eletti et al., 2023] propose an algorithm to estimate the transition probabilities by simulating data from a model that uses ITFs estimated using a generalised link-based additive model.

7.3 Potential violation of the Markov assumption

There might be scenarios where the history of the patient might inform future transitions. We have seen, for example, the case of the dataset in Chapter 6 where we specified two different types of progressive disease states because of the concern that patients who progress after discontinuing treatment prematurely seem to die at a different rate from patients who did not discontinue treatment prematurely. Another scenario where the Markov assumption might be violated in practice is when there are bi-directional transitions or states which can be visited more than once.

For example, if a patient recovers from illness to become healthy and then becomes ill again some time later, he/she may recover at a significantly different rate than the first time he/she contracted the illness. In order to address the potential violation of the Markov assumption, we may consider introducing additional states where appropriate to account for such scenarios. In this specific example, we might remedy the situation by introducing two different types of "Healthy" states - one for patients who have never before contracted the illness, and the other for patients who have previously contracted the illness. Of course, the disadvantage here is the additional complexity. As discussed in Section 7.1, such additional complexity can lead to problems in estimation and inference if the effective sample size associated with each state is not sufficiently large.

To address the violation of the Markov assumption, [Larson and Dinse, 1985, Section 4.3] discuss the possibility of adding the past history as a covariate and then testing formally whether the covariate is significant. If it is not significant, then it might be possible to assert that the Markov assumption is not violated and then proceed without adding extra states. To relate to our earlier "healthy/ill" example, we can consider starting with adding a covariate denoting "has contracted illness in the past". If the covariate parameter is significantly different from zero, then we have evidence that Markov assumption is violated, and we might consider adding another type of "healthy" state in the model as described earlier.

7.4 The assumption of non-informative censoring

It is the case that many studies do not record patients who are censored differently from patients who become lost to follow-up. If we naively treat being lost to follow-up the same as being censored, then it is clear that the censoring process is potentially dependent on the underlying multi-state process since something related to the study itself (which leads to patients withdrawing from the study) could be contributing to the rate of censoring. [Lee and Wolfe, 1998] also acknowledge this, saying "Some study designs are likely to yield independent censoring, . . . Other mechanisms are very likely to yield dependent censoring, *e.g.*, censoring due to subjects selectively dropping out of the study,", and [Ferreira and Patino, 2019] share the different ways such missing data could be classified as *missing completely at random (MCAR)*, *missing at random (MAR)*, or *missing not at random (MNAR)*. Various methods to handle such missing data exist in the literature (see, for example, [Kang, 2013]).

Ideally, one would be able to properly design clinical studies and allocate resources so that patients are never lost to follow-up. Then, we could have "censoring complete" data as described in Section 2.3.3, where patients are right-censored only if they have not yet experienced an event by the end of the observation period. However, this may not be possible in practice. The solution we propose involves treating "lost to follow-up" as a separate state to ensure the censoring process remains non-informative. However, this may have its own issues if, for example, the number of patients lost to follow-up is small (recall transition 3 → 5 in Chapter 5 as a possible comparison, for example). Additionally, it may be difficult to ascertain whether the patients are lost to follow-up because of something potentially related to the multi-state process (*e.g.* excessively toxic drug leading to patient withdrawing from study, or patient's cancer progressing and they decide it is better to withdraw from treatment) or something unrelated (*e.g.* patient deciding to move to a different city early on during the study). It is an area of potential further research to investigate how to best deal with this assumption being violated in practice. It is noted that the work by [Dettoni et al., 2020] implements a generalised link-based additive model (discussed in Section 7.2) that allows for informative censoring.

7.5 Concluding remarks

This thesis has shown that the use of the ubiquitous Cox proportional hazards model (Section 2.3.2) alone is not appropriate for quantifying patient benefit as we have defined it. A more suitable model for quantifying patient benefit is one that answers questions about absolute risk rather than relative risk (the latter of which the Cox model seeks to quantify). While the Fine-Gray proportional hazards model (Section 2.3.3) does quantify absolute risk, it still has certain limitations such as being harder to interpret and its unnatural risk set. Semi-Markov multi-state models address the shortcomings of each of the aforementioned models, and provide the tools necessary to quantify potential patient benefit.

We again emphasise that the current methods (such as the Cox proportional hazards model) are still useful, but only if we are interested in assessing a drug's treatment effect (all else being equal). We believe that the methods we propose could be used alongside, and not necessarily in place of, the current methods used to analyse clinical study data.

While the definition of patient benefit can be subjective, the methods proposed in this thesis provide a natural means to compute probabilities and other statistics which can be considered in the analysis of potential benefit. Once patient benefit is broadly defined, as per Chapter 3.1, to be "an effective treatment that (relatively) slows down patients transitioning to undesirable states", then it can be left to clinicians and other stakeholders to decide how to best quantify this benefit. However, we still propose two different hypothesis tests in Section 4.4 – one based on differences in total average sojourn times between treatment arms (**Test A**), and the other based on differences between the survival functions of holding time in particular states of interest (**Test B**). The latter is based on the fact that the definition of patient benefit entails preventing undesirable events for as long as possible, and so the survival function of the holding time in state i at time t (which is the probability of being event free in state i at time t) becomes a natural quantity to consider.

Overall, we have shown from the results that there are definite merits to using semi-Markov multi-state models to quantify and assess potential patient benefit. A major benefit is the ability to use the estimated model parameters to estimate functions such as the survival function of the holding time, CIFs, or state occupancy probabilities, *etc.*. We also have the flexibility to perform statistical inference and quantify uncertainty on the estimation of any such function using the delta method. However, despite this, we recognise that the methods proposed could be tedious to implement. This is mainly due to the fact that many of the quantities proposed could involve integrals without closed form expressions. Furthermore, the delta method requires partial derivatives with respect to the parameter vector, which can be numerous if there are many parameters in the model. Despite this, it remains the case that quantifying patient benefit can help cancer patients, their doctors, and care-givers to make better-informed decisions about how to best spend the patients' limited remaining lifespan.

References

[Aalen, 1978] Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, pages 701–726.

[Andersen et al., 2012] Andersen, P. K., Geskus, R. B., de Witte, T., and Putter, H. (2012). Competing risks in epidemiology: possibilities and pitfalls. *International Journal of Epidemiology*, 41(3):861–870.

[Andersen and Gill, 1982] Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, pages 1100–1120.

[Anderson and Darling, 1952] Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics*, 23(2):193–212.

[Armstrong and Westerhout, 2017] Armstrong, P. W. and Westerhout, C. M. (2017). Composite end points in clinical research: a time for reappraisal. *Circulation*, 135(23):2299–2307.

[Arnab, 2017] Arnab, R. (2017). *Survey sampling theory and applications.* Academic Press.

[Asanjarani et al., 2021] Asanjarani, A., Liquet, B., and Nazarathy, Y. (2021). Estimation of semi-Markov multi-state models: a comparison of the sojourn times and transition intensities approaches. *The International Journal of Biostatistics.*

[Austin and Fine, 2017] Austin, P. C. and Fine, J. P. (2017). Practical recommendations for reporting Fine-Gray model analyses for competing risk data. *Statistics in Medicine*, 36(27):4391–4400.

[Austin et al., 2021] Austin, P. C., Steyerberg, E. W., and Putter, H. (2021). Fine-Gray subdistribution hazard models to simultaneously estimate the absolute risk of different event types: cumulative total failure probability may exceed 1. *Statistics in Medicine*, 40(19):4200–4212.

[Beyersmann et al., 2009] Beyersmann, J., Latouche, A., Buchholz, A., and Schumacher, M. (2009). Simulating competing risks data in survival analysis. *Statistics in Medicine*, 28(6):956–971.

[Breslow, 1974] Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, pages 89–99.

[Butler and Bronson, 2012] Butler, R. W. and Bronson, D. A. (2012). Bootstrap confidence bands for sojourn distributions in multistate semi-markov models with right censoring. *Biometrika*, 99(4):959–972.

[Castelli et al., 2007] Castelli, C., Combescure, C., Foucher, Y., and Daures, J.-P. (2007). Cost-effectiveness analysis in colorectal cancer using a semi-markov model. *Statistics in Medicine*, 26(30):5557–5571.

[Cordoba et al., 2010] Cordoba, G., Schwartz, L., Woloshin, S., Bae, H., and Gøtzsche, P. C. (2010). Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review. *BMJ*, 341.

[Cox, 1970] Cox, D. (1970). *The analysis of binary data.* London: Chapman and Hall.

[Cox, 1959] Cox, D. R. (1959). The analysis of exponentially distributed life-times with two types of failure. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 21(2):411–421.

[Cox, 1972] Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society: Series B*, 34(2):187–220.

[Crowley and Hu, 1977] Crowley, J. and Hu, M. (1977). Covariance analysis of heart transplant survival data. *Journal of the American Statistical Association*, 72(357):27–36.

[Dettoni et al., 2020] Dettoni, R., Marra, G., and Radice, R. (2020). Generalized link-based additive survival models with informative censoring. *Journal of Computational and Graphical Statistics*, 29(3):503–512.

[Efron, 1977] Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American statistical Association*, 72(359):557–565.

[Efron, 1979] Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1–26.

[Efron, 1982] Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for Industrial and Applied Mathematics.

[Efron, 1985] Efron, B. (1985). Bootstrap confidence intervals for a class of parametric problems. *Biometrika*, 72(1):45–58.

[Eletti et al., 2023] Eletti, A., Marra, G., and Radice, R. (2023). A spline-based framework for the flexible modelling of continuously observed multistate survival processes. *Statistical Modelling*, 23(5-6):495–509.

[Fehrenbacher et al., 2020] Fehrenbacher, L., Cecchini, R. S., Geyer Jr, C. E., Rastogi, P., Costantino, J. P., Atkins, J. N., Crown, J. P., Polikoff, J., Boileau, J.-F., Provencher, L., et al. (2020). NSABP B-47/NRG oncology phase III randomized trial comparing adjuvant chemotherapy with or without trastuzumab in high-risk invasive breast cancer negative for HER2 by FISH and with IHC 1+ or 2+. *Journal of Clinical Oncology*, 38(5):444.

[Ferreira and Patino, 2019] Ferreira, J. C. and Patino, C. M. (2019). Loss to follow-up and missing data: important issues that can affect your study results. *Jornal Brasileiro de Pneumologia*, 45:e20190091.

[Fine and Gray, 1999] Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association*, 94(446):496–509.

[Fine et al., 2001] Fine, J. P., Jiang, H., and Chappell, R. (2001). On semi-competing risks data. *Biometrika*, 88(4):907–919.

[Food and Drug Administration, 2022] Food and Drug Administration (2022). Patient-focused drug development: selecting, developing, or modifying fit-for-purpose clinical outcome assessments (draft guidance). `https://www.fda.gov/regulatory-information/search-fda-guidance-documents/patient-focused-drug-development-selecting-developing-or-modifying-fit-purpose-clinical-outcome`. Accessed: 2024-03-31.

[Ghalanos and Theussl, 2015] Ghalanos, A. and Theussl, S. (2015). *Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method*. R package version 1.16.

[Gingell and Mendivil, 2023] Gingell, K. and Mendivil, F. (2023). Random walks, directed cycles, and Markov chains. *The American Mathematical Monthly*, 130(2):127–144.

