PHD

Joint modelling of longitudinal and time-to-event data applied to group sequential trials

Burdon, Abigail

Award date:
2022

Awarding institution:
University of Bath

Joint modelling of longitudinal and time-to-event data applied to group sequential trials

submitted by

Abigail Jane Burdon

for the degree of Doctor of Philosophy

of the

University of Bath

Department of Mathematical Sciences

October 2021

COPYRIGHT

Declaration of any previous submission of the work

The material presented here for examination for the award of a higher degree by research has not been incorporated into a submission for another degree.

....

Abigail J. Burdon

Declaration of authorship

I am the author of this thesis, and the work described therein was carried out by myself personally, in collaboration with my supervisor, Christopher Jennison.

....

Abigail J. Burdon

Acknowledgements

First and foremost, I would like to thank my supervisor, Christoper Jennison for his expertise and help over the past four years. Thank you for your encouragement and patience, and for introducing me to a subject which I have found so interesting. Thanks also to Lisa Hampson, who has been instrumental in shaping the direction of this research. I'm grateful for the two fantastic summer placements at Novartis in Switzerland, which Lisa has organised and mentored me through, and have helped me to grow in my understanding of the pharmaceutical industry. I am grateful to EPSRC for sponsoring this project.

Thank you to all of my friends in Bath who I have shared an office, bike ride, coffee or storybook with, and thank you to all my friends in Bath who attended my wedding. A special thanks to Lizzi for so kindly giving me accommodation when I commuted to Bath. To my friends in my small group from St Matthias Church Plymouth, thank you for your many prayers during the final stages of this PhD. I wish to acknowledge my colleagues at MASH, particularly Emma and Andrew. This opportunity at MASH has helped me to develop as a Statistician, and your support has encouraged me to have confidence in my abilities.

This PhD project would not have been possible without the support of my family, in particular my parents Phil and Margaret and husband Dom — for your emotional support, for taking an interest in my research, and for all your love and kindness.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 2: Background and methods

Chapter 4: Joint model with treatment directly affecting survival

Chapter 5: Joint model with both longitudinal and survival treatment effects

LIST OF TABLES

Chapter 4: Joint model with treatment directly affecting survival

Chapter 5: Joint model with both longitudinal and survival treatment effects

CHAPTER 1

INTRODUCTION

CHAPTER 2

BACKGROUND AND METHODS

2.1 | Group sequential trials

2.1.1 | Hypothesis testing in group sequential trials

Many clinical trials focus on survival as the primary endpoint. When this is the case, it may take several years to observe enough events to make a statistically sound decision about the null hypothesis. A possible solution to avoid a prolonged trial is to look at, and assess, the accumulating data periodically. These looks are formally known as interim analyses and create a type of group sequential trial. A stopping rule is specified during the design of a group sequential trial and early stopping is when a trial is terminated at an interim analysis because it meets the stopping rule requirements. Stopping rules ensure that trials with negative results are terminated and for trials with positive results, the drug may be brought to market sooner. Some of the benefits of early stopping are categorised as ethical, administrative or economic and often, a clinical trial with a small sample size is desirable. An efficient group sequential design reduces sample sizes through early stopping whilst controlling for type 1 error and without decreasing power. In some cases, the group sequential design reduces the expected sample size to 60% of the fixed sample trial.

The primary aim of a Phase 3 clinical trial is to show that the new treatment is more effective than the standard treatment, which shall be done by hypothesis testing. Throughout, assume that the model describing the relationship between patient covariates and the clinical endpoint is known. Also assume that included in the model is a parameter θ, that defines the difference in outcome distributions between the treatment group and the control group. We are interested in testing the null hypothesis $H_0 : \theta \leq 0$ against the one-sided alternative hypothesis $H_A : \theta > 0$, where a treatment difference $\theta > 0$ means that the new treatment is superior.

To perform the hypothesis test in the fixed sample trial, it is necessary to first find an estimate for the treatment effect, let this estimate be $\hat{\theta}$. We can calculate the information level $\mathcal{I} = [Var(\hat{\theta})]^{-1}$, in order to define the standardised test statistic $Z = \hat{\theta}\sqrt{\mathcal{I}}$. Determining the distribution of $\hat{\theta}$ provides a distribution for Z. Throughout this Thesis, we shall often prove that the treatment effect estimate is normally distributed such that $\hat{\theta} \sim N(\theta, \mathcal{I}^{-1})$, and therefore $Z \sim N(\theta\sqrt{\mathcal{I}}, 1)$. Then, deciding upon a suitable type 1 error rate α, the constant c is calculated such that $\mathbb{P}_{\theta=0}\{Z > c\} = \alpha$. In the fixed sample trial, to perform the hypothesis test: accept H_0 if $Z < c$ and reject H_0 if $Z > c$.

Further, at the design stage, a power requirement is specified which is used to determine the necessary sample size. Suppose that power is required to be $1-\beta$ at a minimum clinically significant effect size $\theta = \delta$, and let $\mathcal{I}_f$ be the information level required in a fixed sample trial in order that $\mathbb{P}_{\theta=\delta}\{Z > c\} = \beta$. Then $\mathcal{I}_f$ is found to be

$$\mathcal{I}_f = \left(\frac{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)}{\delta} \right)^2 . \tag{2.1}$$

Since information is a function of the number of patients n, Equation (2.1) can be used to calculate sample size.

Hypothesis testing in the group sequential trial is more complicated due to the dependency between data at different interim analyses and similarly to many other multiple testing procedures, we wish to control the overall type 1 error rate. A group sequential trial is made up of K analyses, each occurring at different points in calendar time. At each of $k = 1, \dots, K$, the accumulated data at analysis k is used to produce the parameter estimate $\hat{\theta}_k$ and its observed information level $\mathcal{I}_k$, which leads to the standardised test statistic at analysis k given by $Z_k = \hat{\theta}_k \sqrt{\mathcal{I}_k}$. The group sequential trial is designed so that at each interim analysis, the trial is either stopped to accept H_0, stopped to reject H_0 or continued. Therefore, at analysis k there is an interval (a_k, b_k) that splits the real line into three sections, each representing a different option. A one-sided group sequential test with K analyses is defined by Jennison and Turnbull (2000), Definition 4.2.1, and is summarised as:

```
After analysis $k = 1, \dots, K-1$
    if $Z_k \geq b_k$       stop, reject $H_0$
    if $Z_k \leq a_k$       stop, accept $H_0$
    otherwise         continue to group $k+1$,
after group $K$
    if $Z_K \geq b_K$       stop, reject $H_0$
    if $Z_K \leq a_K$       stop, accept $H_0$,
```

where $a_K = b_K$. This restriction applied at the final analysis is necessary to guarantee that the trial terminates before, or at, analysis K.

In the same way that the constant c is calculated to achieve type 1 error of α and power of $1-\beta$ at $\theta = \delta$ in the fixed trial, the constants $a_1, \dots, a_K$ and $b_1, \dots, b_K$ are calculated to attain the same type 1 error and power requirements in the group sequential trial. For this calculation in the group sequential setting, the joint distribution of the sequence of successive estimates $\hat{\theta}_1, \dots, \hat{\theta}_K$ must be known. The constants $a_1, \dots, a_K$ collectively are known as the lower boundary and the constants $b_1, \dots, b_K$ are the upper boundary. Figure 2.1 below represents a typical group sequential trial with 5 analyses. The outcome in this example is that the trial stops early at the fourth interim analysis and H_0 is accepted.

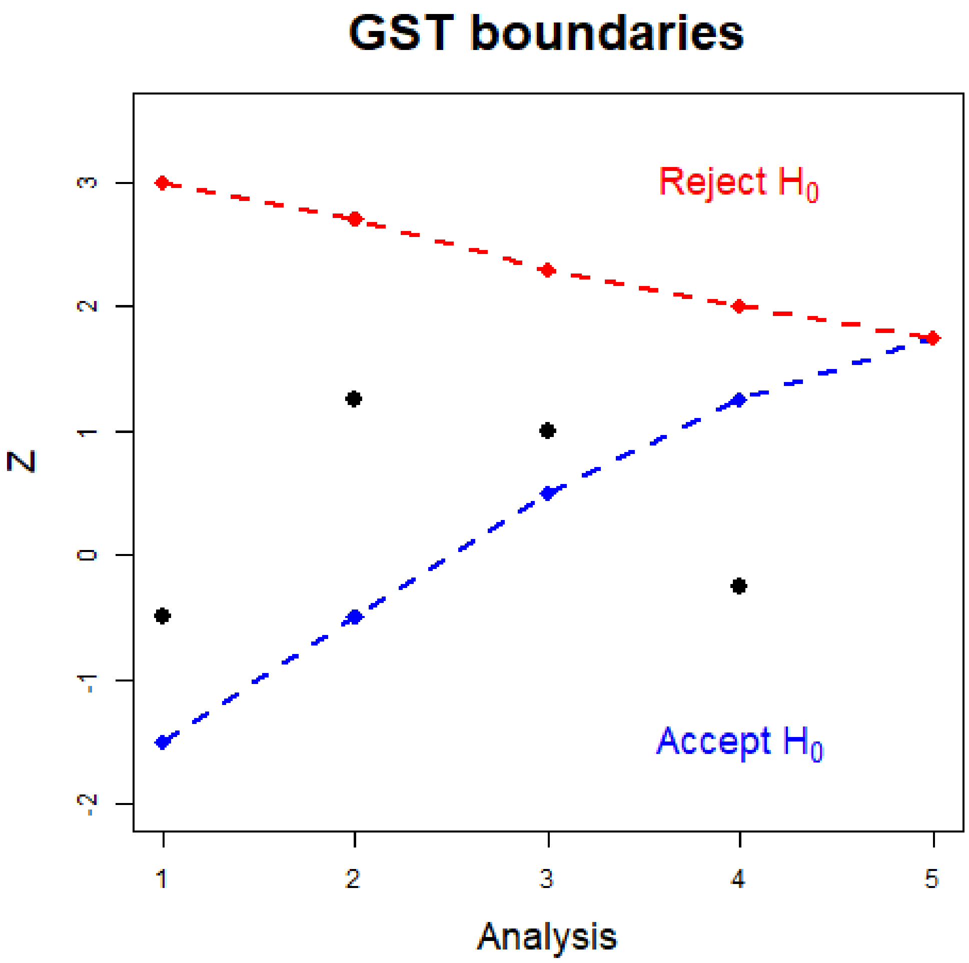


Figure 2.1: Group sequential trial with 5 analyses. Upper boundary constants are represented by red points, lower boundary constants by blue points and black points represent the standardised statistic trajectory $Z_1, \ldots, Z_4$. H_0 is accepted at analysis 4.

2.1.2 | Canonical joint distribution

To perform a group sequential trial, the distribution of the sequence of test statistics $Z_1, \ldots, Z_K$ must be known. Thus far, the group sequential trial has been described assuming that it is possible to calculate certain probabilities for events involving the test statistics. It can be shown that many data types result in a group sequential trial where the sequence of test statistics have a certain natural correlation structure. This particular correlation is known as the canonical joint distribution and simplifies the calculation for boundary constants.

Definition 2.1. *Suppose that a group sequential trial yields standardised test statistics $Z_1, \ldots, Z_K$ from data available at analyses $1, \ldots, K$ respectively and $\mathcal{I}_1, \ldots, \mathcal{I}_K$ are the associated observed information levels. The "canonical joint distribution" for the sequence of statistics $Z_1, \ldots, Z_k$ is such that*

1. *$(Z_1, \ldots, Z_K)$ is multivariate normal*
2. *$Z_k \sim N(\theta\sqrt{\mathcal{I}_k}, 1), \qquad 1 \leq k \leq K$*
3. *$Cov(Z_{k_1}, Z_{k_2}) = \sqrt{\mathcal{I}_{k_1}/\mathcal{I}_{k_2}}, \qquad 1 \leq k_1 \leq k_2 \leq K.$*

Sequences of test statistics with the canonical distribution have a Markov property; the distribution of $Z_{k+1}, \ldots, Z_K$ given $Z_1, \ldots, Z_k$ is the same as the distribution of $Z_{k+1}, \ldots, Z_K$ given Z_k. This property greatly simplifies the calculation of the boundary constants and is therefore very useful for implementing simulation studies. Jennison and Turnbull (1997) prove that the canonical joint distribution holds for many different data types. The focus of Section 3.2 is the proof that the canonical distribution holds for estimates of parameters in Cox's proportional hazards regression model for survival data, fitted by maximum partial likelihood, and in Section 4.1 we show why the canonical joint distribution is an appropriate assumption under the joint modelling framework.

We often refer to the canonical joint distribution of the sequence of treatment effect estimates $\hat{\theta}_1, \ldots, \hat{\theta}_K$ as, in some situations, it may be more convenient to prove that the canonical distribution holds under this parameterisation. The following definition can easily be seen to be equivalent to Definition 2.1.

Definition 2.2. *Let $\hat{\theta}_1, \ldots, \hat{\theta}_K$ be the sequence of treatment effect estimates in a group sequential trial from data available at analyses $1, \ldots, K$ respectively and $\mathcal{I}_1, \ldots, \mathcal{I}_K$ are the associated observed information levels. The "canonical joint distribution" for the sequence of estimates $\hat{\theta}_1, \ldots, \hat{\theta}_K$ is such that*

1. *$(\hat{\theta}_1, \ldots, \hat{\theta}_K)$ is multivariate normal*
2. *$\hat{\theta}_k \sim N(\theta, \mathcal{I}_k^{-1}), \qquad 1 \leq k \leq K$*
3. *$Cov(\hat{\theta}_{k_1}, \hat{\theta}_{k_2}) = \mathcal{I}_{k_2}^{-1}, \qquad 1 \leq k_1 \leq k_2 \leq K.$*

We can define a range of group sequential trials for a given sequence $\mathcal{I}_1, \ldots, \mathcal{I}_K$ of information levels and compute properties of these tests, such as expected number of patients. For all sequences, there is not a unique solution for the calculation of boundary constants. Therefore one can define a group sequential test with a chosen shape of boundary that has a given property, for example a conservative early interim analysis. The shape of the boundary, or the type of test, should be decided upon before the trial commences. The Pocock boundary designed by Pocock (1977) and the P&T by boundary Pampallona and Tsiatis (1994) are popular choices for group sequential trials with known information levels. These kinds of parametric tests may not be appropriate if the information levels $\mathcal{I}_1, \ldots, \mathcal{I}_K$ are unpredictable, hence in the upcoming Section 2.1.3 we present a test incorporating varying information levels.

2.1.3 | Error spending tests

This project focuses on survival data where the randomness of event times results in variable numbers of observed events at planned interim analyses. In turn, this results in unpredictable sequences of information levels which motivates using an error spending design. Different group sequential designs give rise to different properties of a trial; some designs are best suited to particular data types and some are designed to have particular properties. For this project, we focus on the "error spending test", a group sequential trial design that uses a flexible method for dealing with unpredictable sequences of information levels.

The underlying idea for an error spending test is to spend error according to the amount of information that has been observed. We focus on a particular method, which requires specifying the maximum target information, which is $\mathcal{I}_{max}$, and unless early stopping occurs, the trial continues until $\mathcal{I}_{max}$ is reached. The value of $\mathcal{I}_{max}$ is chosen to meet a given power requirement and we shall expand on this choice shortly. A trial is designed with K analysis planned and the calculation of $\mathcal{I}_{max}$ is based on a design with K planned analyses. If we observe $\mathcal{I}_K < \mathcal{I}_{max}$, then additional analyses are performed. The trial can be extended by recruiting more patients or extending the trial's duration to increase follow-up time.

The cumulative amount of type 1 error to spend is given by the function $f(\cdot)$ whose input is the fraction of the target information that has been observed. At or before analysis k, there will be type 1 error $f(\mathcal{I}_k/\mathcal{I}_{max})$ spent. Similarly, the function $g(\cdot)$ denotes the cumulative amount of type 2 error that is spent and by analysis k there will be type 2 error $g(\mathcal{I}_k/\mathcal{I}_{max})$ spent. These functions should be chosen to spend error cumulatively with information and should protect the overall type 1 error rate, so that the amount of cumulative type 1 error spent does not exceed α. Therefore, the functions $f(\cdot)$ and $g(\cdot)$ must be such that:

- $f(t)$ and $g(t)$ are non-decreasing in t
- $f(0) = g(0) = 0$

- $f(t) = \alpha$ for $t \geq 1$ and $g(t) = \beta$ for $t \geq 1$.

The ρ-family of functions are an example of error spending functions and are given by:

$$\begin{aligned} f(t) &= \min\{\alpha t^{\rho}, \alpha\} \\ g(t) &= \min\{\beta t^{\rho}, \beta\}. \end{aligned}$$

Throughout this report, when error spending functions are implemented, we shall use the ρ-family with $\rho = 2$.

The sequence of information levels $\mathcal{I}_1, \dots, \mathcal{I}_K$ does not need to be known before commencing the trial. A decision upon early stopping and the null hypothesis can be made at analysis k knowing the values $\mathcal{I}_1, \dots, \mathcal{I}_k$ and this does not depend on the future values $\mathcal{I}_{k+1}, \dots, \mathcal{I}_K$. Therefore, at analysis k it is possible to calculate boundary constants a_k and b_k without knowing what the future information levels will be. At the first interim analysis the boundary constants a_1 and b_1 are calculated to satisfy

$$\begin{aligned} \mathbb{P}_{\theta=0}\{Z_1 > b_1\} &= f(\mathcal{I}_1/\mathcal{I}_{max}) \\ \mathbb{P}_{\theta=\delta}\{Z_1 < a_1\} &= g(\mathcal{I}_1/\mathcal{I}_{max}). \end{aligned}$$

The trial is stopped if either $Z_1 < a_1$ or $Z_1 > b_1$ and continued otherwise. Further, if the trial is terminated at the first interim analysis, H_0 is accepted if $Z_1 < a_1$ and rejected if $Z_1 > b_1$. To preserve type 1 error, the errors allocated to each analysis create a partition of α. Therefore given that $f(\mathcal{I}_k/\mathcal{I}_{max})$ and $g(\mathcal{I}_k/\mathcal{I}_{max})$ are cumulative type 1 and type 2 errors spent up to and including analysis k, the constants a_k and b_k are calculated such that

$$\begin{aligned} \mathbb{P}_{\theta=0}\{a_1 < Z_1 < b_1, \dots, a_{k-1} < Z_{k-1} < b_{k-1}, Z_k > b_k\} &= f(\mathcal{I}_k/\mathcal{I}_{max}) - f(\mathcal{I}_{k-1}/\mathcal{I}_{max}) \\ \mathbb{P}_{\theta=\delta}\{a_1 < Z_1 < b_1, \dots, a_{k-1} < Z_{k-1} < b_{k-1}, Z_k < a_k\} &= g(\mathcal{I}_k/\mathcal{I}_{max}) - g(\mathcal{I}_{k-1}/\mathcal{I}_{max}) \end{aligned}$$

and a decision upon termination and H_0 is made in accordance with the boundary constants.

In the fixed sample trial, the information level $\mathcal{I}_f$ is calculated to satisfy a power requirement. In a similar manner, $\mathcal{I}_{max}$ in an error spending design is calculated to achieve power of $1 - \beta$ under H_A with $\theta = \delta$. During the design stage of the group sequential trial, we may assume that there are K fixed information levels equally spaced so that for $k = 1, \ldots, K$

$$\mathcal{I}_k = \frac{k}{K}\mathcal{I}_{max}.$$

Under this assumption, there is a unique solution for $\mathcal{I}_{max}$ such that the restriction $a_K = b_K$ is satisfied and the design obtains the correct significance level and power.

Having equally spaced information levels and reaching $\mathcal{I}_{max}$ exactly at the final analysis is a design assumption. However, following the error spending design in practice, by placing boundaries to meet overall error rates, is unlikely to result in boundaries meeting at the final analysis and there must be suitable amendments planned to conclude with the correct type 1 error. Overrunning occurs when the final analysis has a higher information level than expected so that $\mathcal{I}_K > \mathcal{I}_{max}$ and solving for a_K and b_K will result in the boundaries crossing and $a_K > b_K$. This may also happen if $\mathcal{I}_{max}$ is reached but information levels are not equally spaced so that $\mathcal{I}_K = \mathcal{I}_{max}$ but $\mathcal{I}_1, \ldots, \mathcal{I}_{K-1}$ are not at $\mathcal{I}_k = \frac{k}{K}\mathcal{I}_{max}$. It is important to retain type 1 error, so a suitable solution is to reduce a_K to match b_K and the trial has power greater than the planned $1 - \beta$. As previously mentioned, the trial continues unless early stopping occurs so that $\mathcal{I}_K \geq \mathcal{I}_{max}$ however some sequences of information levels which are unequally spaced result in $a_K < b_K$. Again, preserving type 1 error is of most importance, so a_K is increased to match b_K which results in a loss of power.

To conclude, error spending tests provide a flexible method for dealing with sequences of unequally spaced information levels that may occur due to unpredictable data types such as survival data with unknown event times. These methods are particularly useful for controlling type 1 error and ensuring type 2 error is close to the design requirements.

2.2 | Introduction to survival analysis

2.2.1 | Time-to-event data

A time-to-event observation measures the amount of time that passes between entry to a study and the event of interest. In many clinical trials, the event of interest is death and when this is the case, we refer to the analysis as survival analysis.

It may be the case that the event of interest is not observed but the time until another event is recorded. For example when a patient in a clinical trial with survival as the primary endpoint leaves the study before death and their follow-up is ceased upon departure. This is recorded as a (right) censored observation and the time recorded is from entering the study until leaving the study. Censored observations still provide useful information about the primary endpoint as it is known that the event time is at least as great as the censored time. For each patient $i = 1, \dots, n$, let F_i be the time-to-failure random variable for patient i and let C_i be the potential censoring time random variable. The random variable $T_i = \min(F_i, C_i)$ is known as the event time and t_i is the observed event time for patient i. Also, the censoring indicator $\delta_i = \mathbb{I}\{F_i \leq C_i\}$ is observed. We shall focus on the case where the random variables F_i and C_i are independent for each patient $i = 1, \dots, n$. This implies that censoring is non-informative. Lagakos (1979) presents other options for right-censored data.

Figure 2.2 shows example data from a sample of patients in a clinical trial where the primary endpoint is survival. Patients arrive with staggered entry to the study and are assigned to either the control or the new treatment. The patients' event times are either exact, $\delta_i = 1$, or censored, $\delta_i = 0$.

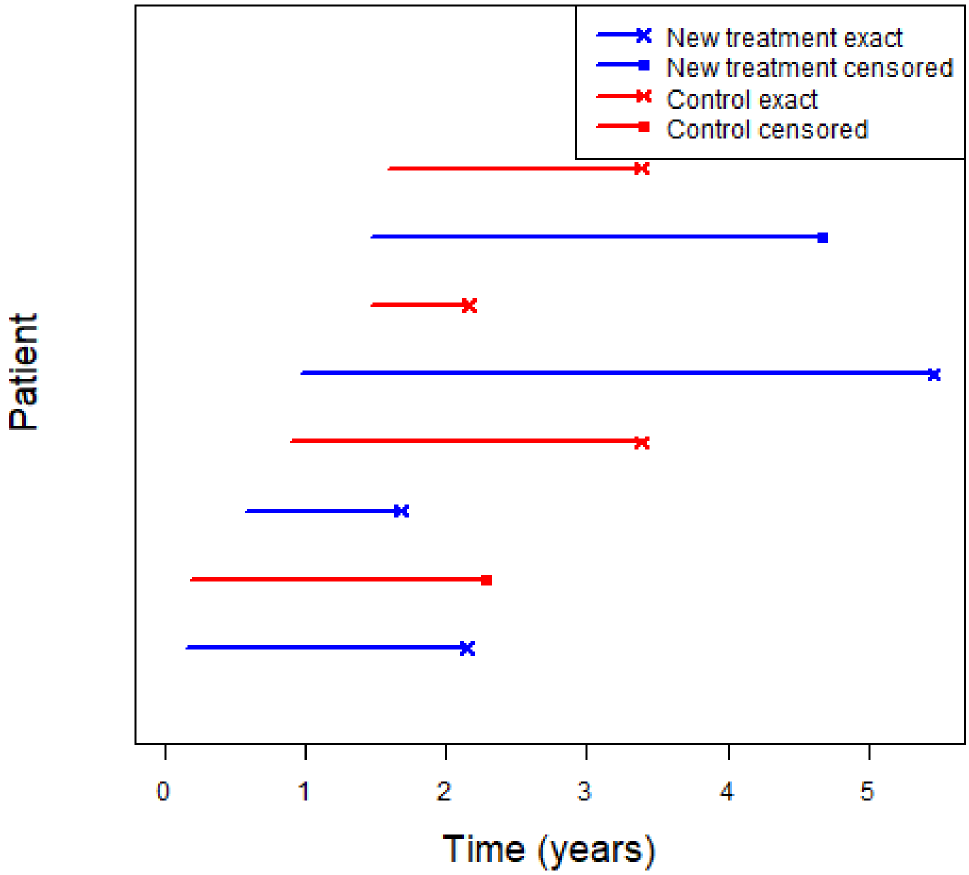


Figure 2.2: Example of a clinical trial using survival data.

The framework for dealing with censored observations is useful in group sequential trials as patients that have survived past an interim analysis will be temporarily marked as censored. Figure 2.3 shows the data from a group sequential trial at the interim analysis which occurs at calendar time 3 years. For the patients who are still alive and not yet censored, their interim observation is marked as censored at 3 years.

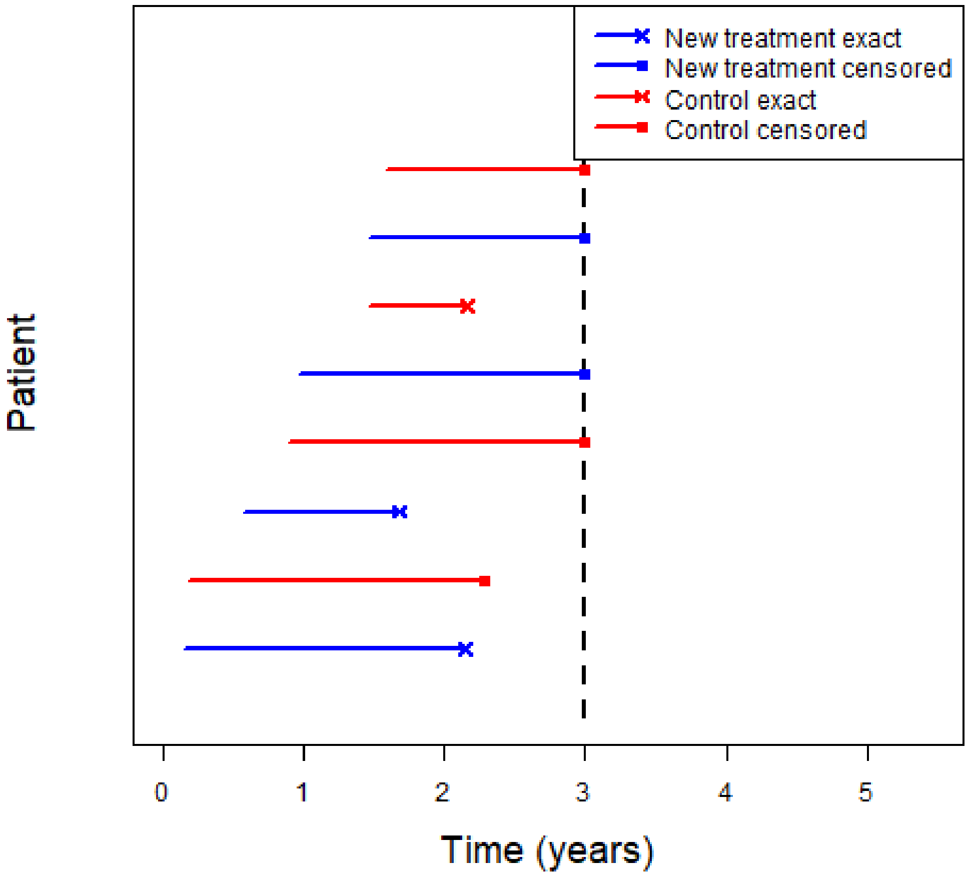


Figure 2.3: Interim analysis for a group sequential clinical trial using survival data.

2.2.2 | Survival functions

In many time-to-event studies, the object of primary interest is called the survival function. With F_i as the time-to-event random variable for patient i, the survival function is defined as

$$S_i(t) = \mathbb{P}(F_i > t).$$

It is often of interest to compare the survival curves for patients on different treatment arms. Let $Z_i = \mathbb{I}(\text{patient } i \text{ is on new treatment})$ be the indicator that patient i receives the experimental drug, then we may wish to compare $S_i(t|Z_i = 1)$ and $S_i(t|Z_i = 0)$. Figure 2.4 shows an example of the survival curves. In this example it is clear that the treatment is working effectively as $S_i(t|Z_i = 1) > S_i(t|Z_i = 0)$ for all values of t.

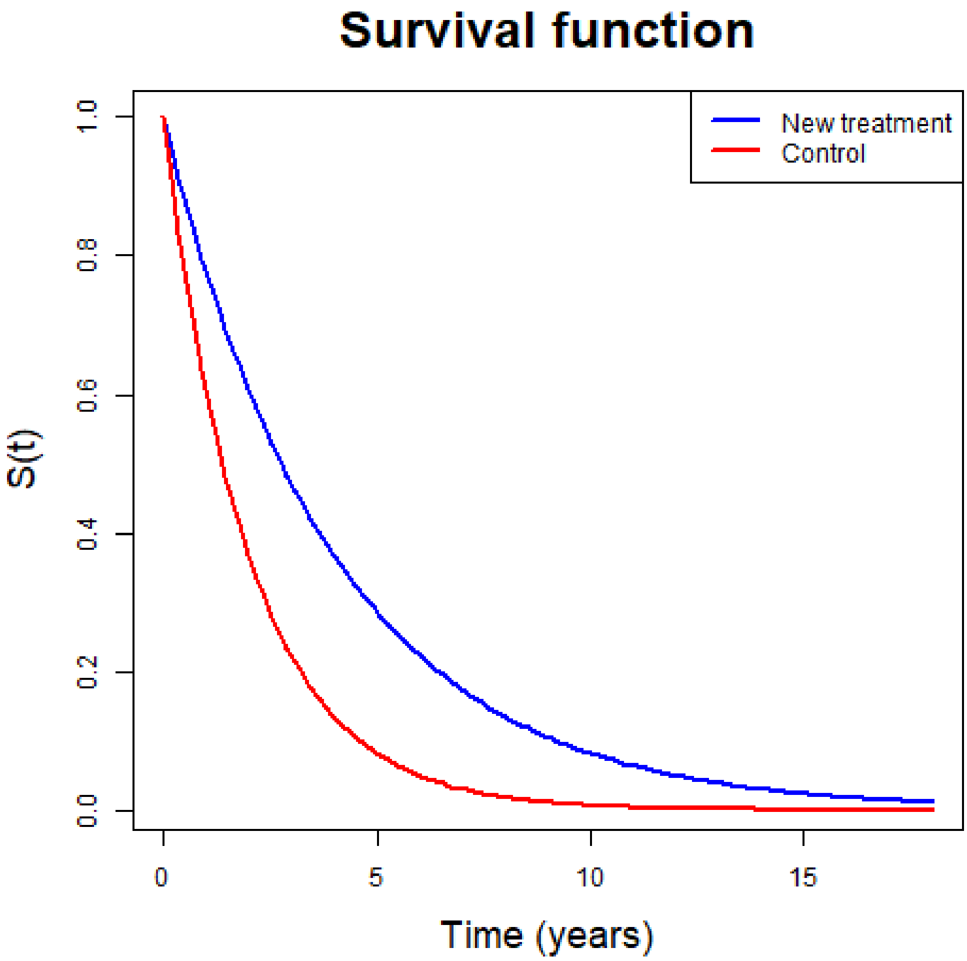


Figure 2.4: Survival functions.

The survival function is a smooth function of time t. In general, the true survival function will not be known. We can use the observed data to estimate survival functions. The Kaplan-Meier estimator, introduced by Kaplan and Meier (1958), estimates the survival probability at time t by considering the proportion of patients that have not yet had an event or been censored at time t. Let $t'_1, \ldots, t'_m$ be the observed event times (whether censored or observed) in order of increasing magnitude. Let d_i be the number of deaths at time t'_i and let r_i be the number of patients at risk just before time t'_i. Then the Kaplan-Meier estimator is defined as

$$\hat{S}(t) = \prod_{i:t'_i \leq t} \frac{r_i - d_i}{r_i}.$$

The Kaplan-Meier estimator for a data set with $n = 50$ patients on each treatment arm is shown in Figure 2.5 below.

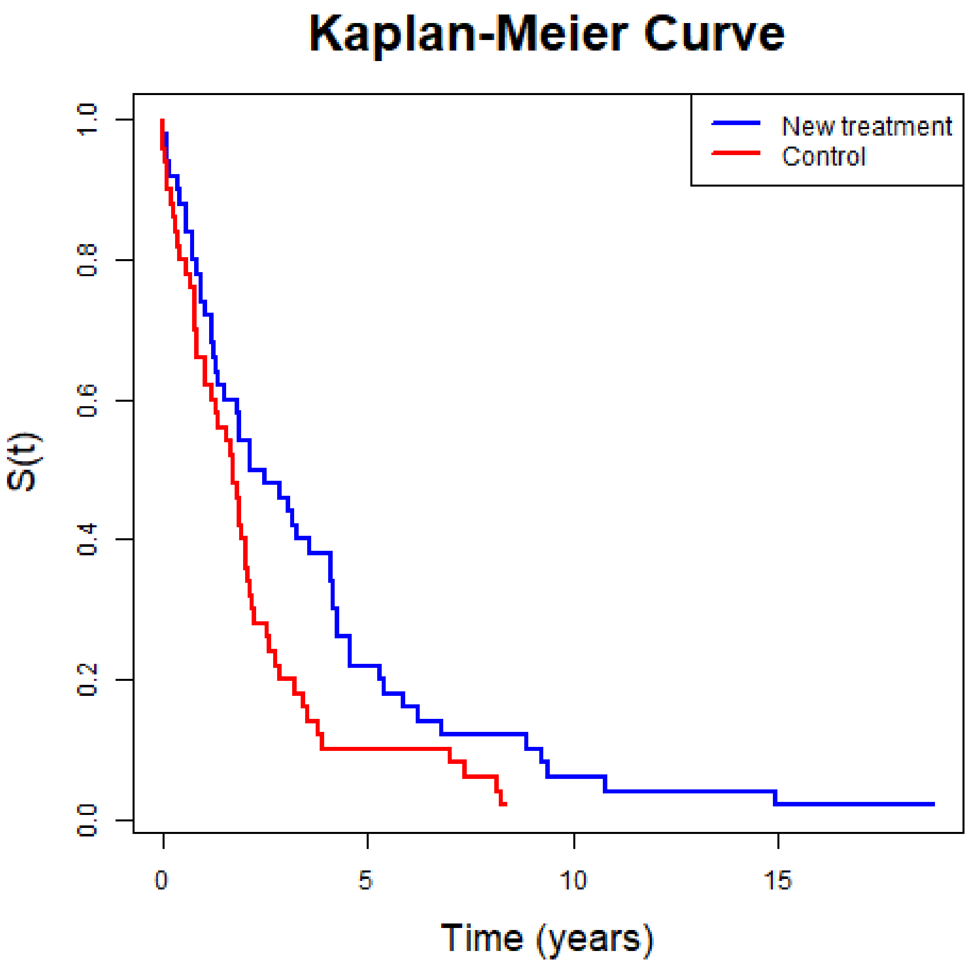


Figure 2.5: Kaplan-Meier estimator.

To model time-to-event data, it is necessary to associate a probability density to the event at any given point in time. Since time is treated as a continuous variable, this probability is for an instantaneous moment in time and the information is summarised by a hazard function. For each patient $i = 1, \dots, n$, let F_i be the time-to-failure random variable, then the hazard for individual i at time t is defined as

$$h_i(t) = lim_{\delta t \downarrow 0} \frac{\mathbb{P}(t \leq F_i < t + \delta t | F_i \geq t)}{\delta t}. \tag{2.2}$$

We can specify specific forms for the hazard rate. For example, we may wish to include covariates that are known to affect survival in the model. Let $\mathbf{x}_i$ be a $p \times 1$ vector of covariates for patient i, let $\boldsymbol{\theta}$ be a $p \times 1$ vector of coefficients and let $h_0(t)$ be a baseline hazard function, then we may wish to specify that

$$h_i(t) = h_0(t) \exp\{\boldsymbol{\theta}^T \mathbf{x}_i\}.$$

This particular model is called the Cox proportional hazards model. This will be explored in Section 3.2.1.