[Grambsch and Therneau, 1994] Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526.

[Gray, 2022] Gray, R. J. (2022). cmprsk: Subdistribution analysis of competing risks. `https://cran.r-project.org/package=cmprsk`.

[Haller, 2014] Haller, B. (2014). *The analysis of competing risks data with a focus on estimation of cause-specific and subdistribution hazard ratios from a mixture model*. PhD thesis, lmu.

[Haneuse and Lee, 2016] Haneuse, S. and Lee, K. H. (2016). Semi-competing risks data analysis: accounting for death as a competing risk when the outcome of interest is nonterminal. *Circulation: Cardiovascular Quality and Outcomes*, 9(3):322–331.

[Hougaard, 1999] Hougaard, P. (1999). Multi-state models: a review. *Lifetime data analysis*, 5(3):239–264.

[Jackson, 2016] Jackson, C. (2016). flexsurv: A platform for parametric survival modeling in R. *Journal of Statistical Software*, 70(8):1–33.

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

[Kang, 2013] Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, 64(5):402.

[Kaplan and Meier, 1958] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

[Król and Saint-Pierre, 2015] Król, A. and Saint-Pierre, P. (2015). SemiMarkov: An R package for parametric estimation in multi-state semi-markov models. *Journal of Statistical Software*, 66(6):1–16.

[Larson and Dinse, 1985] Larson, M. G. and Dinse, G. E. (1985). A mixture model for the regression analysis of competing risks data. *Journal of the Royal Statistical Society: Series C*, 34(3):201–211.

[Latouche et al., 2013] Latouche, A., Allignol, A., Beyersmann, J., Labopin, M., and Fine, J. P. (2013). A competing risks analysis should report results on all cause-specific hazards and cumulative incidence functions. *Journal of Clinical Epidemiology*, 66(6):648–653.

[Lawless, 2003] Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data (Second Edition).* John Wiley & Sons, Inc.

[Lee and Wolfe, 1998] Lee, S.-Y. and Wolfe, R. A. (1998). A simple test for independent censoring under the proportional hazards model. *Biometrics*, pages 1176–1182.

[Mamounas et al., 2019] Mamounas, E. P., Bandos, H., Lembersky, B. C., Jeong, J.-H., Geyer, C. E., Rastogi, P., Fehrenbacher, L., Graham, M. L., Chia, S. K., Brufsky, A. M., et al. (2019). Use of letrozole after aromatase inhibitor-based therapy in postmenopausal breast cancer (NRG Oncology/NSABP B-42): a randomised, double-blind, placebo-controlled, phase 3 trial. *The Lancet Oncology*, 20(1):88–99.

[Meira-Machado et al., 2009] 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. *Statistical Methods in Medical Research*, 18(2):195–222.

[Meng and Rubin, 1991] Meng, X.-L. and Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, 86(416):899–909.

[Milic et al., 2021] Milic, J., Banchelli, F., Meschiari, M., Franceschini, E., Ciusa, G., Gozzi, L., Volpi, S., Faltoni, M., Franceschi, G., Iadisernia, V., et al. (2021). The impact of tocilizumab on respiratory support states transition and clinical outcomes in COVID-19 patients. a Markov model multi-state study. *Plos one*, 16(8):e0251378.

[Miller et al., 2020] Miller, D. S., Filiaci, V. L., Mannel, R. S., Cohn, D. E., Matsumoto, T., Tewari, K. S., DiSilvestro, P., Pearl, M. L., Argenta, P. A., Powell, M. A., et al. (2020). Carboplatin and paclitaxel for advanced endometrial cancer: final overall survival and adverse event analysis of a phase III trial (NRG Oncology/GOG0209). *Journal of Clinical Oncology*, 38(33):3841.

[Morris et al., 2019] Morris, T. P., White, I. R., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine*, 38(11):2074–2102.

[Muehlemann et al., 2023] Muehlemann, N., Zhou, T., Mukherjee, R., Hossain, M. I., Roychoudhury, S., and Russek-Cohen, E. (2023). A tutorial on modern Bayesian methods in clinical trials. *Therapeutic Innovation & Regulatory Science*, 57(3):402–416.

[Nelson, 1969] Nelson, W. (1969). Hazard plotting for incomplete failure data. *Journal of Quality Technology*, 1(1):27–52.

[Nelson, 1972] Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):945–966.

[Oberoi et al., 2020] Oberoi, D., Piedalue, K.-A. L., Pirbhai, H., Guirguis, S., Santa Mina, D., and Carlson, L. E. (2020). Factors related to dropout in integrative oncology clinical trials: interim analysis of an ongoing comparative effectiveness trial of mindfulness-based cancer recovery and Tai chi/Qigong for cancer health (the MATCH study). *BMC Research Notes*, 13(1):1–7.

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

[Putter et al., 2007] Putter, H., Fiocco, M., and Geskus, R. B. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, 26(11):2389–2430.

[Robins and Rotnitzky, 1992] Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology: Methodological Issues*, pages 297–331. Springer.

[Rosenberg, 1995] Rosenberg, P. S. (1995). Hazard function estimation using B-splines. *Biometrics*, pages 874–887.

[Ross, 2007] Ross, S. (2007). Composite outcomes in randomized clinical trials: arguments for and against. *American Journal Of Obstetrics And Gynecology*, 196(2):119–e1.

[Smith et al., 2019] Smith, E., Eleuteri, A., Goilav, B., Lewandowski, L., Phuti, A., Rubinstein, T., Wahezi, D., Jones, C., Marks, S., Corkhill, R., et al. (2019). A Markov multi-state model of lupus nephritis urine biomarker panel dynamics in children: Predicting changes in disease activity. *Clinical Immunology*, 198:71–78.

[Spruance et al., 2004] Spruance, S. L., Reid, J. E., Grace, M., and Samore, M. (2004). Hazard ratio in clinical trials. *Antimicrobial Agents and Chemotherapy*, 48(8):2787–2792.

[Stephens, 1974] Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69(347):730–737.

[Sutradhar and Austin, 2018] Sutradhar, R. and Austin, P. C. (2018). Relative rates not relative risks: addressing a widespread misinterpretation of hazard ratios. *Annals of Epidemiology*, 28(1):54–57.

[Therneau, 2023] Therneau, T. M. (2023). *A Package for Survival Analysis in R*. R package version 3.5-7.

[Touraine, 2019] Touraine, C. (2019). *Illness-Death Model*, pages 1–9. John Wiley & Sons, Ltd.

[Tsiatis, 1975] Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, 72(1):20–22.

[Tsiatis, 1981] Tsiatis, A. A. (1981). A large sample study of Cox's regression model. *The Annals of Statistics*, 9(1):93–108.

[Ursino and Stallard, 2021] Ursino, M. and Stallard, N. (2021). Bayesian approaches for confirmatory trials in rare diseases: opportunities and challenges. *International Journal of Environmental Research and Public Health*, 18(3):1022.

[van der Vaart, 2000] van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge University Press.

[Ventura et al., 2014] Ventura, L., Carreras, G., Puliti, D., Paci, E., Zappa, M., and Miccinesi, G. (2014). Comparison of multi-state Markov models for cancer progression with different procedures for parameters estimation. an application to breast cancer. *Epidemiology, Biostatistics, and Public Health*, 11(1).

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

[Wolbers et al., 2014] Wolbers, M., T. Koller, M., S. Stel, V., Schaer, B., J. Jager, K., Leffondré, K., and Heinze, G. (2014). Competing risks analyses: objectives and approaches. *European Heart Journal*.

[Ye, 1987] Ye, Y. (1987). *Interior Algorithms for Linear, Quadratic, and Linearly Constrained Non-Linear Programming*. PhD thesis, Department of ESS, Stanford University.

[Younes and Lachin, 1997] Younes, N. and Lachin, J. (1997). Link-based models for survival data with interval and continuous time censoring. *Biometrics*, pages 1199–1211.

[Yu, 2015] Yu, S.-Z. (2015). *Hidden Semi-Markov models: theory, algorithms and applications*. Morgan Kaufmann.

[Zhou et al., 2013] Zhou, B., Fine, J., and Laird, G. (2013). Goodness-of-fit test for proportional subdistribution hazards model. *Statistics in Medicine*, 32(22):3804–3811.

Appendix A

Further exploration of results in Section 5.3.2

Section 5.3.2 showed results where a model with significant censoring caused bias in the estimates of the survival function, and made it appear as though the models fitted on censored data perform “better” than those associated with similar data without censoring. We briefly discuss why this can happen through the results of a simple simulation study, and describe under what circumstances the issue becomes most obvious.

A.1 Simulation setup and findings

We consider a model with 2 states and one possible transition (in other words, a standard survival model). The true distribution associated with event times is a Weibull distribution with rate parameter $a = 2$ and shape parameter $b = 0.35$. We simulate 1000 datasets of sample size $m = 10000$ where the data are non-censored, and another 1000 sets of data of the same sample size where the data are censored. The censoring times are simulated from a Uniform(0.001, 10) distribution. This results in an average censoring rate of about 14.55%.

We then fit the (correct) Weibull model on each dataset and obtain an average parameter vector estimate for each of the non-censored datasets as well the censored datasets. After this, we plot the true survival function and compare it with the fitted survival functions associated with average parameter vectors for both types of data (non-censored and censored). Finally, we mis-specify the model with exponential and gamma distributions and repeat the same procedure each time.

Figure A.1 below shows a plot of the relevant survival functions. The solid black line is that of the true survival function. Blue, brown, and orange lines are associated with Weibull, exponential, and gamma model fits respectively. Dot-dash and dotted lines are associated with non-censored and censored data respectively.

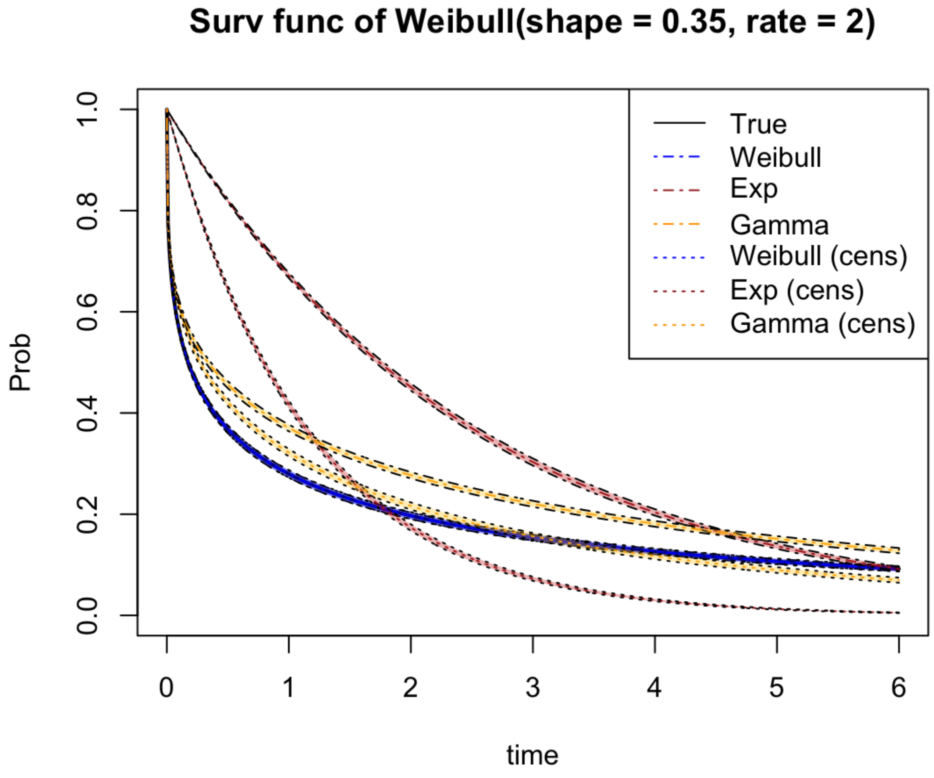


Figure A.1: Survival functions associated with the true model and relevant model fits. The data are from a Weibull model with rate parameter $a = 2$ and shape parameter $b = 0.35$. The censoring times are from a Uniform(0.001, 10) distribution, which results in an average censoring rate of approximately 14.55%. The solid black line is that of the true survival function. Blue, brown, and orange lines are associated with Weibull, exponential, and gamma model fits respectively. Dot-dash and dotted lines are associated with non-censored and censored data respectively. Confidence intervals are obtained using the delta method.