The cumulative hazard function for patient i is defined as $H_i(t) = \int_0^t h_i(u)du$ and the relationship between the hazard rate and the survival function is given by

$$S_i(t) = \exp\{-H_i(t)\}.$$

This relationship is useful for simulating survival data. Note that $S_i(t) = \mathbb{P}(F_i > t)$, so the cumulative distribution of the time-to-event random variable F_i is therefore $\mathbb{P}(F_i \leq t) = 1 - S_i(t)$. Therefore, using the inverse transform theorem, let $u_i \sim U(0,1)$ be sampled from a uniform distribution, then the time-to-event observation F_i is the solution to

$$H_i(F_i) = -\log(1 - u_i).$$

For each of our examples, C_i and F_i are independent. Therefore, these values can be simulated independently and then the value $T_i = \min\{F_i, C_i\}$ is observed.

The survival function can also be used to find the likelihood function of the survival data. For uncensored data, where $T_i = F_i$, the contribution of patient i to the likelihood is given by $d/dt(1 - S_i(t))$, which is the derivative of the cumulative hazard function with respect to t. For censored data, the contribution is $S_i(t_i)$. Suppose that each of these functions depends on a vector of parameters $\boldsymbol{\theta}$, and that the model is specified through the hazard rate $h_i(t, \boldsymbol{\theta})$. Let $t_1, \dots, t_n$ be the observed censored or exact event times and $\delta_1, \dots, \delta_n$ be the observed censoring indicators, then the likelihood function is defined as

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} f_i(t_i, \boldsymbol{\theta})^{\delta_i} S_i(t_i, \boldsymbol{\theta})^{1-\delta_i}.$$

2.3 | Longitudinal data analysis

2.3.1 | Random effects models

Longitudinal data is comprised of repeated measurements of the same variable at different points in time. In a clinical trial, this is particularly useful for tracking trends and changes in patient covariates and risk factors over time. For example, levels of the biomarker, prostate specific antigen (PSA), are monitored in patients with prostate cancer at different points in time using blood tests. In this Section, we shall discuss methods for dealing with longitudinal data, specifically random effects models which allow for within-patient correlations.

Let $W(t)$ be the measurement of the biomarker at time t and let $\epsilon(t)$ be the measurement error at time t. Then a longitudinal data model is given by

$$W(t) = \boldsymbol{\beta}^T \boldsymbol{\rho}(t) + \epsilon(t)$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is a $p \times 1$ vector of coefficients and $\boldsymbol{\rho}(t)$ is a $p \times 1$ vector of functions in t. In general, $\boldsymbol{\rho}(t)$ is not constrained to linear functions in t. The simple example which we shall use throughout this report is the case

$$W(t) = \beta_0 + \beta_1 t + \epsilon(t).$$

Therefore $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ and $\boldsymbol{\rho}(t) = (1, t)^T$.

For model fitting, it is often necessary to assume a distribution for the measurement error. Suppose that the longitudinal observations are measured at times $t_1, \ldots, t_m$, then it is a common assumption that $\epsilon(t_j) \sim N(0, \sigma^2)$ for $j = 1, \ldots, m$ and that $\epsilon(t_k)$ and $\epsilon(t_l)$ are independent for $k \neq l$.

A random effects model is a statistical model where the parameters are random variables. This is useful in a clinical trial as we can model each patient's trajectory separately. Therefore, the global trend in risk factors can be studied while accounting for the correlation of the repeated measurements.

For patients $i = 1, \ldots, n$ let $W_i(t)$ be the biomarker measurement at time t for patient i. The random effects model is given by

$$W_i(t) = b_{i0} + b_{i1} t + \epsilon_i(t).$$

Let $\mathbf{b}_i = (b_{i0}, b_{i1})^T$ be a vector of patient-specific random effects and let $\boldsymbol{\rho}(t) = (1, t)^T$ be a vector of functions in t. Again, in general, $\mathbf{b}_i$ and $\boldsymbol{\rho}(t)$ are vectors of length p

and more general functions of t are possible for $\boldsymbol{\rho}(t)$.

In this model, the random effects are random variables with $\mathbf{b}_1, \ldots, \mathbf{b}_n$ distributed according to the density function $f_{\boldsymbol{b}}(\cdot)$. Similarly to the simple longitudinal data model, it is sometimes useful or necessary to assume a distribution for both the measurement error and the random effects, e.g make an assumption about the form of $f_{\boldsymbol{b}}(\cdot)$. The following are commonly imposed assumptions. Let $\boldsymbol{\mu}$ be a 2×1 vector for the mean of the random effects and let Σ be a 2×2 symmetric matrix for the variance matrix of the random effects. Then we suppose that, for each patient $i = 1, \ldots, n$, the biomarker is measured at times $t_{i1}, \ldots, t_{im_i}$ and

$$\begin{aligned} \mathbf{b}_i &\overset{i.i.d}{\sim} N(\boldsymbol{\mu}, \Sigma) \text{ for } i = 1, \ldots, n \\ \epsilon(t_{ij})|\mathbf{b}_i &\overset{i.i.d}{\sim} N(0, \sigma^2) \text{ for } j = 1, \ldots, m_i. \end{aligned}$$

These assumptions specify that the distribution of the measurement error is common across all patients and time, and that the random effects are independent and identically distributed normal random variables. There are two methods considered in this thesis for analysing joint models. The first method in Section 4.1 does not require any distributional assumptions about the random effects, however for the second model in Section 5.2 we must assume a distribution for the random effects and have chosen this to be a normal distribution.

We now present a likelihood function for the random effects model. For each patient $i = 1, \ldots, n$, biomarker measurements $W_i(t_{i1}), \ldots, W_i(t_{im_i})$ are observed at times $t_{i1}, \ldots, t_{im_i}$. Let $\boldsymbol{\theta}$ be a vector of all parameters in the model, including parameters in the distribution function $f_{\mathbf{b}}(\cdot)$. Then the contribution to the likelihood from patient i can be found by integrating over the random effects, this is

$$\mathcal{L}_i(\boldsymbol{\theta}) = \prod_{j=1}^{m_i} \int \int \mathbb{P}(W_j(t_{ij})|\mathbf{b}_i, \boldsymbol{\theta}) f_{\boldsymbol{b}}(\mathbf{b}_i|\boldsymbol{\theta})\, db_{i0}\, db_{i1}.$$

The full likelihood function is then

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \mathcal{L}_i(\boldsymbol{\theta}). \tag{2.3}$$

2.3.2 | Gauss-Hermite integration

Gauss-Hermite Quadrature is a numerical method for evaluating definite integrals. This numerical method is particularly useful for integration of functions which include the density function of Gaussian random variables. Hence, this quadrature rule will prove particularly useful for evaluating the likelihood function of models which include normally distributed random effects, such as in Equation (2.3). Further, in Section 5.1 we introduce a function called the restricted mean survival time and this function involves integrating over normally distributed random effects.

Gauss-Hermite integration is specifically for evaluating functions of the form

$$\int_{-\infty}^{\infty} e^{-x^2/2} g(x) dx$$

where $g(\cdot)$ is a smooth deterministic function. It is clear that when we wish to integrate over normally distributed random effects then we will have a function of the above form since the density function includes the term $e^{-x^2/2}$.

The quadrature rule approximates the integral by the sum

$$\sum_{i=1}^{k} w_i g(\zeta_i)$$

with nodes $\zeta_1, \ldots, \zeta_k$ and weights $w_1, \ldots, w_k$. Liu and Pierce (1994) give details of this calculation. The nodes are given by the roots of the Hermite polynomial function, and the weights also include the Hermite polynomial function. The nodes and weights can be calculated using computer programs or by looking these up in standard tables such as in Shao et al. (1964).

The advantage of this quadrature rule is that calculation of this integral is computationally efficient. When $g(\cdot)$ is a polynomial function of degree $2k - 1$ or less, the Gaussian quadrature rule is exact. Further, for an accurate calculation, the number of nodes required is usually very small. To demonstrate this, consider approximating the double integral

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left\{\frac{b_0}{b_1}\right\} f(b_0, b_1) db_0 db_1 \tag{2.4}$$

where $f(\cdot)$ is the density function of the random variable

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} \sim N\left(\begin{bmatrix} 1 \\ 3 \end{bmatrix}, \begin{bmatrix} 1.2^2 & 0 \\ 0 & 0.25^2 \end{bmatrix}\right).$$

This example represents a simplified version of the restricted mean survival time calculation that we shall see in Section 5.1.

Table 2.1 shows the calculation of Equation (2.4) using Gauss-Hermite integration with differing numbers of nodes k. The Gauss-Hermite quadrature rule for this double integral example is accurate in the 6^{th} decimal place with $k = 4$ nodes. This approximation is much more accurate than calculating the integral by Monte Carlo simulation and is also computationally efficient.

k	Gauss-Hermite integral
2	1.519241
3	1.519439
4	1.519447
5	1.519447
6	1.519447

Table 2.1: Gauss-Hermite quadrature rule example for different numbers of nodes.

The example integral in Equation (2.4) has a similar structure to the likelihood function for a random effects model in Equation (2.3). Both equations include double integrals over the probability density function for bivariate normal random effects. The likelihood function for a random effects model can be evaluated using Gauss-Hermite integration. This is useful because we need an efficient integration method as there are n double integrals in one calculation of the likelihood. Further, maximum likelihood estimation requires a sequence of such likelihood evaluations.

Note that for this example, the random variables b_0 and b_1 are independent so the double integral can be evaluated by performing the Gauss-Hermite rule on each dimension separately. In the more general setting, bivariate normal random variables can be transformed to the independent case. Consider random variables X_1 and X_2 which have the bivariate normal distribution

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho \\ \rho & \sigma_2^2 \end{bmatrix}\right).$$

Let

$$\begin{aligned} Y_1 &= X_1 \\ Y_2 &= X_1 - \frac{\sigma_1^2}{\rho} X_2. \end{aligned}$$

Then it can be shown that Y_1 and Y_2 are independent and their joint distribution is given by

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_1 - \frac{\sigma_1^2}{\rho}\mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_1^2(\frac{\sigma_1^2\sigma_2^2}{\rho^2} - 1) \end{bmatrix}\right).$$

The Gauss-Hermite integral can be manipulated for correlated normal random variables by creating independent normal random variables.

CHAPTER 3

SURVIVAL ANALYSIS

3.1 | General asymptotic theory

In this section, we introduce and prove a key result in asymptotic distribution theory. The result concerns an estimator which is the root of an "estimating equation", in which a function with expectation zero is set equal to zero. The theorem states that such an estimator is asymptotically normally distributed. Many estimators, such as the maximum likelihood estimate, fit naturally into this category of estimator. Alternatively, if asymptotic normality is thought to be a desirable property, one may construct an estimator to be the solution to an estimating equation. This theorem will be used in Section 3.2 to derive asymptotic normality of the estimator that arises from analysing survival data using partial likelihood. The theorem is used again in Section 4.1 for analysing the joint model. Therefore, the aim of this section is to present and prove asymptotic normality for a broad class of estimators. In doing so, some conditions are presented which are sufficient to prove asymptotic normality of the estimators in later sections.

This general asymptotic theory is well known throughout frequentist statistics. The proof we present was largely developed using Section 2.3 of the book "Bayesian and Frequentist Regression Methods" by Wakefield (2013), Section 9.2 of the book "Theoretical Statistics" by Cox and Hinkley (1979) and Section 4 of Jennison and Turnbull (1997). In Section 2.3 of the book, Wakefield introduces an estimating equation which is the formal definition for an equation where a function with expectation zero is set equal to zero. The author gives an outline derivation of the proof of our Theorem 3.1 in one dimension. We have filled in some details, particularly for multiple dimensions. In Section 2.9 of "Theoretical Statistics", Cox and Hinkley prove that the maximum likelihood estimate is consistent and asymptotically normal. We have adapted this proof to a broader class of estimators, which are solutions to estimating equations. The regularity conditions presented by Cox and Hinkley are given for maximum likelihood estimation but we shall use these conditions with the likelihood function $f_Y(y;\theta)$ replaced by a more general function $\mathbf{G}(\boldsymbol{\theta},\mathbf{x})$. Jennison and Turnbull (1997) prove asymptotic normality of estimates in a general parametric regression model in a group sequential trial. Here, the parameter $\boldsymbol{\theta}$ is multidimensional and the result concerns the joint distribution of estimates $\hat{\boldsymbol{\theta}}_n^{(1)}, \dots, \hat{\boldsymbol{\theta}}_n^{(K)}$ across analyses in the group sequential trial. We shall follow similar theory and present the corresponding result for group sequential trials where the sequence of estimates are solutions to estimating equations.

In this Chapter, we shall begin with some asymptotic distribution theory for general statistical models in Section 3.1. In Section 3.1, we present the theoretical results for a fixed sample trial and then extend this result to group sequential trials. We aim to apply these distributional results to both survival models and the joint model. We are then equipped to present the asymptotic distributional results for survival data in Section 3.2 for both a fixed sample trial and a GST.

Some notation is needed before presenting and proving Theorem 3.1. Suppose that $X_1, \ldots, X_n$ are random variables and that $x_1, \ldots, x_n$ are observations of these random variables. The collections of these objects are labelled as $\mathbf{X}_n = (X_1, \ldots, X_n)$ and $\mathbf{x}_n = (x_1, \ldots, x_n)$ respectively. Let $\boldsymbol{\theta}$ be a $p \times 1$ column vector parameter in a statistical model and let $\boldsymbol{\theta_0}$ be the true value of the parameter. The parameter space will therefore be $\Theta \subset \mathbb{R}^p$. The parameter estimate based on $x_1, \ldots, x_n$ will be denoted by $\hat{\boldsymbol{\theta}}_n$ to show dependence on the number of observations and to highlight that the limits are defined as $n \to \infty$. An "estimating function" is a $p \times 1$ column vector

$$\mathbf{G}_n(\boldsymbol{\theta}, \mathbf{x}_n) = \sum_{i=1}^{n} \mathbf{G}(\boldsymbol{\theta}, x_i) = \sum_{i=1}^{n} \begin{pmatrix} G^1(\boldsymbol{\theta}, x_i) \\ \vdots \\ G^p(\boldsymbol{\theta}, x_i) \end{pmatrix}$$

such that $\mathbb{E}_{\boldsymbol{\theta}}(\mathbf{G}_n(\boldsymbol{\theta}, \mathbf{X}_n)) = \mathbf{0}$ for all $\boldsymbol{\theta}$. The "estimating equation" is then a set of p equations given by

$$\mathbf{G}_n(\hat{\boldsymbol{\theta}}_n, \mathbf{x}_n) = \mathbf{0}.$$

If there are multiple roots to the above estimating equation, we suppose estimates $\hat{\boldsymbol{\theta}}_n, n = 1, 2, \ldots,$ are chosen such that the sequence $\{\hat{\boldsymbol{\theta}}_n\}$ is consistent and Theorem 3.1 shall be applied to this particular consistent sequence.

In applying Theorem 3.1, we shall assume that the following conditions are satisfied.

Conditions 3.1.

1. *$\hat{\boldsymbol{\theta}}_n$ is a consistent estimator for $\boldsymbol{\theta_0}$, that is as $n \to \infty$, $\hat{\boldsymbol{\theta}}_n$ converges in probability to $\boldsymbol{\theta_0}$, written*

$$\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta_0}$$

2. *$n^{-\frac{1}{2}}\mathbf{G}_n(\boldsymbol{\theta_0}, \mathbf{X}_n)$ converges in distribution to a zero-mean Gaussian random variable with finite-valued, positive-definite covariance matrix $\mathbf{B}$, specifically*

$$n^{-\frac{1}{2}}\mathbf{G}_n(\boldsymbol{\theta_0}, \mathbf{X}_n) \xrightarrow{d} N(\mathbf{0}, \mathbf{B})$$

3. *For all $\boldsymbol{\theta}_n^*$ such that $\boldsymbol{\theta}_n^* \xrightarrow{p} \boldsymbol{\theta_0}$, $n^{-1}\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{G}_n(\boldsymbol{\theta}, \mathbf{x}_n)\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_n^*}$ converges in probability*

to a finite-valued, positive-definite matrix $\mathbf{A}$. *That is*

$$n^{-1}\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{G}_n(\boldsymbol{\theta},\mathbf{x}_n)\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_n^*}\xrightarrow{p}\mathbf{A}$$

4. *The parameter space* Θ *is closed and compact and the true parameter value* $\boldsymbol{\theta_0}$ *is an interior point of* Θ

5. *For all* $\mathbf{x}_n$, *the derivatives* $\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{G}_n(\boldsymbol{\theta},\mathbf{x}_n)$ *and* $\frac{\partial^2}{\partial\boldsymbol{\theta}^2}\mathbf{G}_n(\boldsymbol{\theta},\mathbf{x}_n)$ *exist in a neighbourhood of* $\boldsymbol{\theta_0}$

6. $n^{-1}\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{G}_n(\boldsymbol{\theta},\mathbf{x}_n)$ *is uniformly continuous in* $\boldsymbol{\theta}$ *in a neighbourhood of* $\boldsymbol{\theta_0}$.

In Section 3.2 and Section 4.1 where we apply Theorem 3.1, the conditions 1–3 will be checked. Conditions 4–6 will be assumed to hold in order to avoid technical distractions during the proof of Theorem 3.1. We can now prove that the solution to an estimating equation is asymptotically normally distributed.

Theorem 3.1. *Let the estimate* $\hat{\boldsymbol{\theta}}_n$ *be a solution to the equation* $\mathbf{G}_n(\boldsymbol{\theta},\mathbf{x}_n)=\mathbf{0}$ *and let* $\boldsymbol{\theta_0}$ *be the true value of the parameter* $\boldsymbol{\theta}$. *Suppose that Conditions 3.1 hold. Then* $n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta_0})$ *is asymptotically normal, converging in distribution to a Gaussian random variable, given by*

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta_0})\xrightarrow{d}N(\mathbf{0},\mathbf{A}^{-1}\mathbf{B}(\mathbf{A}^{-1})^T).$$

Proof. To derive the asymptotic distribution of $n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta_0})$ we begin by applying a Taylor expansion to each element of $\mathbf{G}_n(\boldsymbol{\theta},\mathbf{x}_n)$. Let $G_n^j(\boldsymbol{\theta},\mathbf{x}_n)$ be the j^{th} element of the $p\times 1$ column vector $\mathbf{G}_n(\boldsymbol{\theta},\mathbf{x}_n)$. Then, $\frac{\partial}{\partial\boldsymbol{\theta}}G_n^j(\boldsymbol{\theta},\mathbf{x}_n)$ is a $1\times p$ row vector and $\boldsymbol{\theta}-\boldsymbol{\theta_0}$ is a $p\times 1$ column vector. Regularity conditions 4 and 5 allow us to perform the following Taylor expansion. For each $j=1,\ldots,p$ the Taylor expansion of $G_n^j(\boldsymbol{\theta},\mathbf{x}_n)$ around $\boldsymbol{\theta_0}$ is

$$G_n^j(\tilde{\boldsymbol{\theta}},\mathbf{x}_n)=G_n^j(\boldsymbol{\theta_0},\mathbf{x}_n)+\frac{\partial}{\partial\boldsymbol{\theta}}G_n^j(\boldsymbol{\theta},\mathbf{x}_n)\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{n,j}^*}\cdot(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta_0})$$

where $\boldsymbol{\theta}_{n,j}^*$ lies on the line segment between $\boldsymbol{\theta_0}$ and $\tilde{\boldsymbol{\theta}}$. Each row of $\mathbf{G}_n(\boldsymbol{\theta},\mathbf{x}_n)$ represents a dimension, so different rows are differing functions of $\boldsymbol{\theta}$ and hence each $\boldsymbol{\theta}_{n,j}^*$ is specific to the row j.