As seen in Section 5.3.2, we find that there is no perceivable difference whether or not there is censoring when the true model is fitted to the data. However, differences in fitted survival functions become apparent when the models are misspecified.

We now repeat the above simulation for each of $b = 0.6$ and $b = 0.8$ (with a left unchanged). A known property of the Weibull distribution is that $b < 0$ results in a hazard function which decreases with time while $b > 0$ results in a hazard function that increases with time (with $b = 1$ resulting in the exponential distribution, which has a hazard function which is constant with respect to time). Hence, the parameters of the censoring distribution are adjusted in each case in order to maintain a similar level of average censoring. When $b = 0.5$, the censoring distribution is Uniform(0.001, 5.9) and when $b = 0.8$ the censoring distribution is Uniform(0.001, 3.8). In each case, the average censoring rate is roughly 14.52% and 14.72% respectively. Table A.1 summarises all the information associated with the simulation study.

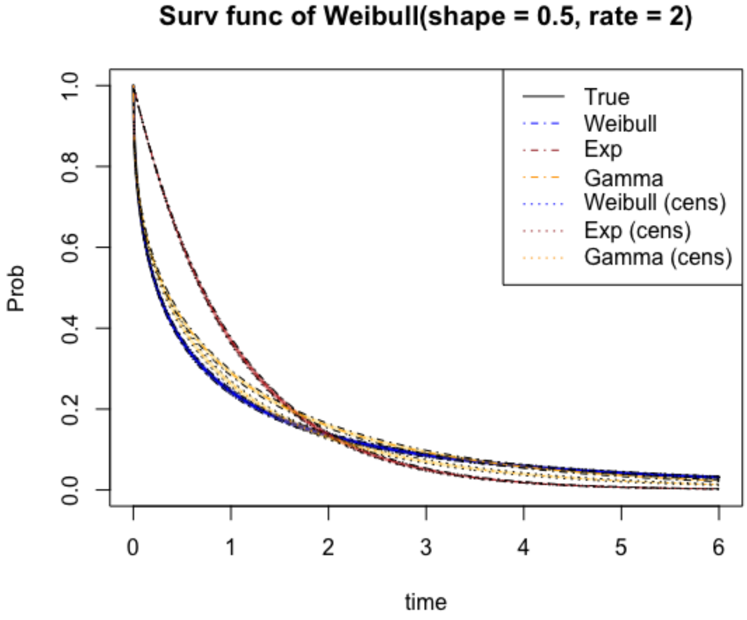
Weibull(rate, shape)	Censoring distribution	Censoring rate
(2, 0.35)	Uniform(0.001, 10)	14.55%
(2, 0.50)	Uniform(0.001, 5.9)	14.52%
(2, 0.80)	Uniform(0.001, 3.8)	14.72%

Table A.1: Summary of parameters and censoring rates associated with simple simulation study involving Weibull data. The data are all from Weibull models with rate parameter 2, but different shape parameters. The censoring distributions are slightly different in each case to ensure a roughly equal amount of average censoring.

Figure A.2 shows survival functions similar to that seen in Figure A.1, but associated with the larger shape parameters of $b = 0.5$ and $b = 0.8$. Other than slight differences when $b = 0.5$ we find that there are no dramatic differences in model fits owing to differences in censoring.

A.2 Discussion

When there is no censoring and the Weibull model has shape parameter $b = 0.35$, there is initially a very high rate of exits out of the initial state. As time passes, the rate of exits out of the initial state decreases. This is due to the fact that $b < 1$



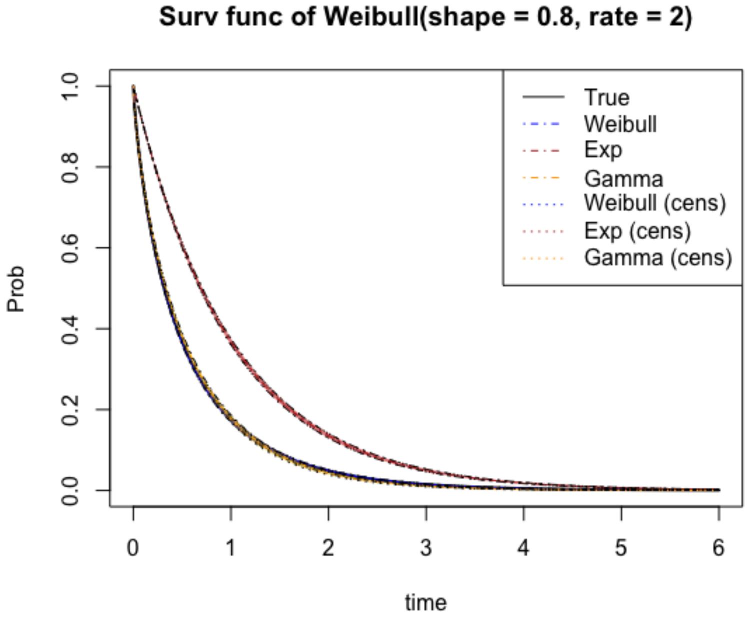


Figure A.2: Survival functions associated with the true model and relevant model fits for different shape parameters. The solid black line is that of the true survival function. Blue, brown, and orange lines are associated with Weibull, exponential, and gamma model fits respectively. Dot-dash and dotted lines are associated with non-censored and censored data respectively. The top figure shows fitted survival functions associated with data that are from a Weibull model with rate parameter $a = 2$ and shape parameter $b = 0.50$. The bottom figure shows fitted survival functions associated with data that are from a Weibull model with rate parameter $a = 2$ and shape parameter $b = 0.80$. The average censoring rate in each case is roughly 14.52% and 14.72% respectively.

and is reflected by the true survival function in Figure A.1 having a steep drop just after $t = 0$ which very quickly flattens out. The exponential and gamma fits are not fully able to capture this aspect of the Weibull model, with the exponential fit performing much worse. However, when censoring is introduced, a significant proportion of the relatively large event times are obscured by censoring. Given that the model is misspecified and now most of the large event times are not observed, there is greater difficulty in estimating the survival function associated with a censored event. The maximum likelihood estimate then tends to suggest an exponential (or a gamma) distribution with a lower rate parameter than would have been otherwise obtained without censoring. This can be seen in Figure A.1.

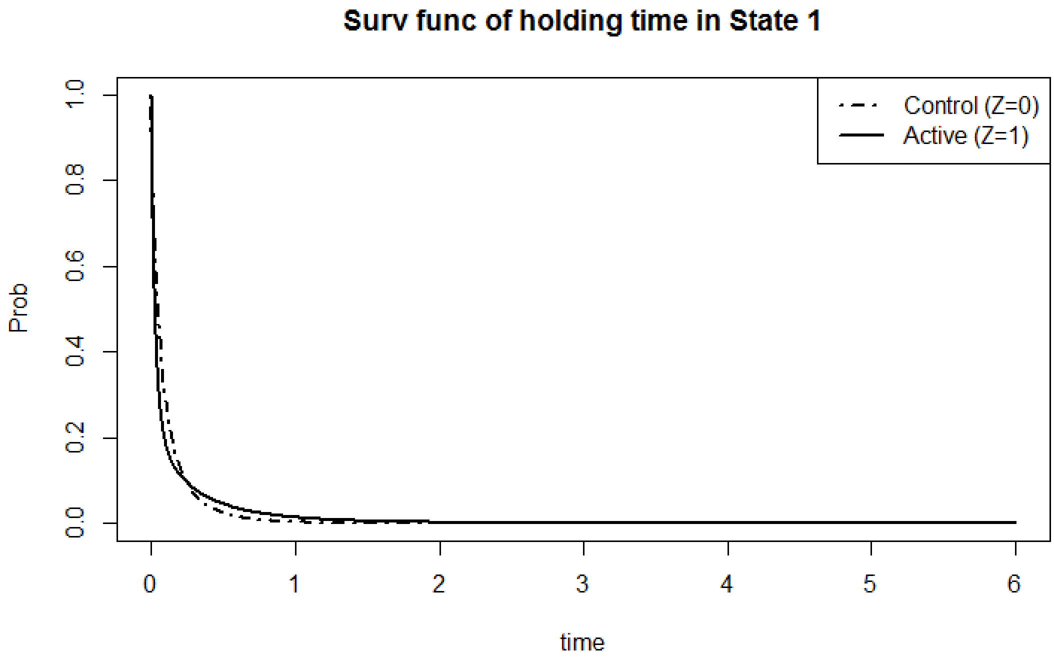
On the other hand, as we increase b towards unity, the Weibull model becomes "closer" to an exponential model and so the rate of exits out of the initial state become relatively more constant. This means that there are less relatively large event times as compared to when $b = 0.35$. Noting that we have kept the rate of censoring relatively similar, any censored observations are now likely to be associated with relatively small event times compared to when $b = 0.35$. For the misspecified models this leads to far less uncertainty in estimating the survival function associated with censored events. We can see from Figure A.2 that when $b = 0.8$, the misspecified model fits are indistinguishable regardless of censoring.

Overall, these results demonstrate that the effects of model misspecification might be greatly exacerbated in certain scenarios where clinicial trials end too prematurely.

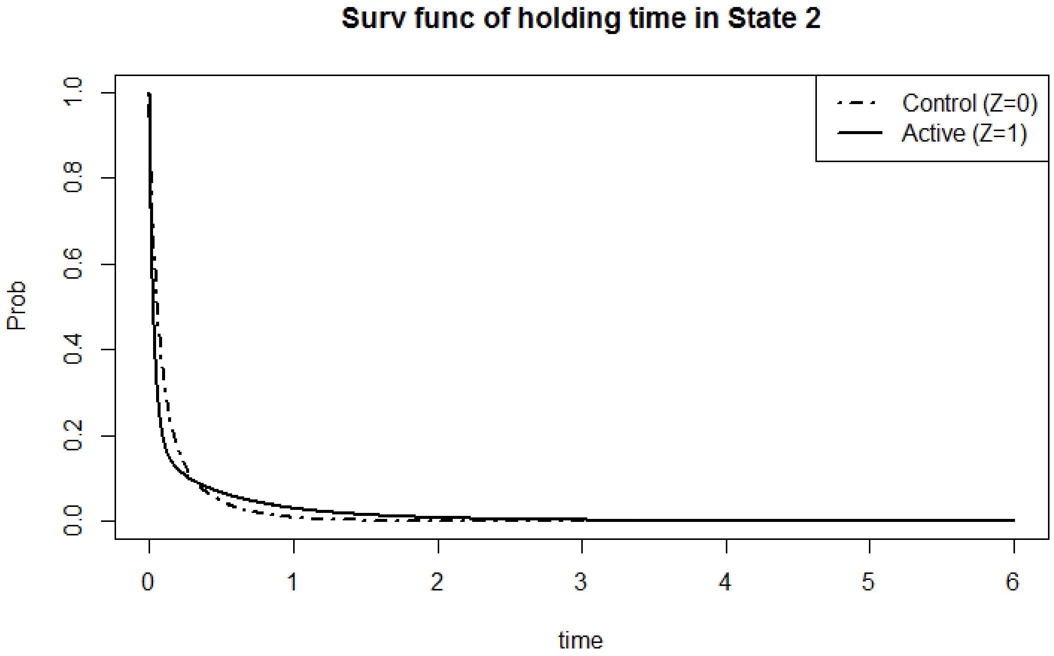
Appendix B

Chapter 5 simulation study additional figures

B.1 True survival functions of holding times associated with state 1 and state 2



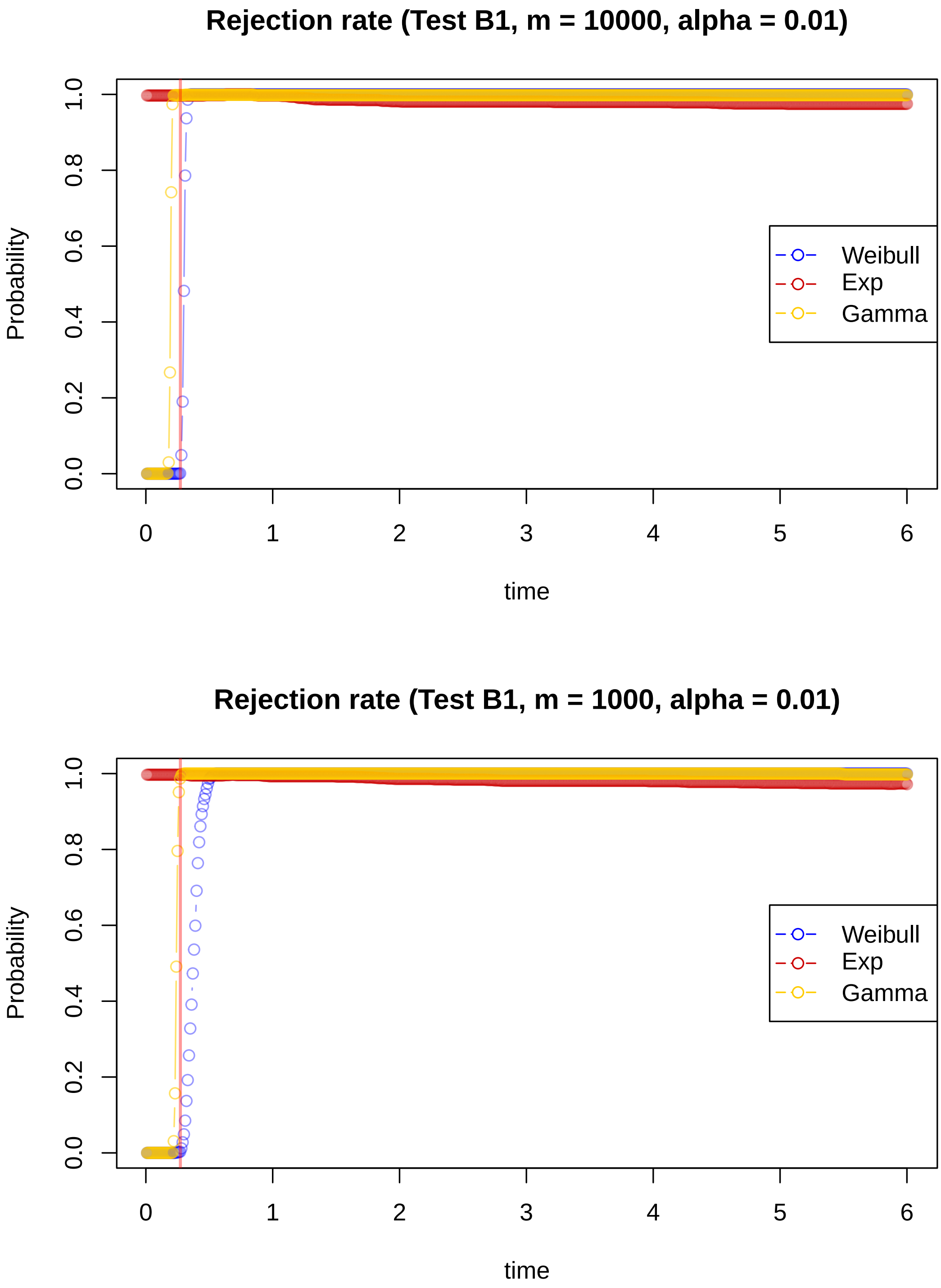
(a) True survival functions of holding times in state 1 for the simulation study in Chapter 5.



(b) True survival functions of holding times in state 2 for the simulation study in Chapter 5.

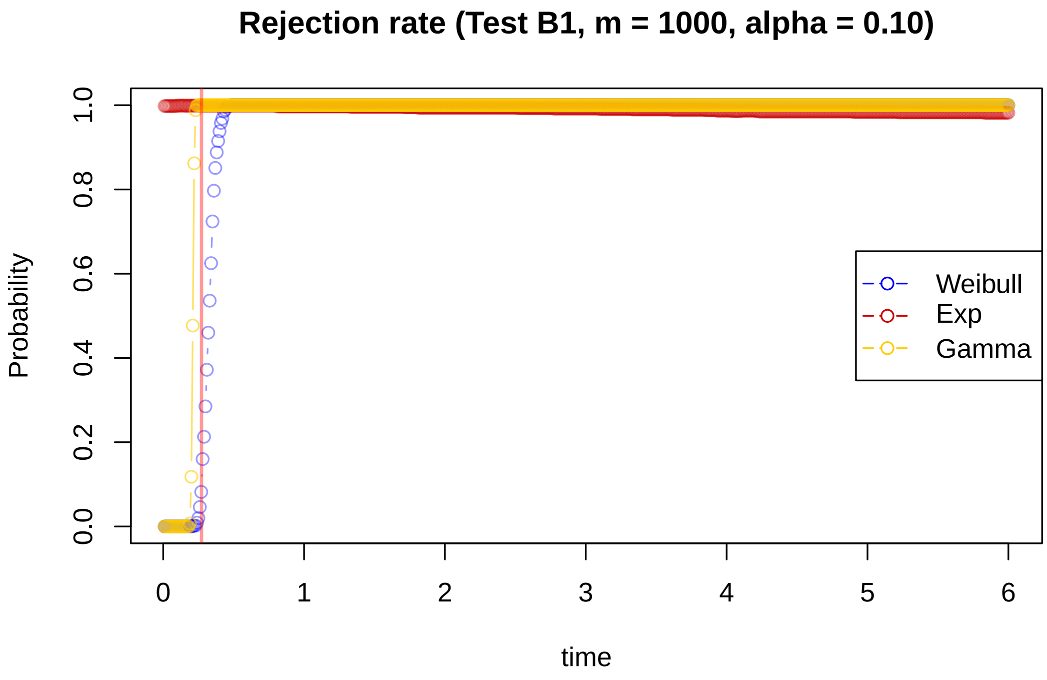
Figure B.1: rue survival functions of holding times in states 1 and 2 for the simulation study in Chapter 5.

B.2 Baseline model: Test B power functions


Rejection rate (Test B1, m = 10000, alpha = 0.01)

Rejection rate (Test B1, m = 1000, alpha = 0.01)

Figure B.2: (Baseline) Estimates of power of **Test B1** as described in Section 5.1.3, for $\alpha = 0.01$.



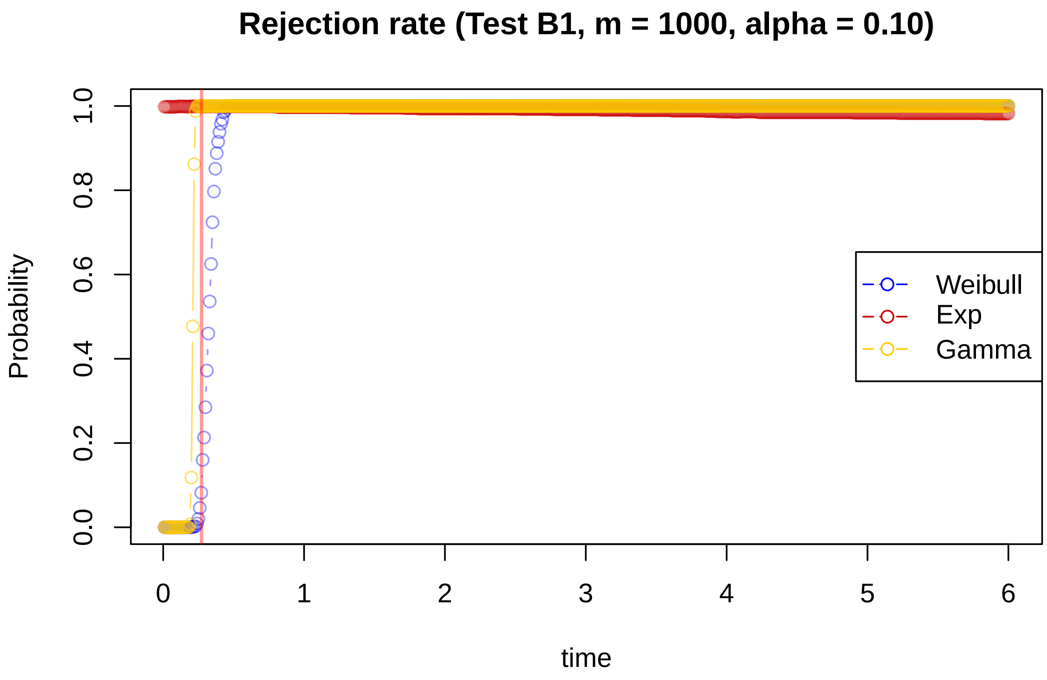
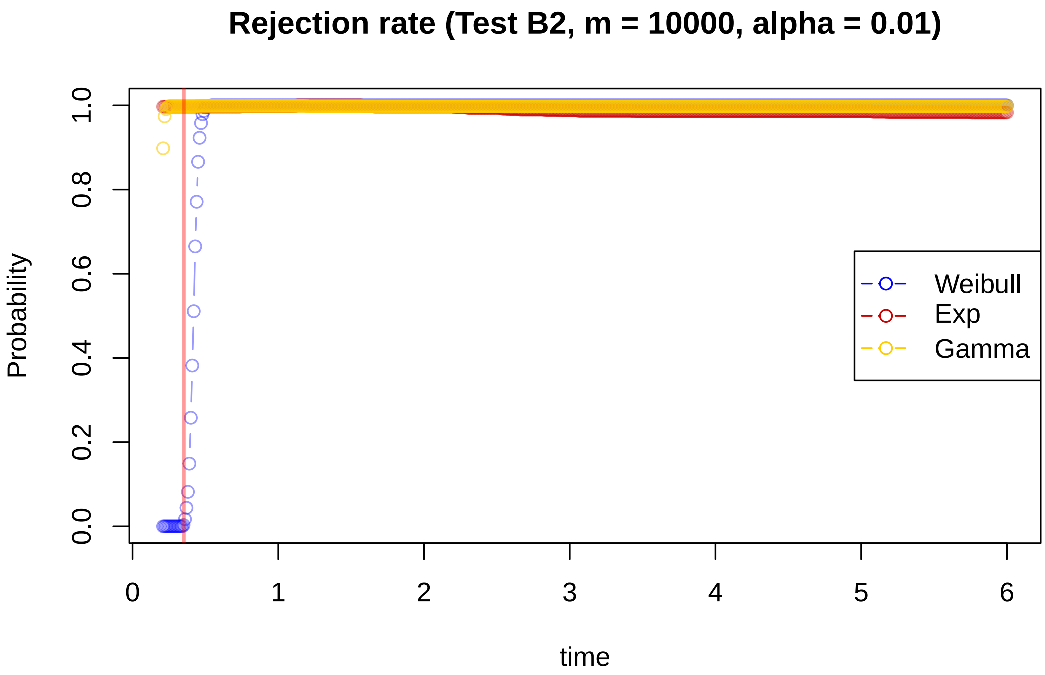