Given the definition of the estimate $\hat{\boldsymbol{\theta}}_n$, we recognise that $\mathbf{G}_n(\hat{\boldsymbol{\theta}}_n,\mathbf{x}_n)=\mathbf{0}$.

Therefore by substituting $\hat{\boldsymbol{\theta}}_n$ for $\boldsymbol{\theta}$, and stacking the rows, we have

$$\mathbf{0} = \begin{bmatrix} G_n^1(\hat{\boldsymbol{\theta}}_n, \mathbf{x}_n) \\ \vdots \\ G_n^p(\hat{\boldsymbol{\theta}}_n, \mathbf{x}_n) \end{bmatrix} = \mathbf{G}_n(\boldsymbol{\theta}_\mathbf{0}, \mathbf{x}_n) + \begin{bmatrix} \frac{\partial}{\partial \boldsymbol{\theta}} G_n^1(\boldsymbol{\theta}, \mathbf{x}_n)\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*_{n,1}} \\ \vdots \\ \frac{\partial}{\partial \boldsymbol{\theta}} G_n^p(\boldsymbol{\theta}, \mathbf{x}_n)\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*_{n,p}} \end{bmatrix} \cdot (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_\mathbf{0})$$

where each $\boldsymbol{\theta}^*_{n,j}$ lies on the line segment between $\boldsymbol{\theta}_\mathbf{0}$ and $\hat{\boldsymbol{\theta}}_n$.

The matrix $\mathbf{A}$ by definition is finite valued, symmetric and positive definite. Therefore $\mathbf{A}$ is invertible with inverse $\mathbf{A}^{-1}$ which is also finite valued, symmetric and positive definite. Multiplying the above equation by $\mathbf{A}^{-1}$ and a simple rearrangement gives

$$-n^{-\frac{1}{2}} \mathbf{A}^{-1} \mathbf{G}_n(\boldsymbol{\theta}_\mathbf{0}, \mathbf{x}) = \mathbf{A}^{-1} \begin{bmatrix} n^{-1} \frac{\partial}{\partial \boldsymbol{\theta}} G_n^1(\boldsymbol{\theta}, \mathbf{x})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*_{n,1}} \\ \vdots \\ n^{-1} \frac{\partial}{\partial \boldsymbol{\theta}} G_n^p(\boldsymbol{\theta}, \mathbf{x})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*_{n,p}} \end{bmatrix} n^{\frac{1}{2}} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_\mathbf{0}). \tag{3.1}$$

Let

$$\begin{aligned} \mathbf{W}_n &= -n^{-\frac{1}{2}} \mathbf{A}^{-1} \mathbf{G}_n(\boldsymbol{\theta}_\mathbf{0}, \mathbf{x}) \\ \mathbf{Y}_n &= \mathbf{A}^{-1} \begin{bmatrix} n^{-1} \frac{\partial}{\partial \boldsymbol{\theta}} G_n^1(\boldsymbol{\theta}, \mathbf{x})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*_{n,1}} \\ \vdots \\ n^{-1} \frac{\partial}{\partial \boldsymbol{\theta}} G_n^p(\boldsymbol{\theta}, \mathbf{x})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*_{n,p}} \end{bmatrix} \\ \mathbf{Z}_n &= n^{\frac{1}{2}} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_\mathbf{0}). \end{aligned}$$

Then Equation (3.1) becomes

$$\mathbf{W}_n = \mathbf{Y}_n \mathbf{Z}_n. \tag{3.2}$$

We now consider the limiting distribution of the objects $\mathbf{W}_n$, $\mathbf{Y}_n$ as $n \to \infty$ in order to determine the limiting distribution of $\mathbf{Z}_n$. We must also consider whether or not the inverse of $\mathbf{Y}_n$ exists and and the limiting distribution of this matrix.

For $\mathbf{W}_n$, regularity condition 2 states that $n^{-\frac{1}{2}} \mathbf{G}_n(\boldsymbol{\theta}_\mathbf{0}, \mathbf{x}_n) \xrightarrow{d} N(\mathbf{0}, \mathbf{B})$ and multiplication by a finite matrix implies

$$\mathbf{W}_n = -n^{-\frac{1}{2}} \mathbf{A}^{-1} \mathbf{G}_n(\boldsymbol{\theta}_\mathbf{0}, \mathbf{x}_n) \xrightarrow{d} N(\mathbf{0}, \mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-1}).$$

Given that $\hat{\boldsymbol{\theta}}_n$ is a consistent estimator for $\boldsymbol{\theta}_\mathbf{0}$ by regularity condition 1, $n^{-1} \frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{G}_n(\boldsymbol{\theta}, \mathbf{x}_n)$ is uniformly continuous in $\boldsymbol{\theta}$ by regularity condition 6 and $\boldsymbol{\theta}^*_{n,j}$ lies on the line segment between $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}_n$, the difference between $\boldsymbol{\theta}^*_{n,j}$ and $\boldsymbol{\theta}_\mathbf{0}$ for each $j = 1, \ldots, p$ is asymptotically negligible. Let A^j denote row j of the matrix $\mathbf{A}$, then by regularity condition 3,

$$n^{-1}\frac{\partial}{\partial\boldsymbol{\theta}}G_n^j(\boldsymbol{\theta},\mathbf{x}_n)\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*_{n,j}} \xrightarrow{p} n^{-1}\frac{\partial}{\partial\boldsymbol{\theta}}G_n^j(\boldsymbol{\theta},\mathbf{x}_n)\big|_{\boldsymbol{\theta}=\boldsymbol{\theta_0}} = A^j \text{ for each } j = 1, \ldots, p.$$

Hence

$$\begin{bmatrix} n^{-1}\frac{\partial}{\partial\boldsymbol{\theta}}G_n^1(\boldsymbol{\theta},\mathbf{x})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*_{n,1}} \\ \vdots \\ n^{-1}\frac{\partial}{\partial\boldsymbol{\theta}}G_n^p(\boldsymbol{\theta},\mathbf{x})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*_{n,p}} \end{bmatrix} \xrightarrow{p} \begin{bmatrix} A^1 \\ \vdots \\ A^p \end{bmatrix} = \mathbf{A}$$

and since $\mathbf{A}^{-1}$ is finite-valued, we can conclude that

$$\mathbf{Y}_n \xrightarrow{p} \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_p$$

where $\mathbf{I}_p$ is the $p \times p$ identity matrix.

The limiting distribution of $\mathbf{Z}_n$ can be determined by considering the cases when $\mathbf{Y}_n^{-1}$ exists and when $\mathbf{Y}_n^{-1}$ does not exist. To do so, define a new matrix

$$\mathbf{C}_n = \begin{cases} \mathbf{Y}_n^{-1} & \text{if } \mathbf{Y}_n^{-1} \text{ exists} \\ \mathbf{I}_p & \text{otherwise} \end{cases}$$

We now show that $\mathbf{C}_n \xrightarrow{p} \mathbf{I}_p$. Note that since $\mathbf{Y}_n \xrightarrow{p} \mathbf{I}_p$, as $n \to \infty$ then the determinant $|\mathbf{Y}_n| \xrightarrow{p} 1$ as $n \to \infty$ and hence $\mathbb{P}(\mathbf{Y}_n^{-1} \text{ exists}) \to 1$ as $n \to \infty$. Given small $\epsilon > 0$, we can find a K such that $\|\mathbf{Y}_n - \mathbf{I}_p\| < \frac{\epsilon}{K}$ implies that $\mathbf{Y}_n^{-1}$ exists and $\|\mathbf{Y}_n^{-1} - \mathbf{I}_p\| < \epsilon$. Now take n_0 such that if $n > n_0$ then $\mathbb{P}(\|\mathbf{Y}_n - \mathbf{I}_p\| > \frac{\epsilon}{K}) < \epsilon$. Then, for $n > n_0$, $\mathbb{P}(\mathbf{Y}_n^{-1} \text{ exists and } \|\mathbf{Y}_n^{-1} - \mathbf{I}_p\| < \epsilon) > 1 - \epsilon$ and this implies that $\mathbf{C}_n \xrightarrow{p} \mathbf{I}_p$.

Now we have

$$\mathbf{Z}_n = \begin{cases} \mathbf{C}_n\mathbf{W}_n & \text{if } \mathbf{Y}_n^{-1} \text{ exists} \\ \mathbf{Z}_n & \text{otherwise} \end{cases} \tag{3.3}$$

To determine the limiting distribution of $\mathbf{Z}_n$, we shall use the Continuous Mapping Theorem. Given that $\mathbf{C}_n \xrightarrow{p} \mathbf{I}_p$ and $\mathbf{W}_n \xrightarrow{d} N(\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1})$, then $\mathbf{C}_n\mathbf{W}_n \xrightarrow{d} N(\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1})$. This, together with definition (3.3) and the fact that $\mathbb{P}(\mathbf{Y}_n^{-1} \text{ exists}) \to 1$ as $n \to \infty$ gives that

$$\mathbf{Z}_n \xrightarrow{d} N(\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}).$$

Finally, by the definition of $\mathbf{Z}_n$, we have

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_\mathbf{0}) \xrightarrow{d} N(\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}).$$

□

We shall now display the corresponding result for the group sequential trial (GST) version. To do so, we define the data that is available at the analyses of the GST and adapt the regularity conditions to take into account that there will be a parameter estimate at each analysis.

Suppose that we wish to perform a GST with K analyses and that the trial is planned with interim analyses occurring at times $\tau_1, \dots, \tau_K$. Define the random variable $\mathbf{X}_n^{(k)} = (X_1^{(k)}, \dots, X_n^{(k)})$ where $X_n^{(k)}$ is the data available at time τ_k. The value n represents the total recruited sample size, which remains constant across analyses, and we aim to determine the asymptotic distribution as $n \to \infty$. The times $\tau_1, \dots, \tau_K$ are fixed and the rate of recruitment is proportional to n, so the number of observations at each analysis increases with n.

For each analysis $k = 1, \dots, K$, the same method is employed to estimate the $p \times 1$ vector of parameters $\boldsymbol{\theta}$ of the statistical model, but using different sets of data. For each $k = 1, \dots, K$, let $\hat{\boldsymbol{\theta}}_n^{(k)}$ be the solution to the estimating equation $\mathbf{G}_n(\boldsymbol{\theta}, \mathbf{x}_n^{(k)}) = \mathbf{0}$. We therefore have that the $pK \times 1$ vector $(\hat{\boldsymbol{\theta}}_n^{(1)T}, \dots, \hat{\boldsymbol{\theta}}_n^{(K)T})^T$ is the solution to the set of pK equations

$$\begin{bmatrix} \mathbf{G}\left(\boldsymbol{\theta}, \mathbf{x}_n^{(1)}\right) \\ \vdots \\ \mathbf{G}\left(\boldsymbol{\theta}, \mathbf{x}_n^{(K)}\right) \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}.$$

The regularity conditions have been adapted for the GST version of results. We also introduce an extra condition requiring the estimating function to have a certain structure for the asymptotic covariance matrix, which is the main condition that will be checked when we apply this theorem. In Section 3.2 and Section 4.1, where Theorem 3.2 below is applied, we shall check conditions 1–4 of these conditions and the remaining conditions 5–7 will be assumed to hold.

Conditions 3.2.

1. *For each $k = 1, \dots, K$, $\hat{\boldsymbol{\theta}}_n^{(k)}$ is a consistent estimator for $\boldsymbol{\theta}_\mathbf{0}$, that is as $n \to \infty$, $\hat{\boldsymbol{\theta}}_n^{(k)}$ converges in probability to $\boldsymbol{\theta}_\mathbf{0}$, written*

$$\hat{\boldsymbol{\theta}}_n^{(k)} \xrightarrow{p} \boldsymbol{\theta}_\mathbf{0}.$$

2. *For each* $k = 1, \ldots, K$, $n^{-\frac{1}{2}}\mathbf{G}_n\left(\boldsymbol{\theta}_\mathbf{0}, \mathbf{X}_n^{(k)}\right)$ *converges in distribution to a zero-mean Gaussian random variable with finite-valued, positive-definite covariance matrix* $\mathbf{B}^{(k)}$*, specifically*

$$n^{-\frac{1}{2}}\mathbf{G}_n\left(\boldsymbol{\theta}_\mathbf{0}, \mathbf{X}_n^{(k)}\right) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{B}^{(k)}\right)$$

3. *For all* $\boldsymbol{\theta}_n^*$ *such that* $\boldsymbol{\theta}_n^* \xrightarrow{p} \boldsymbol{\theta}_\mathbf{0}$, $n^{-1}\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{G}_n(\boldsymbol{\theta}, \mathbf{x}_n^{(k)})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_n^*}$ *converges in probability to a finite-valued, positive-definite matrix* $\mathbf{A}^{(k)}$ *for each* $k = 1, \ldots, K$*. That is*

$$n^{-1}\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{G}_n\left(\boldsymbol{\theta}, \mathbf{x}_n^{(k)}\right)\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_n^*} \xrightarrow{p} \mathbf{A}^{(k)}, \qquad \text{for } k = 1, \ldots, K.$$

4. *For* $1 \leq k_1 \leq k_2 \leq K$*, we require*

$$Cov\left(n^{-\frac{1}{2}}\mathbf{G}_n\left(\boldsymbol{\theta}_\mathbf{0}, \mathbf{X}_n^{(k_1)}\right), n^{-\frac{1}{2}}\mathbf{G}_n\left(\boldsymbol{\theta}_\mathbf{0}, \mathbf{X}_n^{(k_2)}\right)\right) \xrightarrow{p} \mathbf{B}^{(k_1)}.$$

5. *The sequence of random variables* $n^{-\frac{1}{2}}\mathbf{G}_n\left(\boldsymbol{\theta}_\mathbf{0}, \mathbf{X}_n^{(1)}\right), \ldots, n^{-\frac{1}{2}}\mathbf{G}_n\left(\boldsymbol{\theta}_\mathbf{0}, \mathbf{X}_n^{(K)}\right)$ *is asymptotically multivariate normally distributed.*

6. *The parameter space* Θ *is closed and compact and the true parameter value* $\boldsymbol{\theta}_\mathbf{0}$ *is an interior point of* Θ*.*

7. *For all* $\mathbf{x}_n^{(k)}$*, the derivatives* $\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{G}_n(\boldsymbol{\theta}, \mathbf{x}_n^{(k)})$ *and* $\frac{\partial^2}{\partial\boldsymbol{\theta}^2}\mathbf{G}_n(\boldsymbol{\theta}, \mathbf{x}_n^{(k)})$ *exist in a neighbourhood of* $\boldsymbol{\theta}_\mathbf{0}$ *for each* $k = 1, \ldots, K$*.*

8. $n^{-1}\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{G}_n(\boldsymbol{\theta}, \mathbf{x}_n^{(k)})$ *is uniformly continuous in* $\boldsymbol{\theta}$ *in a neighbourhood of* $\boldsymbol{\theta}_\mathbf{0}$*.*

We now prove that for a group sequential trial with K analyses, when estimating equations are used as a method for estimating the treatment effect, the sequence of treatment effect estimates is asymptotically multivariate normal. The proof for this Theorem closely follows the proof of Theorem 3.1, and we often refer back to the details of steps in Theorem 3.1.

Theorem 3.2. *For each* $k = 1, \ldots, K$ *let the estimate* $\hat{\boldsymbol{\theta}}_n^{(k)}$ *be a solution to the estimating equation* $\mathbf{G}_n(\boldsymbol{\theta}, \mathbf{x}_n^{(k)}) = \mathbf{0}$ *and let* $\boldsymbol{\theta}_\mathbf{0}$ *be the true value of the parameter* $\boldsymbol{\theta}$*. Suppose that Conditions 3.2 hold. Then the sequence of estimates* $\hat{\boldsymbol{\theta}}_n^{(1)}, \ldots, \hat{\boldsymbol{\theta}}_n^{(K)}$ *is asymptotically normal, converging in distribution to a Gaussian random variable,* *given by*

$$n^{\frac{1}{2}}\begin{bmatrix}\hat{\boldsymbol{\theta}}_n^{(1)}-\boldsymbol{\theta_0}\\ \hat{\boldsymbol{\theta}}_n^{(2)}-\boldsymbol{\theta_0}\\ \vdots\\ \hat{\boldsymbol{\theta}}_n^{(K)}-\boldsymbol{\theta_0}\end{bmatrix}\xrightarrow{d}N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\\ \vdots\\ \mathbf{0}\end{bmatrix},\Sigma=\begin{bmatrix}\Sigma_{11}&\Sigma_{12}&\cdots&\Sigma_{1K}\\ \Sigma_{12}&\Sigma_{22}&\cdots&\Sigma_{2K}\\ \vdots&\vdots&\ddots&\vdots\\ \Sigma_{1K}&\Sigma_{2K}&\cdots&\Sigma_{KK}\end{bmatrix}\right)$$

where

$$\Sigma_{k_1k_2}=(\mathbf{A}^{(k1)})^{-1}\mathbf{B}^{(k1)}((\mathbf{A}^{(k2)})^{-1})^T.$$

Proof. Following the proof of Theorem 3.1 up to Equation (3.1), we have for each $k=1,\ldots,K$,

$$-n^{-\frac{1}{2}}(\mathbf{A}^{(k)})^{-1}\mathbf{G}_n(\boldsymbol{\theta_0},\mathbf{x}^{(k)})=(\mathbf{A}^{(k)})^{-1}\begin{bmatrix}n^{-1}\frac{\partial}{\partial\boldsymbol{\theta}}G_n^1(\boldsymbol{\theta},\mathbf{x}^{(k)})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{*(k)}_{n,1}}\\ \vdots\\ n^{-1}\frac{\partial}{\partial\boldsymbol{\theta}}G_n^p(\boldsymbol{\theta},\mathbf{x}^{(k)})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{*(k)}_{n,p}}\end{bmatrix}n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}_n^{(k)}-\boldsymbol{\theta_0})$$

where $\boldsymbol{\theta}^{*(k)}_{n,j}$ lies on the line segment between $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_n^{(k)}$.