Figure B.3: (Baseline) Estimates of power of **Test B1** as described in Section 5.1.3, $\alpha = 0.10$.



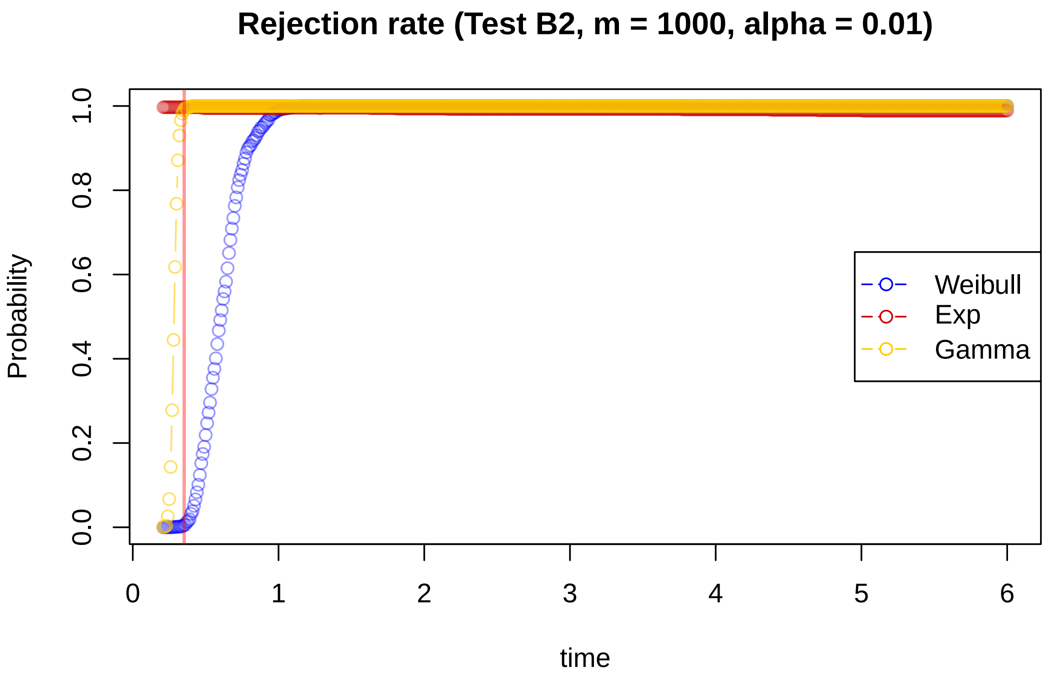


Figure B.4: (Baseline) Estimates of power of **Test B2** as described in Section 5.1.3, for $\alpha = 0.01$.

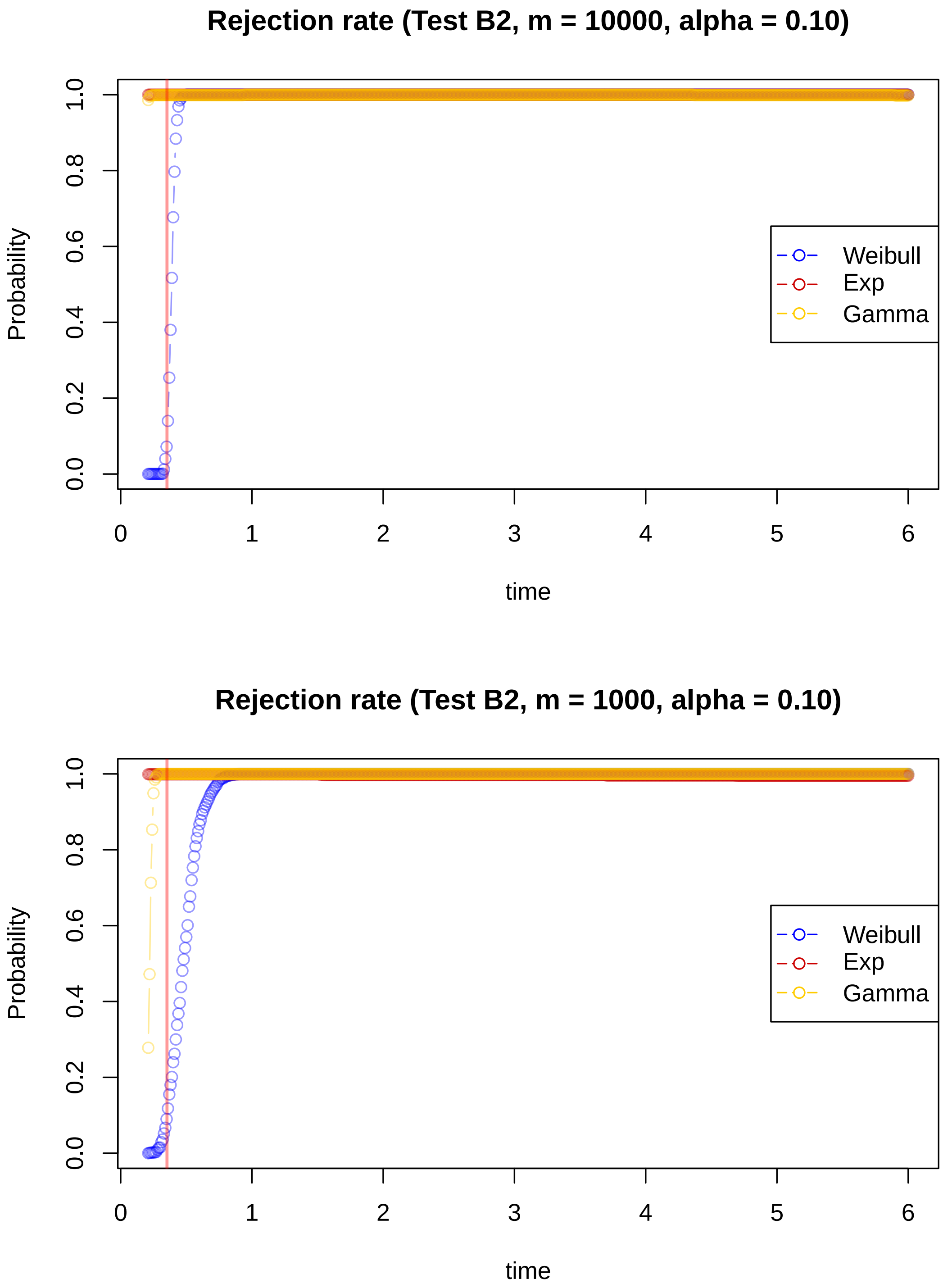
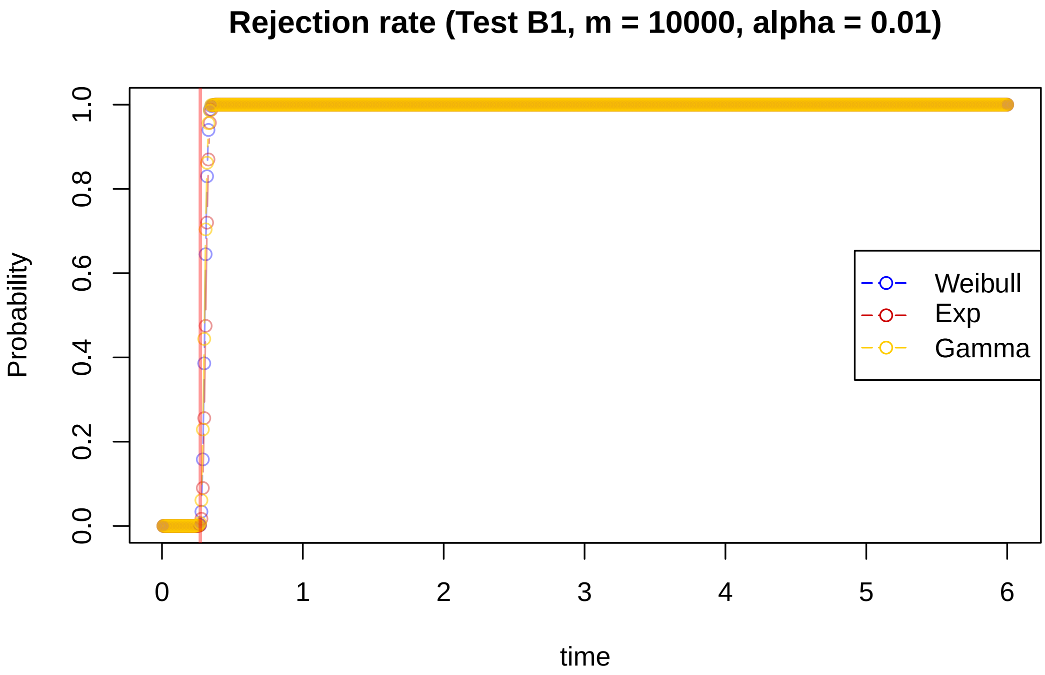


Figure B.5: (Baseline) Estimates of power of **Test B1** as described in Section 5.1.3, for $\alpha = 0.10$.