Let $\bar{\mathbf{A}}$ be the block diagonal matrix whose k^{th} diagonal matrix is the $p\times p$ matrix $\mathbf{A}^{(k)}$. Then, aggregating the above equation for $k=1,\ldots,K$, we have that

$$-n^{-\frac{1}{2}}\bar{\mathbf{A}}^{-1}\begin{bmatrix}\mathbf{G}_n(\boldsymbol{\theta_0},\mathbf{x}^{(1)})\\ \vdots\\ \mathbf{G}_n(\boldsymbol{\theta_0},\mathbf{x}^{(K)})\end{bmatrix}=\bar{\mathbf{A}}^{-1}\begin{bmatrix}n^{-1}\frac{\partial}{\partial\boldsymbol{\theta}}G_n^1(\boldsymbol{\theta},\mathbf{x}^{(1)})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{*(1)}_{n,1}}\\ \vdots\\ n^{-1}\frac{\partial}{\partial\boldsymbol{\theta}}G_n^p(\boldsymbol{\theta},\mathbf{x}^{(1)})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{*(1)}_{n,p}}\\ \vdots\\ n^{-1}\frac{\partial}{\partial\boldsymbol{\theta}}G_n^1(\boldsymbol{\theta},\mathbf{x}^{(K)})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{*(K)}_{n,1}}\\ \vdots\\ n^{-1}\frac{\partial}{\partial\boldsymbol{\theta}}G_n^p(\boldsymbol{\theta},\mathbf{x}^{(K)})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{*(K)}_{n,p}}\end{bmatrix}\cdot n^{\frac{1}{2}}\begin{bmatrix}\hat{\boldsymbol{\theta}}_n^{(1)}-\boldsymbol{\theta_0}\\ \hat{\boldsymbol{\theta}}_n^{(2)}-\boldsymbol{\theta_0}\\ \vdots\\ \hat{\boldsymbol{\theta}}_n^{(K)}-\boldsymbol{\theta_0}\end{bmatrix}. \tag{3.4}$$

The remaining steps in this proof follow the final steps of Theorem 3.1 and the details are omitted. We shall now summarise the final steps. We see that since the estimates $\hat{\boldsymbol{\theta}}^{(1)},\ldots,\hat{\boldsymbol{\theta}}^{(K)}$ are consistent, the difference between $\boldsymbol{\theta}^{*(k)}_{n,j}$ and $\boldsymbol{\theta}_0$ is asymptotically negligible for all $k=1,\ldots,K$ and $j=1,\ldots,p$, and by condition 3, we have

$$\begin{bmatrix} n^{-1}\frac{\partial}{\partial\boldsymbol{\theta}}G_n^1(\boldsymbol{\theta},\mathbf{x}^{(1)})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{*(1)}_{n,1}} \\ \vdots \\ n^{-1}\frac{\partial}{\partial\boldsymbol{\theta}}G_n^p(\boldsymbol{\theta},\mathbf{x}^{(1)})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{*(1)}_{n,p}} \\ \vdots \\ n^{-1}\frac{\partial}{\partial\boldsymbol{\theta}}G_n^1(\boldsymbol{\theta},\mathbf{x}^{(K)})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{*(K)}_{n,1}} \\ \vdots \\ n^{-1}\frac{\partial}{\partial\boldsymbol{\theta}}G_n^p(\boldsymbol{\theta},\mathbf{x}^{(K)})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{*(K)}_{n,p}} \end{bmatrix} \xrightarrow{p} \bar{\mathbf{A}}.$$

Condition 2 states that $n^{-\frac{1}{2}}\mathbf{G}_n(\boldsymbol{\theta}_\mathbf{0},\mathbf{X}_n^{(k)})$ is asymptotically multivariate normal for each $k = 1,\dots,K$. Combining this with condition 4 and condition 5, we have that

$$n^{-\frac{1}{2}}\begin{bmatrix}\mathbf{G}_n(\boldsymbol{\theta}_\mathbf{0},\mathbf{X}_n^{(1)})\\ \mathbf{G}_n(\boldsymbol{\theta}_\mathbf{0},\mathbf{X}_n^{(2)})\\ \vdots \\ \mathbf{G}_n(\boldsymbol{\theta}_\mathbf{0},\mathbf{X}_n^{(K)})\end{bmatrix} \xrightarrow{d} N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\\ \vdots\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{B}^{(1)} & \mathbf{B}^{(1)} & \cdots & \mathbf{B}^{(1)}\\ \mathbf{B}^{(1)} & \mathbf{B}^{(2)} & \cdots & \mathbf{B}^{(2)}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{B}^{(1)} & \mathbf{B}^{(2)} & \cdots & \mathbf{B}^{(K)}\end{bmatrix}\right).$$

Therefore, the left hand side of Equation (3.4) converges in distribution to a Gaussian distribution given by

$$-n^{-\frac{1}{2}}\bar{\mathbf{A}}^{-1}\begin{bmatrix}\mathbf{G}_n(\boldsymbol{\theta}_\mathbf{0},\mathbf{X}_n^{(1)})\\ \mathbf{G}_n(\boldsymbol{\theta}_\mathbf{0},\mathbf{X}_n^{(2)})\\ \vdots \\ \mathbf{G}_n(\boldsymbol{\theta}_\mathbf{0},\mathbf{X}_n^{(K)})\end{bmatrix} \xrightarrow{d} N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\\ \vdots\\ \mathbf{0}\end{bmatrix}, \bar{\mathbf{A}}^{-1}\begin{bmatrix}\mathbf{B}^{(1)} & \mathbf{B}^{(1)} & \cdots & \mathbf{B}^{(1)}\\ \mathbf{B}^{(1)} & \mathbf{B}^{(2)} & \cdots & \mathbf{B}^{(2)}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{B}^{(1)} & \mathbf{B}^{(2)} & \cdots & \mathbf{B}^{(K)}\end{bmatrix}\bar{\mathbf{A}}^{-1}\right).$$

Finally, by matrix manipulation and noting the block-diagonal structure for $\bar{\mathbf{A}}$, we have the result

$$n^{\frac{1}{2}}\begin{bmatrix}\hat{\boldsymbol{\theta}}_n^{(1)} - \boldsymbol{\theta}_\mathbf{0}\\ \hat{\boldsymbol{\theta}}_n^{(2)} - \boldsymbol{\theta}_\mathbf{0}\\ \vdots\\ \hat{\boldsymbol{\theta}}_n^{(K)} - \boldsymbol{\theta}_\mathbf{0}\end{bmatrix} \xrightarrow{d} N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\\ \vdots\\ \mathbf{0}\end{bmatrix}, \Sigma = \begin{bmatrix}\Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1K}\\ \Sigma_{12} & \Sigma_{22} & \cdots & \Sigma_{2K}\\ \vdots & \vdots & \ddots & \vdots\\ \Sigma_{1K} & \Sigma_{2K} & \cdots & \Sigma_{KK}\end{bmatrix}\right)$$

where

$$\Sigma_{k_1k_2} = (\mathbf{A}^{(k1)})^{-1}\mathbf{B}^{(k1)}((\mathbf{A}^{(k2)})^{-1})^T.$$

□

We now note the relationship between Theorem 3.2 and the canonical joint distribution of Definition 2.2. Suppose that $\mathbf{A}^{(k)} = \mathbf{B}^{(k)}$ for each $k = 1, \ldots, K$, then we have that

$$Cov\left(\hat{\boldsymbol{\theta}}_n^{(k_1)}, \hat{\boldsymbol{\theta}}_n^{(k_2)}\right) = \Sigma_{k_1 k_2} = \left(\mathbf{B}^{(k_2)}\right)^{-1} = Var\left(\hat{\boldsymbol{\theta}}_n^{(k_2)}\right)$$

and condition 3 of the canonical joint distribution holds. For many estimates that are the solution to an estimating equation, this property and hence the canonical joint distribution holds. Consider for example, the maximum likelihood estimate (MLE), $\hat{\boldsymbol{\theta}}_n$. At analysis k of a GST with K analyses, $\hat{\boldsymbol{\theta}}_n^{(k)}$ is defined as the value of $\boldsymbol{\theta}$ which maximises the log-likelihood function $\ell(\boldsymbol{\theta}, \mathbf{x}^{(k)})$. Alternatively, $\hat{\boldsymbol{\theta}}_n^{(k)}$ is defined as the value of $\boldsymbol{\theta}$ which is the solution to the equation

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}, \mathbf{x}_n^{(k)}) = \mathbf{0} \tag{3.5}$$

and the Fisher information matrix at analysis k is defined as

$$\mathcal{I}^{(k)}(\boldsymbol{\theta}) = -\mathbb{E}\left[\frac{1}{n}\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \ell(\boldsymbol{\theta}, \mathbf{X}_n^{(k)})\right]. \tag{3.6}$$

Proofs of the following results can be found in Section 2.4 of Wakefield (2013):

- For each $k = 1, \ldots, K$, $\mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}, \mathbf{X}_n^{(k)})\right] = \mathbf{0}$
- For each $k = 1, \ldots, K$, $n^{-\frac{1}{2}} \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}_0, \mathbf{x}_n^{(k)}) \xrightarrow{d} N(\mathbf{0}, \mathcal{I}^{(k)}(\boldsymbol{\theta}_0))$
- For all $\boldsymbol{\theta}_n^*$ such that $\boldsymbol{\theta}_n^* \xrightarrow{p} \boldsymbol{\theta}_0$, we have that $n^{-1} \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}, \mathbf{x}_n^{(k)})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_n^*} \xrightarrow{p} \mathcal{I}^{(k)}(\boldsymbol{\theta}_0)$.

Comparing these results with the regularity Conditions 3.2, it is clear that for the MLE, we have $\mathbf{A}^{(k)} = \mathcal{I}^{(k)}(\boldsymbol{\theta}) = \mathbf{B}^{(k)}$ for each $k = 1, \ldots, K$. Hence, the canonical joint distribution holds for the sequence of MLEs, $\hat{\boldsymbol{\theta}}_n^{(1)}, \ldots, \hat{\boldsymbol{\theta}}_n^{(K)}$.

In Section 3.2, we consider the Cox proportional hazards model. We shall show that the canonical joint distribution holds for the sequence of treatment effect estimates that are found using maximum partial likelihood. To do so, we prove that Conditions 3.2 hold and also use the property that $\mathbf{A}^{(k)} = \mathbf{B}^{(k)}$ for each $k = 1, \ldots, K$.

Theorem 3.2 also provides a basis for the theory that follows in Section 4.1. A joint model for longitudinal and survival data is considered and an analysis method called the conditional score is used to find a sequence of treatment effect estimates. In this example, the canonical joint distribution does not hold because $\mathbf{A}^{(k)} \neq \mathbf{B}^{(k)}$, however we show that Conditions 3.2 hold and hence we are able to derive the asymptotic distribution for the sequence of treatment effect estimates.

3.2 | Survival models

3.2.1 | The cox proportional hazards model

In Section 2.1.1 it was noted that to perform a hypothesis test one must find a statistical model for the endpoint, a method for estimating the parameters of the model and the distribution of the parameter estimates. To model the survival data we shall use the well-studied Cox proportional hazards model which naturally leads to estimating parameters using partial likelihood. Finally, we shall introduce counting processes which will then be used to derive the distribution of the parameter estimates.

The paper "Regression models and life-tables" is very popular within medical statistics. Cox (1972) elegantly formalises a model for survival data and introduces a function called "conditional likelihood". Later, Cox (1975) calls this function the "partial likelihood" and proves some large sample properties of the the estimator that maximises the partial likelihood function. In this section, we present some definitions and results of Cox (1972) that are used in the set-up prior to proving some asymptotic results.

We shall use a Cox proportional hazards model to specify how the covariates are associated with the time-to-event endpoint. By allowing the hazard function to include covariate information, we can specify which patient groups are at a higher risk of the event at any given time. The Cox proportional hazards model assumes that covariates have a multiplicative effect on the rate of death, which is an attractive feature because the parameters have straightforward and useful interpretations. For each patient $i = 1, \dots, n$, let F_i be the time-to-failure random variable for patient i and let C_i be the potential censoring time random variable. The random variable $T_i = \min(F_i, C_i)$ is known as the event time and t_i is the observed event time for patient i. Also, the censoring indicator $\delta_i = \mathbb{I}\{F_i \leq C_i\}$ is observed. The Cox proportional hazard model allows us to specify our beliefs about survival through the hazard function

$$h_i(t) = h_0(t) \exp\{\theta^T Z_i(t)\},$$

where $h_0(\cdot)$ is the baseline hazard function which is left unspecified, $Z_i(t)$ is a $p \times 1$ column vector of time-varying covariates for patient i and θ is a $p \times 1$ vector of regression parameters.

3.2.2 | Cox's partial likelihood

For analysis of survival data and parameter estimation in the Cox proportional hazards model, the partial likelihood function has many benefits. Partial likelihood plays a similar role to the full likelihood function and under certain regularity conditions, estimates derived using partial likelihood are asymptotically as efficient as those derived using full likelihood. However the difference is that the form and parameters of the baseline hazard function are not included in the partial likelihood. The baseline hazard function can be thought of as an infinite-dimensional nuisance parameter and estimation of the parameters brings no statistical benefit. Cox (1975) first presented partial likelihood with the title "conditional likelihood" because the contribution to the partial likelihood of the i^{th} individual is the conditional probability of that individual failing given all individuals that are at risk of failing at the time that individual i fails.

For $i = 1, \ldots, n$, let t_i be the time at which patient i fails or is censored, and let $R(t_i)$ be the set of all individuals with event time greater than or equal to t_i, known as the at-risk set. If patient i is indeed observed to fail, then the contribution to the partial likelihood from the failure at time t_i is

$$L_i(\theta) = \frac{h_i(t_i)}{\sum_{j \in R(t_i)} h_j(t_i)} = \frac{\exp\{\theta^T Z_i(t_i)\}}{\sum_{j \in R(t_i)} \exp\{\theta^T Z_j(t_i)\}}.$$

The partial likelihood function is the product of these conditional probabilities for all failure times, which is given by

$$L(\theta) = \prod_{i=1}^{n} \left(\frac{\exp\{\theta^T Z_i(t_i)\}}{\sum_{j \in R(t_i)} \exp\{\theta^T Z_j(t_i)\}} \right)^{\delta_i}. \tag{3.7}$$

3.2.3 | Parameter estimation in survival analysis

In section 2.1.1, it was shown how the treatment effect estimate and the fixed sample information level can be used to define a standardised test statistic which is needed to perform a hypothesis test. For many statistical models, a convenient approach for finding the parameter estimate and the information level is through the use of a "score statistic". Further, the score statistic is also essential for deriving the asymptotic distribution of the parameter estimates. The score statistic is the first derivative of the log-likelihood with respect to the parameter. We define a score statistic for standard parametric models and then apply this definition to the Cox partial likelihood to define the parameter estimate and information matrix.

Definition 3.3. *Suppose that $L(\theta)$ is a likelihood function for the statistical model with observed data $\mathbf{x} = \{x_1, \ldots, x_n\}$ and $p \times 1$ column vector θ. The "score statistic" is the $p \times 1$ column vector given by*

$$U(\theta) = \frac{\partial}{\partial \theta} \log L(\theta).$$

Setting the score statistic set equal to zero defines an estimating equation. To see this, we write the likelihood function as $L(\theta) = f(\mathbf{x}; \theta)$, and the sample space as $\mathcal{X}$. Then, by the assumption that the orders of integration and differentiation can be exchanged, we assess the expectation of the score statistic.

$$\begin{aligned}
\mathbb{E}(U(\theta)) &= \int_{\mathcal{X}} \frac{\partial}{\partial \theta} (\log f(\mathbf{x}; \theta)) f(\mathbf{x}; \theta) d\mathbf{x} \\
&= \int_{\mathcal{X}} \frac{1}{f(\mathbf{x}; \theta)} \frac{\partial f(\mathbf{x}; \theta)}{\partial \theta} f(\mathbf{x}; \theta) d\mathbf{x} \\
&= \frac{\partial}{\partial \theta} \int_{\mathcal{X}} f(\mathbf{x}; \theta) d\mathbf{x} \\
&= \frac{\partial}{\partial \theta} 1 \\
&= 0.
\end{aligned}$$

The score statistic has expectation zero, which shows that $U(\theta) = 0$ defines an estimating equation and the asymptotic distributional results of Theorem 3.1 hold.

For the Cox proportional hazards model, an estimating function can be created by taking the score of the partial likelihood, Equation (3.7). The log-partial likelihood is a scalar function of the vector θ and the score statistic is a $p \times 1$ column vector. These functions are given by

$$\begin{aligned}
\log L(\theta) &= \sum_{i=1}^{n} \delta_i \left(\theta^T Z_i(t_i) - \log \sum_{j \in R(t_i)} \exp\{\theta^T Z_j(t_i)\} \right) \\
U(\theta) &= \sum_{i=1}^{n} \delta_i \left(Z_i(t_i) - \frac{\sum_{j \in R(t_i)} Z_j(t_i) \exp\{\theta^T Z_j(t_i)\}}{\sum_{j \in R(t_i)} \exp\{\theta^T Z_j(t_i)\}} \right). \qquad (3.8)
\end{aligned}$$

Estimates for parameters of the Cox proportional hazards model are found as the root of the equation that sets the score statistic equal to zero. The p-dimensional estimate $\hat{\theta}$ for θ is the solution to the equation $U(\theta) = 0$. It is not always true that the solution is unique however it can be shown that the probability of a unique root converges to one as $n \to \infty$. Further, in practice we have not found multiple roots and so do not discuss uniqueness further.

The final object of importance is the information matrix, which is -1 times the first derivative of the score statistic with respect to θ. For any column vector X we use the shorthand notation $X^{\otimes 2} = XX^T$. Differentiating (3.8), we see that the information is a $p \times p$ dimensional matrix given by

$$\mathcal{I}(\theta) = \sum_{i=1}^{n} \delta_i \left(\frac{\sum_{j \in R(t_i)} Z_j(t_i)^{\otimes 2} \exp\{\theta^T Z_j(t_i)\}}{\sum_{j \in R(t_i)} \exp\{\theta^T Z_j(t_i)\}} - \left(\frac{\sum_{j \in R(t_i)} Z_j(t_i) \exp\{\theta^T Z_j(t_i)\}}{\sum_{j \in R(t_i)} \exp\{\theta^T Z_j(t_i)\}} \right)^{\otimes 2} \right). \tag{3.9}$$

Later we shall see that asymptotically $\mathcal{I}(\theta) = Var(\hat{\theta})^{-1}$ for score statistics and maximum partial likelihood estimates in the Cox model.