B.3 Baseline model with censoring: Test B power functions



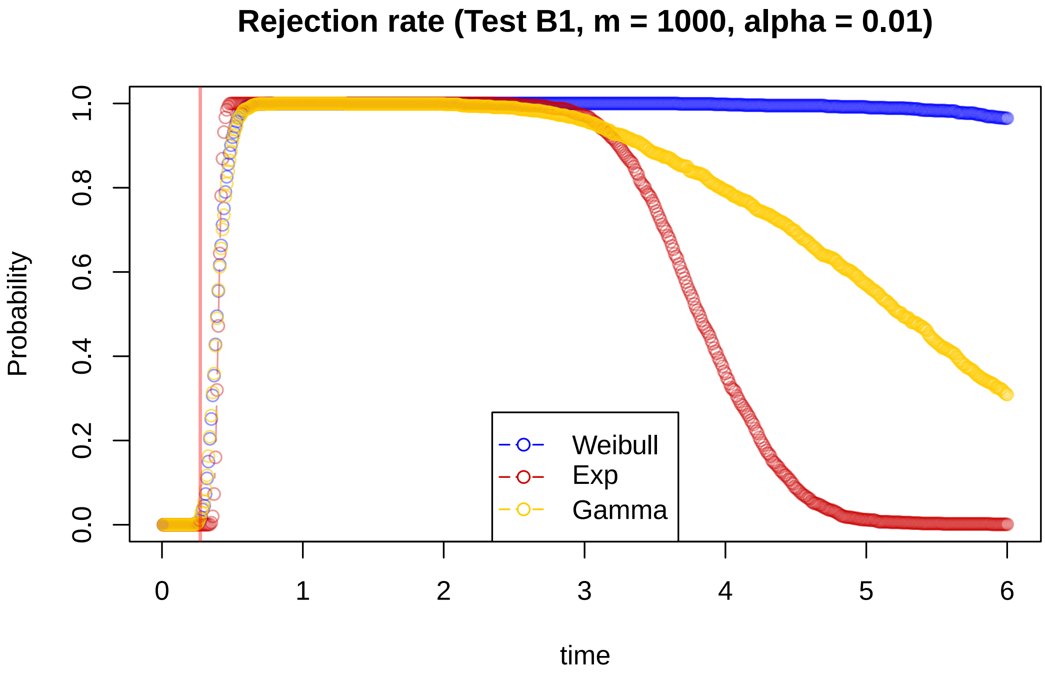


Figure B.6: (Baseline with censoring) Estimates of power of **Test B1** as described in Section 5.1.3, for $\alpha = 0.01$.

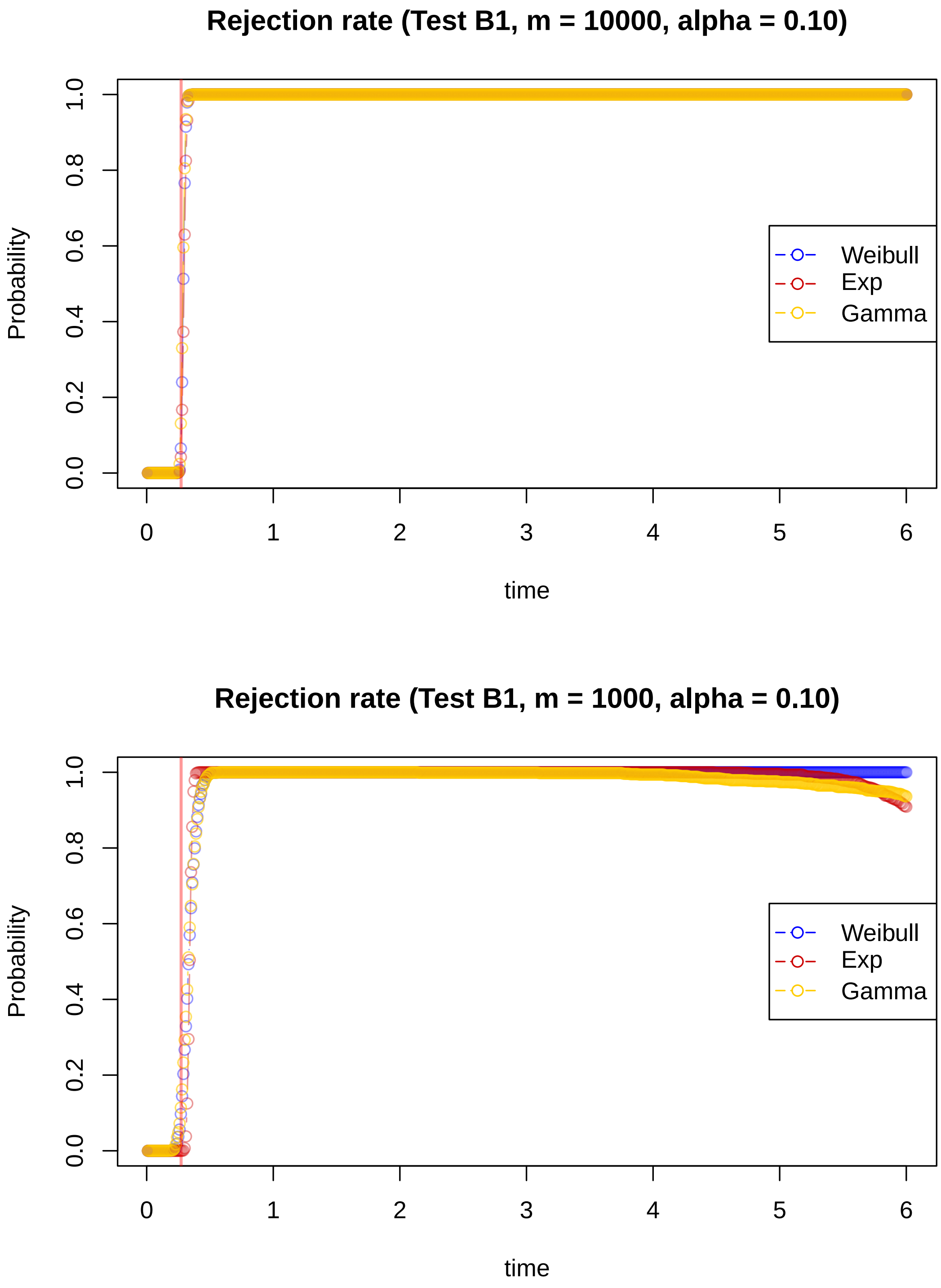
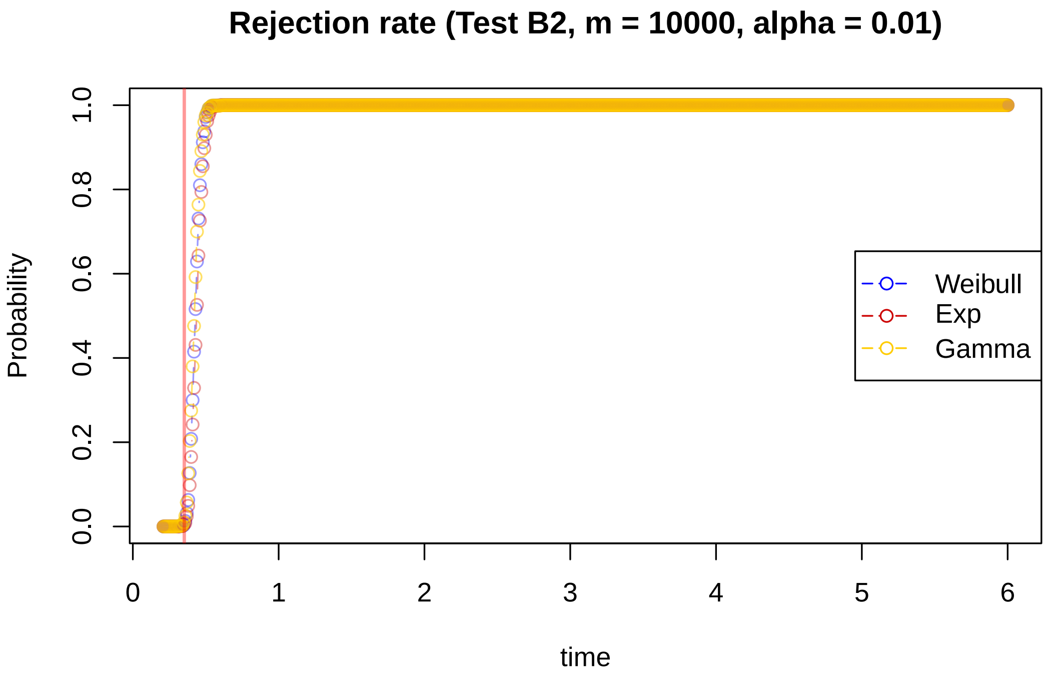


Figure B.7: (Baseline with censoring) Estimates of power of **Test B1** as described in Section 5.1.3, $\alpha = 0.10$.



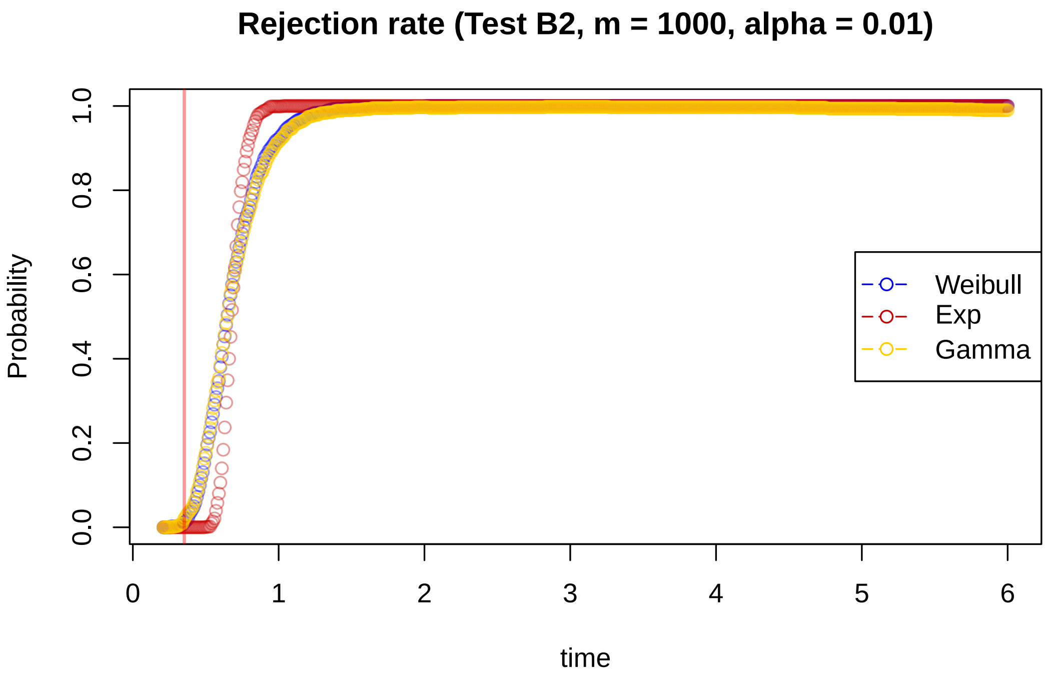
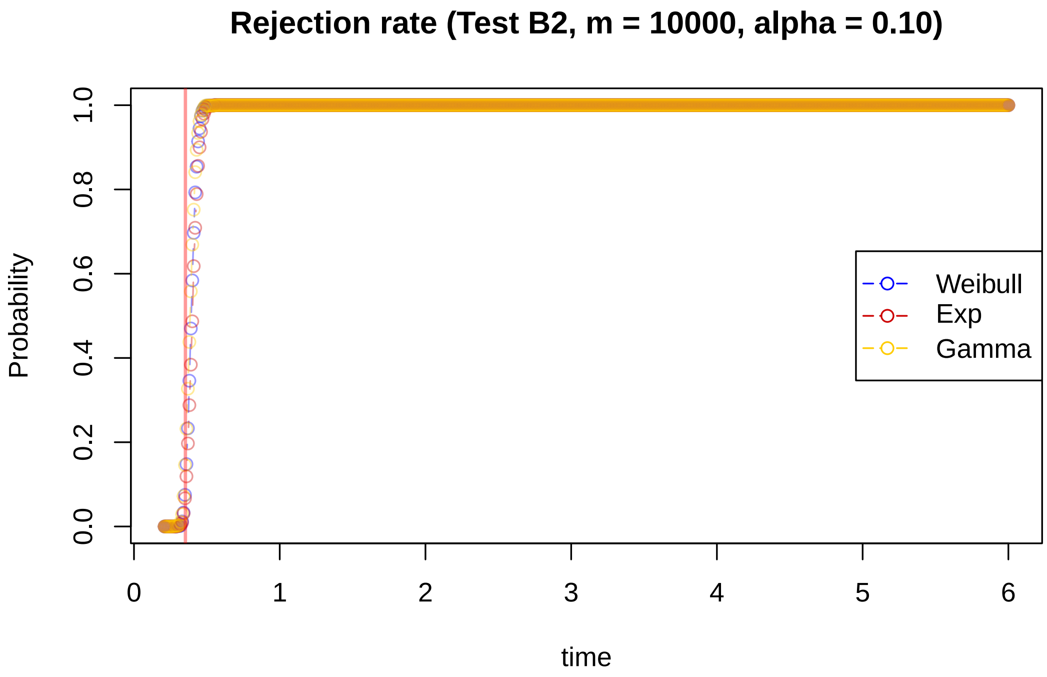