3.2.4 | Counting processes

Commonly throughout the survival data literature, the Cox proportional hazards model and partial likelihood are formulated under the framework of counting processes. The aim of this section is to prove the asymptotic distribution of the treatment effect estimate obtained from the Cox proportional hazards model and to do so we shall introduce counting processes and some useful results about them.

A counting process is analysed under the martingale framework. Andersen et al. (2012) present and prove many results for stochastic processes, martingales and counting processes in their book “Statistical models based on counting processes”. In an earlier paper, Andersen and Gill (1982) show how the Cox proportional hazards model can be presented in the counting process framework. The authors present a proof of the asymptotic normality and consistency of the estimator of interest which we follow for our proof of Theorem 3.5. Our aim is to summarise the necessary results providing the basis for new theory in a later section. An interested reader should refer to Andersen et al. (2012) for further details and discussion of regularity conditions.

As in section 2.2, for $i = 1, \dots, n$, F_i is the time-to-failure random variable for patient i and the hazard function is given by

$$\begin{aligned} h_i(t) &= lim_{\delta t \downarrow 0} \frac{\mathbb{P}(t \leq F_i < t + \delta t | F_i > t)}{\delta t} \\ &= h_0(t) \exp\{\theta^T Z_i(t)\}. \end{aligned}$$

The hazard function defines the probability of an event happening at time t given that patient i has not yet experienced the event or been censored before time t. Suppose that patient i has experienced the event or been censored before time t, then the probability that the event is observed at time t is zero. To formalise this idea, we introduce the "at-risk" process which is given by

$$Y_i(t) = \mathbb{I}\{T_i \geq t\}. \tag{3.10}$$

Then we can introduce the "intensity process" which gives the unconditional probability of the event being observed to occur at time t. The intensity process is given by

$$\lambda_i(t) = h_i(t)Y_i(t) = h_0(t)\exp\{\theta^T Z_i(t)\}Y_i(t). \tag{3.11}$$

The intensity process is a measure of the rate of change of a counting process and we shall shortly introduce the counting process to which Equation (3.11) relates. In general, a counting process is an increasing stochastic process $\{N(t), t \geq 0\}$ taking integer values. Andersen et al. (2012, Sec 11.4.1) present a formal definition. The survival counting process is an object under the classification of counting process.The restriction is that the survival counting process can only take values in $\{0, 1\}$ where 0 means that the event has not yet happened and 1 means the event has happened. Since our analysis only concerns survival data, we restrict our attention to the survival counting process.

Definition 3.4. *Let F_i and C_i be the time-to-failure and time-to-censoring random variables respectively for patient i where censoring is non-informative and let $T_i = \min(F_i, C_i)$ be the event time random variable. Let t_i be the observed event time and $\delta_i = \mathbb{I}\{T_i \leq C_i\}$ the censoring indicator for patient i. Then the survival counting processes is the stochastic process*

$$N_i(t) = \mathbb{I}\{t_i \leq t, \delta_i = 1\}.$$

A counting process, in the general definition, is a step-function increasing in integer increments. For the survival counting process, $N_i(t)$ is a step function jumping from 0 to 1 at the failure time t_i for an uncensored observation. The intensity process measures the rate of change in the increments and can be interpreted as the instantaneous probability of the jump. The intensity process is defined by

$$\lambda_i(t)dt = \mathbb{P}(N_i(t) \text{ jumps in the interval } [t, t + dt)|\mathcal{F}_{t-})$$

where $\mathcal{F}_{t-}$ denotes everything that has happened until just before time t and so determines $Y_i(t)$. The object dt has a special meaning in stochastic calculus and is required for the integration of a stochastic function, see for example Andersen et al.

(2012) Section III.1. For the purpose of this thesis, the object dt can be interpreted as an infinitesimally small value of time. This probability is best understood through a function

$$\begin{aligned} dN_i(t) &= N_i(t+dt) - N_i(t^-) \\ &= \mathbb{I}\{t \leq t_i < t+dt, \delta_i = 1\}. \end{aligned}$$

It can now be seen that the intensity process is such that

$$\lambda_i(t)dt = \mathbb{P}(dN_i(t) = 1 | \mathcal{F}_{t^-}). \tag{3.12}$$

The above function $dN_i(t)$ also presents us with useful notation: for any function or stochastic process $f(\cdot)$, the stochastic integral

$$\int_0^\infty f(u)dN_i(u) = f(t_i) \tag{3.13}$$

is f evaluated at the place where N_i jumps from 0 to 1 if $\delta_i = 1$, and 0 otherwise. Further, the function $dN_i(t)$ is the increment of $N_i(t)$ over a small interval dt and the following relationship holds

$$N_i(t) = \int_0^t dN_i(u).$$

A martingale is a sequence of random variables for which the conditional expectation of a future value is equal to the current value of the sequence, and a submartingale is a sequence of random variables for which the conditional expectation at a future time point is greater than or equal to the current value of the sequence. It can be seen that the counting process $N_i(t)$ is a submartingale and we can create an object $M_i(t)$ called the "compensated counting process" which will be a martingale. The compensated counting process is given by

$$M_i(t) = N_i(t) - \int_0^t \lambda_i(u)du. \tag{3.14}$$

In a similar manner to the survival counting process, it is useful to define the following function

$$\begin{aligned} dM_i(t) &= M_i(t+dt) - M_i(t^-) \\ &= dN_i(t) - \lambda_i(t)dt. \end{aligned} \tag{3.15}$$

This satisfies the relationship

$$M_i(t) = \int_0^t dM_i(u).$$

In Lemma 3.3, we shall show that $M_i(t)$ satisfies the defining feature of a martingale in that the conditional expectation of a future observation of $M_i(\cdot)$ is equal to the current value of $M_i(\cdot)$. To do so, we need to introduce the notion of a predictable process. This is simply a stochastic process that is fixed at time t given what has happened before time t. Note that $Y_i(t)$ and $\lambda_i(t)dt$ are predictable. The score statistic for partial likelihood can be seen to be a predictable process integrated with respect to a martingale, so we shall show that integrating a predicable process gives rise to a martingale.

The final object of importance for the proof of the asymptotic distribution of the score for partial likelihood, is the predictable covariation process for martingales. This is the covariance of two martingales evaluated at time t and conditional on $\mathcal{F}_{t^-}$. We shall use the notation

$$\langle M_i, M_j \rangle(t) = Cov(M_i(t), M_j(t)|\mathcal{F}_{t^-}).$$

We derive the predictable covariation process for $M_i(t)$ and the predictable covariation for predictable processes integrated with respect to martingales.

Lemma 3.3. *Let $M_i(t)$ be the compensated counting process given by Equation* (3.14) *and let $H_i(t)$ be a predictable process. Then the stochastic process given by*

$$M_i'(t) = \int_0^t H_i(u) dM_i(u)$$

is a martingale. Further, for independent martingales $M_i(t)$ and $M_j(t)$, the covariance process is such that

$$\langle M_i', M_i' \rangle(t) = \int_0^t H_i^2(u)\lambda_i(u)du$$
$$\langle M_i', M_j' \rangle(t) = 0 \text{ for } i \neq j.$$

Proof. To prove that the stochastic process $M_i'(t)$ is a martingale, we show the defining property which is that $\mathbb{E}(M_i'(v) - M_i'(u)|\mathcal{F}_u) = 0$. The second line of the proof below uses the facts that the processes $H_i(t)$ and $\lambda_i(t)dt$ are predictable and so are fixed conditional on $\mathcal{F}_{t^-}$, and that $dN_i(t)$ is an indicator function. Then, the final line uses the definition of the intensity process given by Equation (3.12).

$$\begin{aligned}\mathbb{E}(dM_i'(t)|\mathcal{F}_{t^-}) &= \mathbb{E}(H_i(t)dM_i(t)|\mathcal{F}_{t^-})\\ &= H_i(t)\left[\mathbb{P}(dN_i(t)=1|\mathcal{F}_{t^-}) - \lambda_i(t)dt\right]\\ &= H_i(t)\left[\lambda_i(t)dt - \lambda_i(t)dt\right] = 0\end{aligned}$$

This is equivalent to the result that $\mathbb{E}(M_i'(v) - M_i'(u)|\mathcal{F}_u) = 0$ for $v > u$.

For the predictable covariance process, we compute $\langle dM_i', dM_j'\rangle(t) = \mathbb{E}(dM_i'(t)dM_j'(t)|\mathcal{F}_{t^-})$. This is because $\mathbb{E}(dM_i'(t)|\mathcal{F}_{t^-}) = \mathbb{E}(dM_j'(t)|\mathcal{F}_{t^-}) = 0$. In the second and third lines of the proof below, we use the fact that the processes $H_i(t), H_j(t), \lambda_i(t)dt$ and $\lambda_j(t)dt$ are predictable and Equation (3.12) which is the definition of the intensity process. Then, since dt is a small interval, in the final line we restrict attention to terms up to order dt.

$$\begin{aligned}&\mathbb{E}(H_i(t)dM_i(t)H_j(t)dM_j(t)|\mathcal{F}_{t-})\\ =&H_i(t)H_j(t)\mathbb{E}(dN_i(t)dN_j(t) - \lambda_i(t)dN_j(t)dt - \lambda_j(t)dN_i(t)dt + \lambda_i(t)\lambda_j(t)(dt)^2|\mathcal{F}_{t^-})\\ =&H_i(t)H_j(t)\left[\mathbb{E}(dN_i(t)dN_j(t)|\mathcal{F}_{t^-}) - \lambda_i(t)\lambda_j(t)(dt)^2\right]\\ \approx&H_i(t)H_j(t)\mathbb{E}(dN_i(t)dN_j(t)|\mathcal{F}_{t^-})\end{aligned}$$

The function $dN_i(t)$ is an indicator function and event times have probability zero of being tied, this implies that $dN_i(t)^2 = dN_i(t)$ and also that $dN_i(t)dN_j(t) = 0$ for $i \neq j$. Therefore, we have that

$$\begin{aligned}\mathbb{E}(H_i(t)^2 dM_i(t)^2|\mathcal{F}_{t^-}) &= H_i(t)^2\lambda_i(t)dt\\ \mathbb{E}(H_i(t)dM_i(t)H_j(t)dM_j(t)|\mathcal{F}_{t^-}) &= 0 \text{ for } i \neq j.\end{aligned}$$

This is equivalent to the result that

$$\begin{aligned}\langle M_i', M_i'\rangle(t) &= \int_0^t H_i^2(u)\lambda_i(u)du\\ \langle M_i', M_j'\rangle(t) &= 0 \text{ for } i \neq j.\end{aligned}$$

□

The above Lemma 3.3 with $H_i(t) = 1$ proves the result that the compensated counting process $M_i(t)$ is a martingale. This result will be called upon in the proof of the asymptotic distribution of the score statistic for partial likelihood.

3.2.5 | Partial likelihood under the counting process framework

Using the definition of the survival counting process and the at-risk process, we can reformulate the score statistic for partial likelihood and the associated information. This is useful because we can appeal to the Martingale central limit theorem to prove the distribution of the treatment effect estimate.

The following functions are notationally convenient in definitions of the score statistic and information matrix. The main idea is that the summation over the set of individuals at risk is replaced by a summation over the full set of individuals with the at-risk indicator function giving zero weight to those no longer at risk. In the asymptotic distribution theory we shall study these functions as $n \to \infty$. The object $S^{(0)}(\theta, t)$ is a scalar and $S^{(1)}(\theta, t)$ and $E(\theta, t)$ are column vectors of length p, where p is the length of the covariate vector $Z_i(t)$. The objects $S^{(2)}(\theta, t)$ and $V(\theta, t)$ are $p \times p$ dimensional matrices. The definitions of these terms are

$$\begin{aligned}
S^{(0)}(\theta, t) &= \frac{1}{n} \sum_{i=1}^{n} Y_i(t) \exp\{\theta^T Z_i(t)\} \\
S^{(1)}(\theta, t) &= \frac{1}{n} \sum_{i=1}^{n} Z_i(t) Y_i(t) \exp\{\theta^T Z_i(t)\} \\
S^{(2)}(\theta, t) &= \frac{1}{n} \sum_{i=1}^{n} Z_i(t) Z_i(t)^T Y_i(t) \exp\{\theta^T Z_i(t)\} \\
E(\theta, t) &= \frac{S^{(1)}(\theta, t)}{S^{(0)}(\theta, t)} \\
V(\theta, t) &= \frac{S^{(2)}(\theta, t)}{S^{(0)}(\theta, t)} - \frac{S^{(1)}(\theta, t) S^{(1)}(\theta, t)^T}{[S^{(0)}(\theta, t)]^2}.
\end{aligned}$$

The function $E(\theta, t)$ can be interpreted as the expectation of the covariate vector $Z_i(t)$, if we select an individual with probability proportional to $\exp\{\theta^T Z_i(t)\}$ from the set of individuals at risk at time t and $V(\theta, t)$ is the variance of the covariate vector $Z_i(t)$ in this case.

In section 3.2.4 it was noted that the function $dN_i(t)$ is used to replace a summation with a stochastic integral. This feature is essential for formulating the score statistic for partial likelihood in terms of counting processes. Applying

Equation (3.13) to the summand in Equation (3.8), we have

$$\begin{aligned}\delta_i \left(Z_i(t_i) - \frac{\sum_{j \in R(t_i)} Z_j(t_i) \exp\{\theta^T Z_j(t_i)\}}{\sum_{j \in R(t_i)} \exp\{\theta^T Z_j(t_i)\}} \right) &= \int_0^\infty \left(Z_i(u) - \frac{S^{(1)}(\theta, u)}{S^{(0)}(\theta, u)} \right) dN_i(u) \\ &= \int_0^\infty (Z_i(u) - E(\theta, u) dN_i(u)).\end{aligned}$$

Thus, Equation (3.8) is

$$U(\theta) = \int_0^\infty \sum_{i=1}^n \left(Z_i(u) - E(\theta, u)\right) dN_i(u). \tag{3.16}$$

Similarly Equation (3.9) can be written as

$$\mathcal{I}(\theta) = \int_0^\infty \sum_{i=1}^n V(\theta, u) dN_i(u). \tag{3.17}$$

It is not yet obvious that the asymptotic theory of section 3.1 can be applied here. To do this we seek a function with expectation zero from which to construct an estimating equation. In the following lemma, we show how the counting process in the score statistic can be replaced with its compensated version, producing another martingale.

Lemma 3.4.

$$U(\theta) = \int_0^\infty \sum_{i=1}^n \left(Z_i(u) - E(\theta, u)\right) dN_i(u) = \int_0^\infty \sum_{i=1}^n \left(Z_i(u) - E(\theta, u)\right) dM_i(u).$$

Proof. By Equation (3.15), we have that $dM_i(t) = dN_i(t) - \lambda_i(t)dt$. Thus, the stated result is equivalent to proving that

$$\int_0^\infty \sum_{i=1}^n \left(Z_i(u) - E(\theta, u)\right) \lambda_i(u) du = 0.$$

For $i = 1, \ldots, n$, let $E_i(u) = \exp\{\theta^T Z_i(u)\} Y_i(u)$. Then it is easily seen that

$$\begin{aligned}
&\int_0^\infty \sum_{i=1}^n \left(Z_i(u) - E(\theta, u)\right) \lambda_i(u) du \\
=&\int_0^\infty \sum_{i=1}^n \left(Z_i(u) - E(\theta, u)\right) h_i(u) Y_i(u) du \\
=&\int_0^\infty \sum_{i=1}^n \left(Z_i(u) - \frac{\sum_{j=1}^n Z_j(u) E_j(u)}{\sum_{j=1}^n E_j(u)}\right) h_0(u) E_i(u) du \\
=&\int_0^\infty \sum_{i=1}^n \left(\frac{\sum_{j=1}^n Z_i(u) E_i(u) E_j(u) - \sum_{j=1}^n Z_j(u) E_i(u) E_j(u)}{\sum_{j=1}^n E_j(u)}\right) h_0(u) du \\
=&\int_0^\infty \left(\frac{\sum_{i=1}^n \sum_{j=1}^n Z_i(u) E_i(u) E_j(u) - \sum_{i=1}^n \sum_{j=1}^n Z_j(u) E_i(u) E_j(u)}{\sum_{j=1}^n E_j(u)}\right) h_0(u) du \\
=&\int_0^\infty \left(\frac{\sum_{i=1}^n \sum_{j=1}^n Z_j(u) E_i(u) E_j(u) - \sum_{i=1}^n \sum_{j=1}^n Z_j(u) E_i(u) E_j(u)}{\sum_{j=1}^n E_j(u)}\right) h_0(u) du \\
=&0.
\end{aligned}$$

□

By Lemma 3.4, we have

$$U(\theta) = \int_0^\infty \sum_{i=1}^n \left(Z_i(u) - E(\theta, u)\right) dM_i(u).$$

Then, by application of Lemma 3.3, the process

$$U_t(\theta) = \int_0^t \sum_{i=1}^n \left(Z_i(u) - E(\theta, u)\right) dM_i(u)$$

is a martingale. Therefore, this process has expectation 0 for all values of t, and letting $t \to \infty$, we see that $\mathbb{E}(U(\theta)) = 0$. Hence, we shall use this property to define an estimating equation.

3.3 | Asymptotic theory for survival analysis

3.3.1 | Fixed sample results

The parameter estimate $\hat{\theta}$ is a p-dimensional vector which is the solution to the set of p equations $U(\theta) = 0$ and $\hat{\theta}$ has information matrix $\mathcal{I}(\theta)$. We shall denote these objects by $\hat{\theta}_n$, $U_n(\theta)$ and $\mathcal{I}_n(\theta)$ to show dependency on the number of patients n. In this section we derive an asymptotic distribution for $\hat{\theta}_n$ which requires assessing the behaviour of $\hat{\theta}_n$ as $n \to \infty$. The asymptotic setting is that as n increases, the rate of recruitment increases in the study and we have more survival observations with the same hazard rate. In Section 3.1 we showed that an estimator that is the solution to an estimating equation converges to a Gaussian distribution. We have shown that $\mathbb{E}(U_n(\theta)) = 0$, hence the conclusion of Theorem 3.1 applies to $\hat{\theta}_n$ if we can establish that Conditions 3.1 hold in this case. From this list, we shall prove that conditions 2 and 3 hold. We direct the reader to Andersen et al. (2012) Lemma 3.1 for the proof that condition 1 holds. The remaining conditions, 4, 5 and 6, are assumed to hold.

Some further regularity conditions, which relate directly to survival data, are needed for the proof of consistency and asymptotic normality of the estimate $\hat{\theta}_n$. Andersen and Gill (1982) present a list of conditions which we shall assume hold and the purpose of these conditions is to avoid technical distractions. Similarly, we shall assume that the following conditions are satisfied.

Conditions 3.5.

1. $\int_0^\tau h_0(u)du < \infty$ *where* τ *is a censoring time applied to all observations.*

2. *There exists a neighbourhood* Θ *of* θ_0 *and functions* $s^{(0)}(\theta, t)$, $s^{(1)}(\theta, t)$ *and* $s^{(2)}(\theta, t)$ *defined on* $\Theta \times [0, \infty)$ *such that*

$$\sup_{t \in [0,\infty), \theta \in \Theta} \left\| S^{(j)}(\theta, t) - s^{(j)}(\theta, t) \right\| \xrightarrow{p} 0 \text{ for } j = 0, 1, 2.$$

Each $s^{(j)}(\theta, t)$ *is a continuous function of* $\theta \in \Theta$ *uniformly in* $t \in [0, \infty)$, *and bounded on* $\Theta \times [0, \infty)$. *Also, the function* $s^{(0)}$ *is bounded away from zero on* $\Theta \times [0, \infty)$.

3. There exists $\delta > 0$ such that

$$n^{-1/2} \sup\{|Z_i(t)|\mathbb{I}(\theta_0^T Z_i(t) > -\delta|Z_i(t)|); i = 1, \ldots, n\} \xrightarrow{p} 0 \text{ as } n \to \infty.$$

Asymptotic probabilistic limits of the score statistic and information matrix are specified through limits of their components. Conditions 3.5 define probabilistic limits of the functions $S^{(j)}(\theta, t)$ for $j = 0, 1, 2$ and specify that these limits exist. It is clear, given the definitions of $S^{(j)}(\theta, t)$ for $j = 0, 1, 2$ that the following relationships hold:

$$\begin{aligned} s^{(1)}(\theta, t) &= \frac{\partial}{\partial \theta} s^{(0)}(\theta, t), \\ s^{(2)}(\theta, t) &= \frac{\partial^2}{\partial \theta^2} s^{(0)}(\theta, t). \end{aligned}$$

Further, the matrices $E(\theta, t)$ and $V(\theta, t)$ converge in probability to $e(\theta, t)$ and $v(\theta, t)$ respectively which are given by

$$\begin{aligned} e(\theta, t) &= \frac{s^{(1)}(\theta, t)}{s^{(0)}(\theta, t)} \\ v(\theta, t) &= \frac{s^{(2)}(\theta, t)}{s^{(0)}(\theta, t)} - e(\theta, t) e(\theta, t)^T. \end{aligned}$$

It is now possible to determine the asymptotic distribution of the parameter estimate $\hat{\theta}_n$. We shall call upon results in Section 3.1 to show that the estimate is asymptotically normally distributed. The following is a heuristic sketch of the proofs given by Andersen et al. (2012) who prove consistency in their Lemma 3.1 and asymptotic normality in their Theorem 3.2.

Theorem 3.5. *Let the estimate $\hat{\theta}_n$ be the solution to the equation $U_n(\theta) = 0$ with $U_n(\theta)$ defined in Equation* (3.16)*. Suppose that θ_0 is the true value of the parameter θ and that Conditions 3.5 hold. Then $\hat{\theta}_n$ converges in distribution to a Gaussian random variable, specifically*

$$n^{\frac{1}{2}}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \Sigma)$$

where the covariance matrix Σ is given by

$$\Sigma = \int_0^\infty v(\theta_0, u) s^{(0)}(\theta_0, u) h_0(u) du.$$

Proof. In this Theorem, the score statistic $U_n(\theta)$ plays the role of the estimating function $\mathbf{G}_n(\boldsymbol{\theta}, \mathbf{x}_n)$. In applying Theorem 3.1, we show the following conditions are satisfied:

A1 $\hat{\theta}_n$ is consistent for θ, written $\hat{\theta}_n \xrightarrow{p} \theta_0$

A2 $n^{-\frac{1}{2}} U_n(\theta_0) \xrightarrow{d} N(0, \Sigma)$ where Σ is a finite valued semi-definite matrix

A3 $n^{-1} \frac{\partial}{\partial \theta} U_n(\theta)|_{\theta=\theta^*} = n^{-1} \mathcal{I}_n(\theta^*) \xrightarrow{p} \Sigma$ for all θ^* consistent for θ.

Note that these conditions are the first three requirements in Conditions 3.1 with $\mathbf{A} = \mathbf{B} = \Sigma$. We shall focus on the proof that conditions A2 and A3 are satisfied since we shall later build upon this proof in the group sequential case. Andersen et al. (2012) prove consistency in their Lemma 3.1.

Condition A2 can be shown by applying Rebolledo's Central Limit Theorem for square integrable martingales. For each $j = 1, \ldots, p$, define

$$W_j^{(n)}(\theta, t) = \int_0^t \sum_{i=1}^n H_{ij}^{(n)}(u) dM_i(u)$$

where

$$H_{ij}^{(n)}(u) = n^{-\frac{1}{2}} (Z_{ij}(u) - E_j(\theta, u)).$$

Then H is an $n \times p$ matrix of locally bounded predictable processes. Thus, $n^{-\frac{1}{2}} U_n(\theta) = W^{(n)}(\theta, \infty)$. For Rebolledo's Central Limit Theorem to hold, we must demonstrate two conditions:

$$\langle W_{j_1}^{(n)}(\theta_0), W_{j_2}^{(n)}(\theta_0) \rangle(t) \xrightarrow{p} \Sigma_{j_1 j_2}$$

and

$$\int_0^\infty \sum_{i=1}^n H_{ij}^{(n)}(u)^2 \lambda_i(u) \mathbb{I}\{|H_{ij}^{(n)}(u)| > \epsilon\} du \xrightarrow{p} 0 \text{ for } j = 1, \ldots, p \text{ and all } \epsilon > 0.$$

For the first part, we note that patients are independent, so that the martingales $M_{i_1}(t)$ and $M_{i_2}(t)$ are independent for $i_1 \neq i_2$. By Lemma 3.3, we have

$$Cov\left(\int_0^t H_{ij_1}^{(n)}(u) dM_i(u), \int_0^t H_{ij_2}^{(n)}(u) dM_i(u) \middle| \mathcal{F}_{t-} \right) = \int_0^t H_{ij_1}^{(n)}(u) H_{ij_2}^{(n)}(u) \lambda_i(u) du$$

$$Cov\left(\int_0^t H_{i_1 j_1}^{(n)}(u) dM_{i_1}(u), \int_0^t H_{i_1 j_2}^{(n)}(u) dM_{i_2}(u) \middle| \mathcal{F}_{t-} \right) = 0 \text{ for } i_1 \neq i_2.$$

Therefore, summing over all patients, the predictable covariation process of the square-integrable martingales $W_j^{(n)}(\theta_0, t)$ is

$$\begin{aligned}\langle W_{j_1}^{(n)}(\theta_0), W_{j_2}^{(n)}(\theta_0)\rangle(t) &= \int_0^t \sum_{i=1}^n H_{ij_1}^{(n)}(u) H_{ij_2}^{(n)}(u)\lambda_i(u)du \\ &= \left(\int_0^t V(\theta_0, u) S^{(0)}(\theta_0, u) h_0(u) du\right)_{j_1 j_2} \\ &\xrightarrow{p} \left(\int_0^t v(\theta_0, u) s^{(0)}(\theta_0, u) h_0(u) du\right)_{j_1 j_2} \\ &= \Sigma_{j_1 j_2}.\end{aligned}$$

This is a direct result of the regularity conditions and definitions of $s^{(j)}$ for $j = 0, 1, 2$.

The second condition for Rebolledo's Central Limit Theorem makes use of the following simple inequality

$$|a-b|^2 \mathbb{I}\{|a-b| > \epsilon\} \leq 4|a|^2 \mathbb{I}\left\{|a| > \frac{\epsilon}{2}\right\} + 4|b|^2 \mathbb{I}\left\{|b| > \frac{\epsilon}{2}\right\}.$$

Substituting $a = n^{-\frac{1}{2}} Z_i(t)$ and $b = n^{-\frac{1}{2}} E(\theta_0, u)$, it is then enough to show that for each j and $\epsilon > 0$

$$\int_0^\infty \sum_{i=1}^n 4|n^{-\frac{1}{2}} Z_{ij}(u)|^2 \mathbb{I}\left\{|n^{-\frac{1}{2}} Z_{ij}(u)| > \frac{\epsilon}{2}\right\} \lambda_i(u) du \xrightarrow{p} 0$$
$$\int_0^\infty \sum_{i=1}^n 4|n^{-\frac{1}{2}} E_j(\theta_0, u)|^2 \mathbb{I}\left\{|n^{-\frac{1}{2}} E_j(\theta_0, u)| > \frac{\epsilon}{2}\right\} \lambda_i(u) du \xrightarrow{p} 0.$$

The proof of these statements is analytically simple and will not be proved here, for further details, we refer the reader to Andersen and Gill (1982) Theorem 3.2.

For condition A3, the first derivative of the score statistic or the information matrix, equation (3.17), we add and subtract common terms then use the triangle inequality to obtain the following inequality:

$$\begin{aligned}\left\|n^{-1} I(\theta^*) - \Sigma\right\| &\leq \left\| n^{-1} \int_0^\infty \sum_{i=1}^n V(\theta^*, u) dN_i(u) - \int_0^\infty v(\theta_0, u) s^{(0)}(\theta_0, u) \lambda_0(u) du \right\| \\ &= \left\| n^{-1} \int_0^\infty \sum_{i=1}^n \{V(\theta^*, u) - v(\theta^*, u)\} dN_i(u) \right\| \\ &+ \left\| n^{-1} \int_0^\infty \sum_{i=1}^n \{v(\theta^*, u) - v(\theta_0, u)\} dN_i(u) \right\| \\ &+ \left\| \int_0^\infty v(\theta_0, u)\{S^{(0)}(\theta_0, u) - s^{(0)}(\theta_0, u)\} \lambda_0(u) du \right\|.\end{aligned}$$

Each term on the right can be shown to converge to zero in probability which proves that $\mathcal{I}_n(\theta) \xrightarrow{p} \Sigma$. The regularity conditions ensure that the matrix Σ is automatically positive semi-definite. □

3.3.2 | Group sequential results

The analysis of survival data in a fixed sample trial transfers nicely to the group sequential trial. The aim for this section is to prove that the sequence of treatment effect estimates obtained in a group sequential trial with survival as the primary endpoint satisfy the canonical joint distribution of Definition 2.1. To do so, we define a sequence of score statistics indexed by k the interim analysis number. We show how the score statistic at analysis k uses all the available information at the time of the analysis. The canonical joint distribution of the sequence of estimates is then easily found after deriving the joint distribution of the sequence of score statistics.

Jennison and Turnbull (1997) prove that the canonical joint distribution of Definition 2.1 holds for for a variety of data types including survival data. We shall explain how the proof by Jennison and Turnbull (1997) is derived so that we can apply similar methods for the joint model in Section 4.1. The proof for the joint model will be a new result and builds upon theory in this section.

There is an additional form of censoring in a group sequential trial. Patients who have entered the trial but for whom the event of interest has not been observed at an interim analysis are marked as censored. For patient $i = 1, \dots, n$ with time-to-failure random variable F_i, let $C_i(k)$ be the time-to-censoring random variable at analysis k, which is similar to end of study censoring in a fixed sample trial. This value $C_i(k)$ is the minimum of the usual censoring random variable C_i and the follow-up time at the interim analysis for patient i. Then at analysis k the time-to-event random variable is $T_i(k) = \min\{F_i, C_i(k)\}$ which has observed event time $t_i(k)$ and the observed censoring indicator is $\delta_i(k) = \mathbb{I}\{F_i \leq C_i(k)\}$.

Section 3.2.5 defined all objects needed to write the score statistic for partial likelihood in terms of counting processes. It is necessary to define these objects for a given interim analysis in order to create a group sequential version of the score statistic. The at-risk process will now be an indicator for not yet observing the event, non-informative censoring or being censored for an interim analysis. For patient i at analysis k the at-risk process is $Y_i(k, t) = \mathbb{I}\{t_i(k) \geq t\}$. The corresponding counting process and compensated counting process for analysis k are

$$\begin{aligned}
N_i(k,t) &= \mathbb{I}\{t_i(k) \leq t, \delta_i(k) = 1\} \\
M_i(k,t) &= N_i(k,t) - \int_0^t \lambda_0(u) \exp\{\theta^T Z_i(u)\} Y_i(k,u) du.
\end{aligned}$$

The functions $S^{(j)}(k,\theta,t)$ for $j = 0,1,2$ along with $E(k,\theta,t)$ and $V(k,\theta,t)$ have obvious definitions which follow from the fixed sample theory of subsection 3.2.5. For completeness, these are

$$\begin{aligned}
S^{(0)}(k,\theta,t) &= \frac{1}{n}\sum_{i=1}^n Y_i(k,t)\exp\{\theta^T Z_i(t)\} \\
S^{(1)}(k,\theta,t) &= \frac{1}{n}\sum_{i=1}^n Z_i(t) Y_i(k,t)\exp\{\theta^T Z_i(t)\} \\
S^{(2)}(k,\theta,t) &= \frac{1}{n}\sum_{i=1}^n Z_i(t) Z_i(t)^T Y_i(k,t)\exp\{\theta^T Z_i(t)\} \\
E(k,\theta,t) &= \frac{S^{(1)}(k,\theta,t)}{S^{(0)}(k,\theta,t)} \\
V(k,\theta,t) &= \frac{S^{(2)}(k,\theta,t)}{S^{(0)}(k,\theta,t)} - \frac{S^{(1)}(k,\theta,t) S^{(1)}(k,\theta,t)^T}{[S^{(0)}(k,\theta,t)]^2}.
\end{aligned}$$

The group sequential score statistics and information matrices are to be calculated during the clinical trial. Let τ_k be the time from the start of the study of interim analysis k, so that all information recorded up to time τ_k is included in score statistic k. Also note that the same reasoning in the proof of Lemma 3.4 can be used to replace $dN_i(k,u)$ with $dM_i(k,u)$ in the score statistic. Therefore, at analysis k the score statistic and information matrices are respectively

$$\begin{aligned}
U(k,\theta) &= \int_0^{\tau_k} \sum_{i=1}^n (Z_i(u) - E(k,\theta,u)) dN_i(k,u) \\
&= \int_0^{\tau_k} \sum_{i=1}^n (Z_i(u) - E(k,\theta,u)) dM_i(k,u) \\
\mathcal{I}(k,\theta) &= \int_0^{\tau_k} \sum_{i=1}^n V(k,\theta,u) dN_i(k,u).
\end{aligned}$$

Some thought should be given as to why the at-risk process $Y_i(k,t)$ depends upon k. This is because, with staggered entry, the maximum follow-up time for each patient at analysis k is the difference between their time of entry to the trial and the time of interim analysis k.

In the proof of the asymptotic distribution of the sequence of treatment effect estimates, we follow the work of Tsiatis et al. (1995) and create new counting processes that allow the score statistic to be divided into distinct increments. For $k = 1, \ldots, K$ the counting processes are

$$\begin{aligned} DN_i(0,t) &= 0 \\ DN_i(k,t) &= N_i(k,t) - N_i(k-1,t). \end{aligned}$$

For $k = 1, \ldots, K$, the equivalent compensated versions are

$$\begin{aligned} &DM_i(0,t) = 0 \\ &DM_i(k,t) = DN_i(k,t) - \int_0^t \lambda_0(u) \exp\{\theta^T Z_i(u)\}(Y_i(k,u) - Y_i(k-1,u))du. \end{aligned}$$

Since the event for patient i can only happen once, we have $N_i(k,t) = \sum_{l=1}^k DN_i(l,t)$. The score statistic at analysis k is the sum of increments up to and including analysis k. Therefore, the score statistic is written:

$$U_n(k,\theta) = \int_0^{\tau_k} \sum_{i=1}^n \sum_{l=1}^k (Z_i(u) - E(k,\theta,u))dDN_i(l,u). \tag{3.18}$$

Regularity conditions for the group sequential trial have few differences to the regularity conditions of the fixed sample case. Similarly to the fixed sample case, we present the conditions given by Andersen and Gill (1982) and we shall assume that these conditions hold to avoid technical distractions.

Conditions 3.6.

1. *For $k = 1, \ldots, K$, $\int_0^{\tau_k} h_0(u)du < \infty$ where τ_k is the calendar time of analysis k.*

2. *There exists a neighbourhood Θ of θ_0 and for each $k = 1, \ldots, K$ there are functions $s^{(0)}(k,\theta,t)$, $s^{(1)}(k,\theta,t)$ and $s^{(2)}(k,\theta,t)$ defined on $\Theta \times [0,\infty)$ such that*

   $$\sup_{t\in[0,\infty),\theta\in\Theta} \left\| S^{(j)}(k,\theta,t) - s^{(j)}(k,\theta,t) \right\| \xrightarrow{p} 0 \text{ for } j = 0,1,2.$$

   Each $s^{(j)}(k,\theta,t)$ is a continuous function of $\theta \in \Theta$ uniformly in $t \in [0,\infty)$, and bounded on $\Theta \times [0,\infty)$. For each $k = 1, \ldots, K$, $s^{(0)}$ is bounded away from zero on $\Theta \times [0,\infty)$.