Figure B.8: (Baseline with censoring) Estimates of power of **Test B2** as described in Section 5.1.3, for $\alpha = 0.01$.



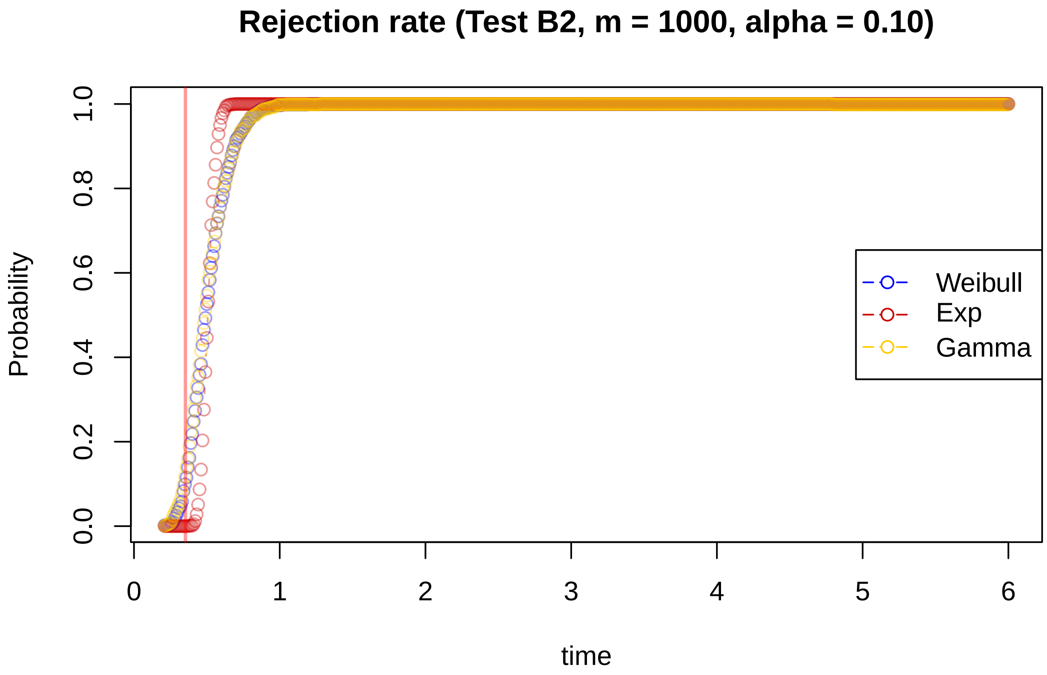
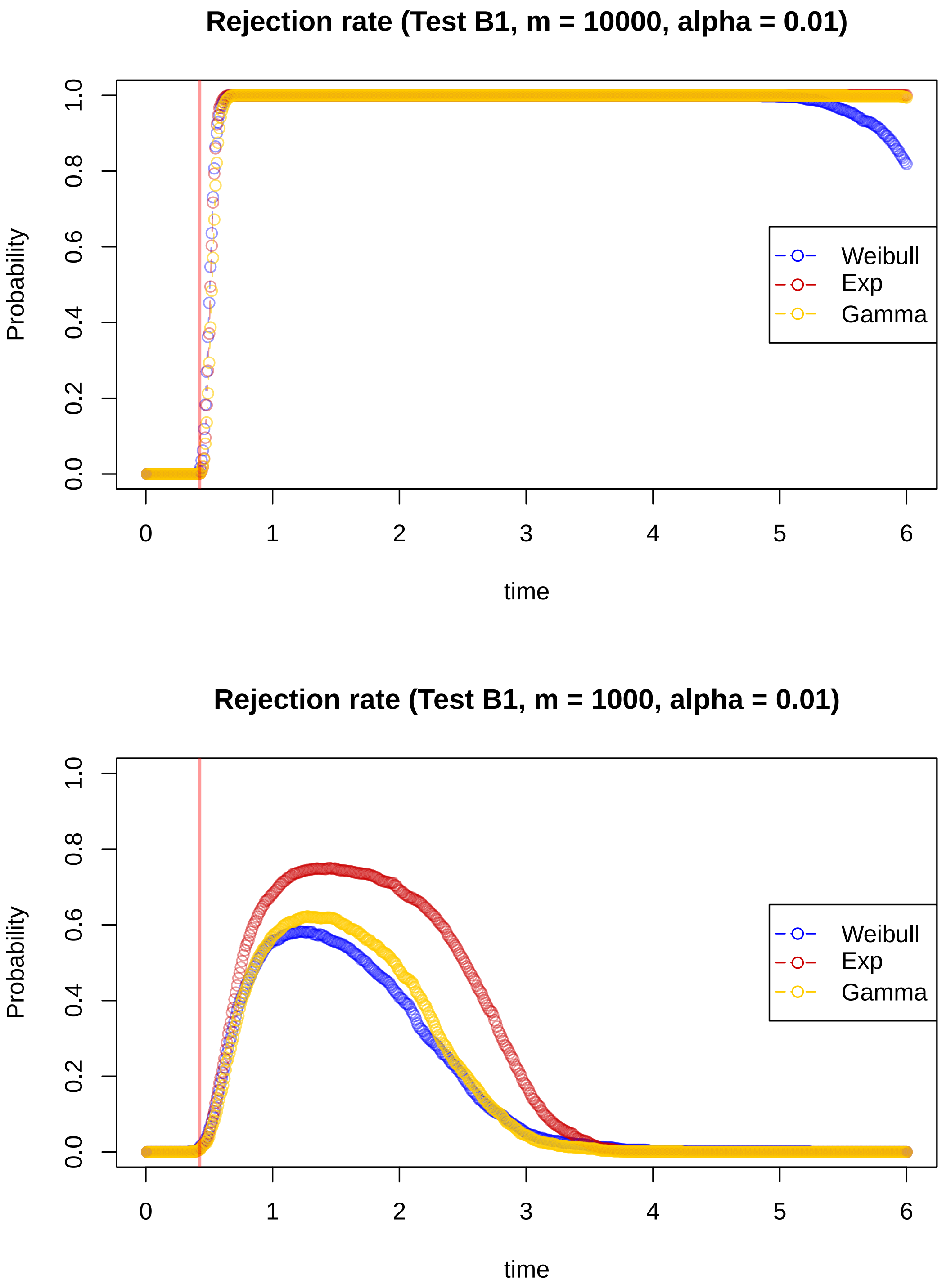


Figure B.9: (Baseline with censoring) Estimates of power of **Test B2** as described in Section 5.1.3, for $\alpha = 0.10$.

B.4 Model with less benefit and with censoring: Test B power functions


Rejection rate (Test B1, m = 10000, alpha = 0.01)

Rejection rate (Test B1, m = 1000, alpha = 0.01)

Figure B.10: (Less benefit with censoring) Estimates of power of **Test B1** as described in Section 5.2.2, for $m = 10000$ and $\alpha = 0.01$.

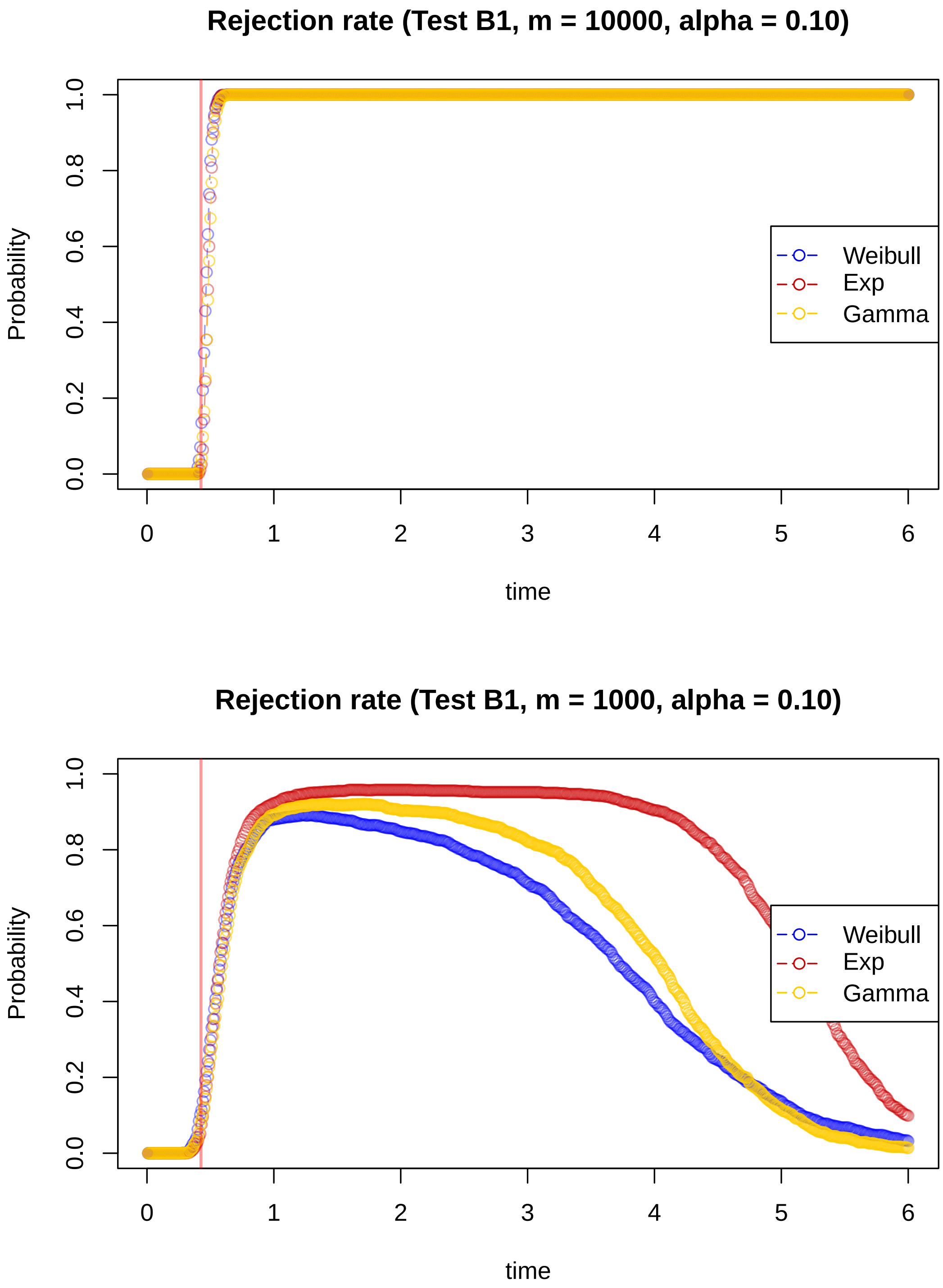
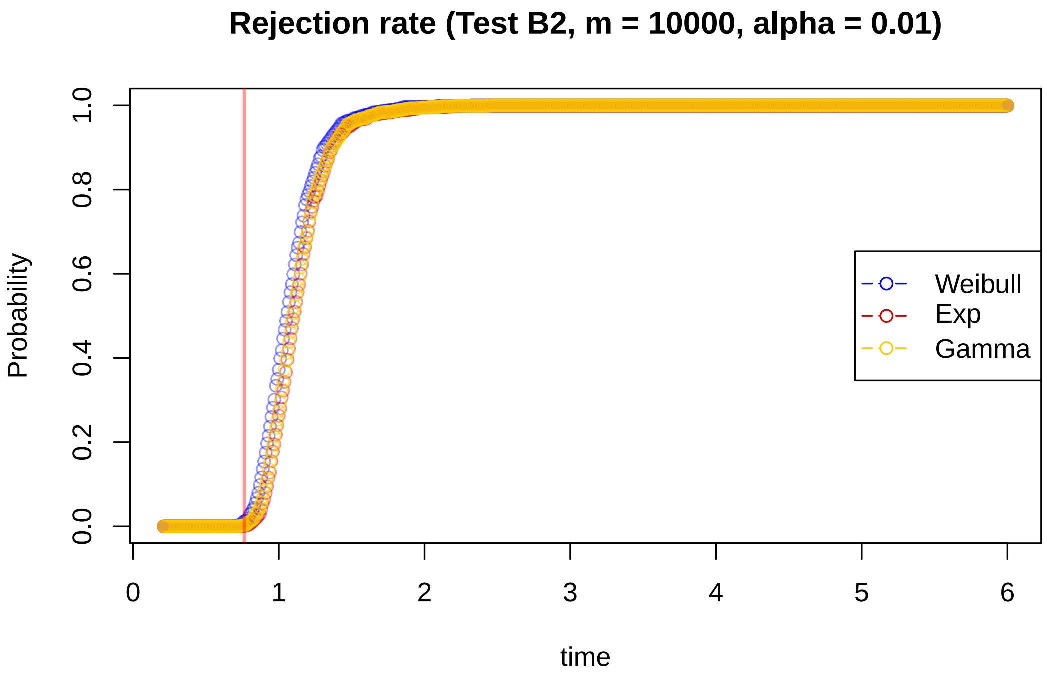


Figure B.11: (Less benefit with censoring) Estimates of power of **Test B1** as described in Section 5.1.3, for $\alpha = 0.10$.



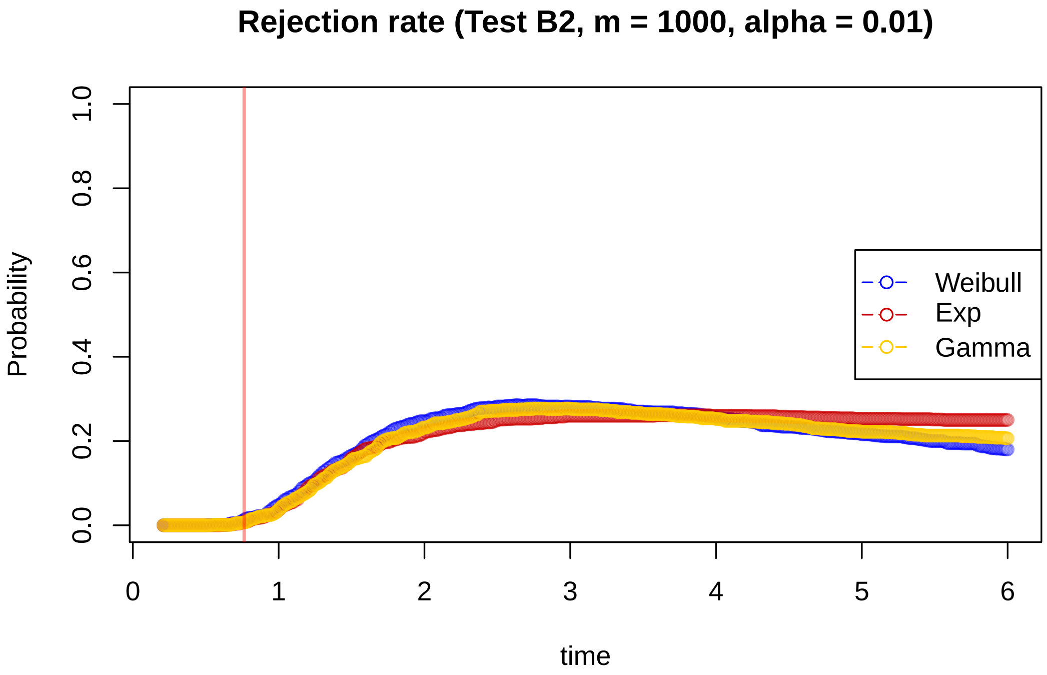
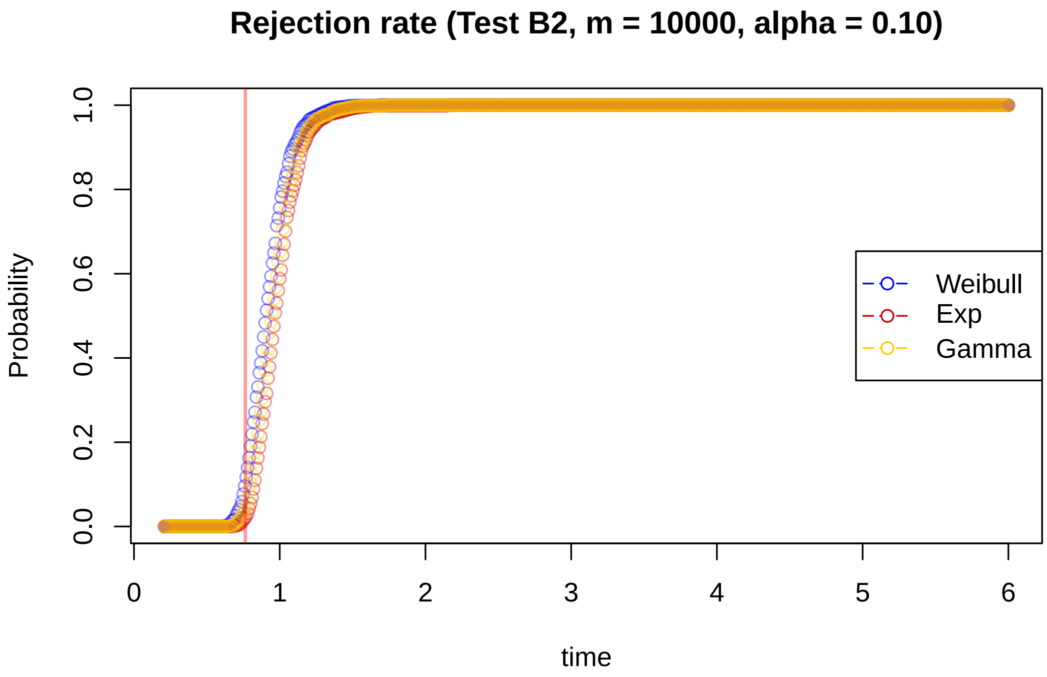


Figure B.12: (Less benefit with censoring) Estimates of power of **Test B2** as described in Section 5.1.3, $\alpha = 0.01$.



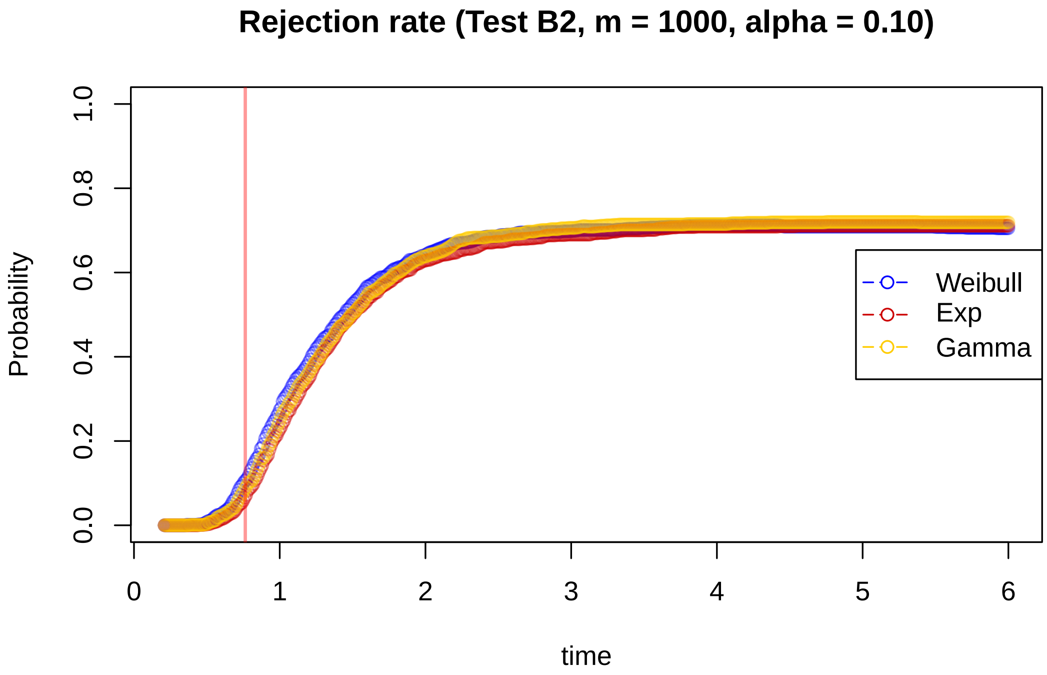


Figure B.13: (Less benefit with censoring) Estimates of power of **Test B2** as described in Section 5.1.3, for $\alpha = 0.10$.