3. *There exists $\delta > 0$ such that*

$$n^{-1/2}\sup\{|Z_i(t)|\mathbb{I}(\theta_0^T Z_i(t) > -\delta|Z_i(t)|); i = 1, \ldots, n\} \xrightarrow{p} 0 \text{ as } n \to \infty.$$

In a similar manner to the definition of the covariance matrix in the fixed sample trial, the covariance matrix at analysis k denoted $\Sigma^{(k)}$ is derived by defining $e(k, \theta, t)$ and $v(k, \theta, t)$ as products of $s^{(0)}(k, \theta, t), s^{(1)}(k, \theta, t)$ and $s^{(2)}(k, \theta, t)$. The definitions of these functions are as follows:

$$\begin{aligned} e(k, \theta, t) &= \frac{s^{(1)}(k, \theta, t)}{s^{(0)}(k, \theta, t)} \\ v(k, \theta, t) &= \frac{s^{(2)}(k, \theta, t)}{s^{(0)}(k, \theta, t)} - e(k, \theta, t)e(k, \theta, t)^T. \end{aligned}$$

Theorem 3.6. *Let the vector of estimates $(\hat{\theta}_n^{(1)}, \ldots, \hat{\theta}_n^{(K)})^T$ be the solution to the equation $(U_n(1, \theta), \ldots, U_n(K, \theta))^T = 0$. Suppose that θ_0 is the true value of the parameter θ and that Conditions 3.6 hold. Then $(\hat{\theta}_n^{(1)}, \ldots, \hat{\theta}_n^{(K)})^T$ converges in distribution to a multivariate Gaussian random variable, specifically*

$$\begin{pmatrix} n^{\frac{1}{2}}(\hat{\theta}_n^{(1)} - \theta_0) \\ n^{\frac{1}{2}}(\hat{\theta}_n^{(2)} - \theta_0) \\ \vdots \\ n^{\frac{1}{2}}(\hat{\theta}_n^{(K)} - \theta_0) \end{pmatrix} \xrightarrow{d} N\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma^{(1)} & \Sigma^{(2)} & \cdots & \Sigma^{(K)} \\ \Sigma^{(2)} & \Sigma^{(2)} & \cdots & \Sigma^{(K)} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma^{(K)} & \Sigma^{(K)} & \cdots & \Sigma^{(K)} \end{bmatrix} \right)$$

where

$$\Sigma^{(k)} = \int_0^\infty v(k, \theta_0, u)s^{(0)}(k, \theta_0, u)h_0(u)du.$$

Proof. In Theorem 3.2, we proved that the sequence of estimates which are the solution to estimating equations, are asymptotically normally distributed. We also showed that the canonical joint distribution holds when the matrices $\mathbf{A}^{(k)}$ and $\mathbf{B}^{(k)}$ of the regularity conditions are equal. Therefore, to prove that the canonical joint distribution of Definition 2.2 holds for this sequence of estimates which are the maximum partial likelihood estimates, we shall prove that the following conditions are satisfied:

1. For each $k = 1, \ldots, K$, $\hat{\theta}_n^{(k)} \xrightarrow{p} \theta_0$.
2. For each $k = 1, \ldots, K$, $n^{-\frac{1}{2}}U_n(k, \theta_0) \xrightarrow{d} N(0, \Sigma^{(k)})$.

3. For all $\theta^{*(k)}$ such that $\theta^{*(k)} \xrightarrow{p} \theta_0$,

$$n^{-1}\frac{\partial}{\partial\theta}U_n(k,\theta)|_{\theta=\theta^{*(k)}} = n^{-1}\mathcal{I}_n(k,\theta^{*(k)}) \xrightarrow{p} \Sigma^{(k)}$$

   for each $k = 1, \ldots, K$.

4. For $1 \leq k_1 \leq k_2 \leq K$,

$$Cov\left(n^{-\frac{1}{2}}U_n(k_1,\theta_0), n^{-\frac{1}{2}}U_n(k_2,\theta_0)\right) \xrightarrow{p} \Sigma^{(k_1)}$$

   .

The fixed sample results given in Theorem 3.5 are enough to prove that conditions 1–3 hold in this group sequential version.

It remains to show that condition 4 holds. Following the work of Tsiatis et al. (1995), we can split the counting process into distinct increments. We shall use the form of the score statistic at analysis k given by Equation (3.18). This is

$$U_n(k,\theta) = \int_0^{\tau_k}\sum_{i=1}^{n}\sum_{l=1}^{k}(Z_i(u) - E(k,\theta,u))dDN_i(l,u).$$

Similarly to the fixed sample proof, we can write the score statistic in martingale form. The j^{th} element of the martingale version of the score statistic times $n^{-\frac{1}{2}}$ with dependence on t is

$$W_j^{(n)}(k,\theta,t) = \int_0^{t}\sum_{i=1}^{n}\sum_{l=1}^{k}H_{ij}^{(n)}(k,u)dDM_i(l,u)$$

where

$$H_{ij}^{(n)}(k,u) = n^{-\frac{1}{2}}(Z_{ij}(u) - E_j(k,\theta,u)).$$

Then H is an $n \times p$ matrix of locally bounded predictable processes. Thus, $n^{-\frac{1}{2}}U_n(k,\theta) = W^{(n)}(k,\theta,\tau_k)$.

To assess the asymptotic limit of the predictable covariation process of the score statistic, we must first determine the range over which two processes are orthogonal; $DM_{i_1}(l_1,t)$ and $DM_{i_2}(l_2,t)$ are orthogonal for $i_1 \neq i_2$ because patients are independent and these processes are also orthogonal if $l_1 \neq l_2$ because the jump is unique for each patient and so cannot happen in in two different analyses. Therefore, the predictable covariation process for $k_1 \leq k_2$ is

$$\begin{aligned}
&\langle W_{j_1}^{(n)}(k_1,\theta_0), W_{j_2}^{(n)}(k_2,\theta_0)\rangle(t) \\
&= \int_0^t \sum_{i=1}^n \sum_{l=1}^{k_1} H_{ij_1}^{(n)}(k_1,u) H_{ij_2}^{(n)}(k_2,u) h_0(u) \exp\{\theta^T Z_i(u)\}(Y_i(l,u) - Y_i(l-1,u))du \\
&= \int_0^t \sum_{i=1}^n H_{ij_1}^{(n)}(k_1,u) H_{ij_2}^{(n)}(k_2,u) h_0(u) \exp\{\theta^T Z_i(u)\} Y_i(k_1,u)du \\
&= \int_0^t \sum_{i=1}^n H_{ij_1}^{(n)}(k_1,u) H_{ij_2}^{(n)}(k_1,u) h_0(u) \exp\{\theta^T Z_i(u)\} Y_i(k_1,u)du \\
&= \left(\int_0^t V(k_1,\theta_0,u) S^{(0)}(k_1,\theta_0,u) h_0(u)du \right)_{j_1 j_2} \\
&\xrightarrow{p} \Sigma_{j_1 j_2}^{(k_1)}.
\end{aligned}$$

Therefore we have the result that for $k_1 \leq k_2$

$$Cov\left(n^{-\frac{1}{2}}U_n(k_1,\theta_0), n^{-\frac{1}{2}}U_n(k_2,\theta_0)\right) \xrightarrow{p} \Sigma^{(k_1)}.$$

The second condition for Rebolledo's central limit theorem for martingales can be proven using the same reasoning as for the fixed sample case. □

Theorem 3.6 shows that the sequence of estimates $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(K)}$ follow the multivariate version of the canonical joint distribution. The ingredients of this known result are what we shall use in Chapter 4 to obtain the asymptotic distribution for the sequence of treatment effect estimates which are obtained using the "conditional score method". The conditional score method is used to obtain estimates for parameters in a joint model for longitudinal and survival data.

CHAPTER 4

JOINT MODEL WITH TREATMENT DIRECTLY AFFECTING SURVIVAL

4.1 | Joint modelling

4.1.1 | Motivation for combining survival and longitudinal information

The focus of this Chapter is to develop methods for designing and analysing a group sequential trial based on a joint model for longitudinal and survival data. We may believe that a trend in the trajectory of the biomarker is predictive of survival, and we would like to know whether this additional longitudinal data can be used in monitoring the trial, leading to early stopping.

Interest in joint modelling is motivated by clinical trials where the biomarker is predictive of survival. For example, Goldman et al. (1996) use CD4 lymphocyte cell counts as a surrogate endpoint for survival in a clinical trial comparing the efficacy and safety of two antiviral drugs for HIV-infected patients. Taylor et al. (2013) use a joint model to predict survival times of patients with prostate cancer based on prostate specific antigen (PSA) levels measured by blood tests at multiple hospital visits.

Suppose that the biomarker observations are available but have not been used in the analysis. The topic of this Chapter is to assess the change in efficiency of the trial when these observations are included in the analysis. We shall focus on efficiency measured in terms of the number of patients that need be recruited to achieve a certain power, and we show that, in some scenarios, the trial using the longitudinal data is up to 1.67 times as efficient as the trial which discards the longitudinal data. That is, when the longitudinal data is not used, 1.67 times as many patients are recruited to achieve the same power as the trial which does use the longitudinal data. We shall present these results in Section 4.5.

Tsiatis and Davidian (2001) present the joint model that we shall use here. Then, Tsiatis and Davidian (2004) give an overview of possible methods of inference. Rizopoulos (2012) gives further detailed theory for some of the inference options. We focus on a method called the “conditional score” method. This was first introduced by Tsiatis and Davidian (2001), who present the asymptotic theory for a fixed sample trial. Lu and Tsiatis (2008) use the theory of semiparametrics to find an estimator for the treatment effect in a survival model when there are auxiliary covariates incorporated. In this case, the auxiliary variables are known to be correlated with the time-to-event outcome and the information for these covariates is collected at baseline and throughout the trail. The authors derive an estimator that is more efficient than the maximum partial likelihood estimator for the model where information arises solely from time-to-event observations. We seek a similar result for the estimator of the conditional score method for analysing the joint model. The conditional score function is seen to have similar properties to the Estimating Equation (11) of Lu and Tsiatis (2008) where the biomarker observations play a similar role to the auxiliary variables. However, the reason why these results cannot be directly applied is because of the measurement error in the longitudinal data.

The conditional score requires an understanding of survival analysis because it is an analogy to partial likelihood. The ideas and methodology for survival analysis which appeared in Section 3.2 are extended upon in this section. The novel aspect of the research in this chapter is the design and analysis of group sequential trials based on a joint model for longitudinal and time-to-event data. Our main result, given by Theorem 4.4, determines the asymptotic distribution of the sequence of treatment effect estimates in a group sequential test based on the joint model.

4.1.2 | Joint model

The joint model that we consider is given in Equation (2) of Tsiatis and Davidian (2001). There are two processes in this model which represent the survival and longitudinal parts separately, and these processes are linked through the hazard rate of the survival process. First we consider the longitudinal data. Suppose that $X_i(t)$ is the true value of the biomarker at time t for subject i and that $W_i(t)$ is the observed value of the biomarker at time t for patient i. Then the longitudinal model takes the form

$$X_i(t) = b_{0i} + b_{1i}t \tag{4.1}$$

$$W_i(t) = X_i(t) + \epsilon_i(t) \tag{4.2}$$

where $\mathbf{b}_i = (b_{i0}, b_{i1})$ is a vector of patient specific random effects and $\epsilon_i(t)$ is the measurement error. In general, the vector $\mathbf{b}_i$ can have dimension p and the function $X_i(t)$ need not be constrained to linear functions in t. However, we shall concentrate on the case given in Equations (4.1) and (4.2). We consider a random effects model where each $\mathbf{b}_i$ is a random variable with density function $f(\cdot)$. The measurement errors are assumed to be independent and if the biomarker for patient i is measured at times $t_{i1}, \dots t_{im_i}$, then $\epsilon_i(t_{ij})|\mathbf{b}_i \sim N(0, \sigma^2)$ for $j = 1, \dots, m_i$ and $\epsilon(t)$ and $\epsilon(t')$ are independent for $t \neq t'$.

The model for the survival endpoint is a Cox proportional hazards model in which the longitudinal variable $X_i(t)$ acts as a time-varying covariate with coefficient γ. The remaining covariates for patient i at time t are given by the $p \times 1$ column vector $\mathbf{Z}_i$ which has corresponding coefficient vector $\boldsymbol{\eta}$ of length p. In the upcoming examples, we only consider the case where $p = 1$ so that Z_i and η are scalars, hence we do not use bold notation for these parameters. Then the hazard function is given by

$$h_i(t) = h_0(t) \exp\{\gamma X_i(t) + \eta^T Z_i\}, \tag{4.3}$$

where $h_0(\cdot)$ is the baseline hazard function. Note that it is the true underlying trajectory $X_i(t)$ that is included as a covariate in the proportional hazards model, whereas the measurements $W_i(t)$ with added error are observed. Together, Equations (4.1), (4.2) and (4.3) define the joint model.

It is surprising, perhaps, to assume treatment has no effect on the biomarker. However, this model appears to be widely accepted and there are many results for the more simple model presented above. Hence, we shall build upon existing literature to develop the group sequential methodology for the model (4.1)-(4.3).

The data collected for each individual $i = 1, \ldots, n$ are the vector $(W_i(t), Z_i(t), t_i, \delta_i)$ where

- $W_i(t_{ij})$ are biomarker measurements at times $t_{i1}, \ldots, t_{im_i}$
- Z_i are known covariates,
- t_i is the observed event time,
- $\delta_i = \mathbb{I}\{F_i \leq C_i\}$ is the indicator function for censoring, so that $\delta_i = 1$ implies an exact observation.

4.1.3 | Model likelihood

An expression for the full likelihood for the joint model is now presented. The full likelihood is not used in the analysis of the joint model, however understanding the structure reveals the difficulties with maximum likelihood based analyses for the joint model.

Tsiatis and Davidian (2004) derive an expression for the full likelihood function for a general joint model presented in their Section 3, Equation (7). We build upon this result and derive the likelihood function for our model Equations (4.1)—(4.3) as

$$\prod_{i=1}^{n} \int h_0(t_i) \exp\{\gamma(b_{0i} + b_{1i}t_i) + \eta^T Z_i\}^{\delta_i} \exp\left[-\int_0^{t_i} h_0(u) \exp\{\gamma(b_{0i} + b_{1i}u) + \eta^T Z_i\} du\right]$$
$$\times \frac{1}{(2\pi\sigma^2)^{m_i/2}} \exp\left[-\sum_{j=1}^{m_i} \frac{(W_i(t_{ij}) - (b_{i0} + b_{i1}t_{ij}))^2}{2\sigma^2}\right] f(\mathbf{b}_i) d\mathbf{b}_i.$$

Analytic expressions for calculation of the above likelihood are rarely available, so we rely on numerical integration techniques. The vector $\mathbf{b}_i$ has length $p = 2$ and hence this likelihood is a product of n 2-dimensional integrals. Calculation of the maximum likelihood estimate requires a series of these computationally expensive calculations. Further, note that calculation of this likelihood function requires that the baseline hazard function $h_0(t)$ is known.

To overcome these computational challenges, we shall introduce the conditional score method which is an analogy to maximising the partial likelihood in survival analysis. The similarity with maximum likelihood is that the analysis is "semi-parametric" and does not require specification of the form of $h_0(t)$. In Section 5.3.1 we discuss some of the complications of using a fully parametric approach with regards to parameter identifiability and robustness. For example, we specify that the model has a piecewise constant baseline hazard function and we must ensure that there are a sufficient number of events occurring between analysis times and knot points. Further, using the conditional score method, there is no integration over the distribution of the random effects which means that we do not need to make any assumptions about the distribution of the random effects.

In comparison with the fully parametric maximum likelihood estimation, the conditional score method has the disadvantage that the resulting estimator is not efficient. This has the consequence that the canonical joint distribution does not hold for the sequence of estimates in the group sequential test. However, we believe that the advantages of a semi-parametric estimator and being able to bypass any assumptions about the random effects outweigh the relative disadvantage of the conditional score estimator not being efficient.

4.1.4 | Counting processes for the joint model

In Section 3.2.5, it was shown how counting processes can be used to formulate the partial likelihood for the Cox proportional hazards model. The re-parameterisation of the data into the counting process framework allowed martingale results to be used in the derivation of the asymptotic distribution of the treatment effect estimate. We now show the analogous counting process for the joint model.

With our choice of model $X_i(t) = b_{0i} + b_{1i}t$ for longitudinal data, the joint model counting process is defined by the following functions:

$$\begin{aligned} N_i(t) &= \mathbb{I}(t \leq t_i, \delta_i = 1, t_{i2} \leq t) \\ dN_i(t) &= \mathbb{I}(t \leq t_i \leq t + dt, \delta_i = 1, t_{i2} \leq t). \end{aligned}$$

Note that here t_i is the time of the exact or censored event for patient i while t_{i2} is the second term in the sequence of times $t_{i1}, t_{i2}, \ldots, t_{im_i}$ at which we observe the biomarker. The difference between this counting process and the counting process for standard partial likelihood is the inclusion criterion $t_{i2} \leq t$. This ensures that there are a sufficient number of longitudinal observations to perform linear regression of the longitudinal data at a time t where the process $N_i(t)$ jumps. Note that this criterion would be modified for a different longitudinal model. This gives rise to a modified at-risk function. The criterion for inclusion in the risk set is now dependent on having enough longitudinal measurements, so the at-risk function is given by

$$Y_i(t) = \mathbb{I}\{t_i \geq t, t_{i2} \leq t\}.$$

By analogy to survival analysis, we seek a compensated counted process with expectation zero. This property leads us to define an estimating equation from which we can obtain treatment effect estimates that are asymptotically normally distributed. In the usual survival analysis setting, we can calculate the compensated counting process by subtracting the intensity process of the counting process. In the joint modelling setting, this is not as simple because the randomness of the nuisance parameters $\mathbf{b}$ mean that the intensity process will not be predictable. To overcome this, Tsiatis and Davidian (2001) introduce a "conditional intensity process" which is conditional on a certain "sufficient statistic". The origins of the conditional intensity process and sufficient statistic are not crucial for our purpose. What we actually use are the definitions and properties that are derived from these. In what follows, we shall introduce these two functions, given by Tsiatis and Davidian (2001), and the compensated counting process. We shall then show that this compensated counting process has expectation zero.

For patient i, let $t_i(u)$ be set of all time points for measurements of the biomarker, up to and including time u. Let $\hat{X}_i(u)$ be the ordinary least squares estimate of $X_i(u)$ for patient i based on the set of measurements taken at times $t_i(u)$. That is, calculate $\hat{b}_{0i}(u)$ and $\hat{b}_{1i}(u)$ based on measurements taken at times $t_i(u)$, then $\hat{X}_i(u) = \hat{b}_{0i}(u) + \hat{b}_{1i}(u)u$. As we pass time t_{ij}, a new observation W_{ij} is included and the formula for $\hat{X}_i(u)$ is updated for larger values of u. This seems strange since at early time points, not all of the available data is used for calculation of $\hat{X}_i(u)$ however this is necessary for the martingale property to hold in later results. Suppose that $\sigma^2\theta_i(u)$ is the variance of the estimator $\hat{X}_i(u)$ at time u. Tsiatis and

Davidian (2001) define the sufficient statistic to be the function

$$\begin{aligned} S_i(t,\gamma,\sigma^2) &= \hat{X}_i(t) + \gamma\sigma^2\theta_i(t)dN_i(t) \\ &= \hat{b}_{0i}(t) + \hat{b}_{1i}(t)t + \gamma\sigma^2\theta_i(t)dN_i(t) \end{aligned}$$

which is defined for all $t \in (t_{i2}, t_i)$ for patient i. The authors give details for the derivation of the conditional intensity process in their appendix, which is given by

$$\lambda_i^C(t) = lim_{dt\downarrow 0}\frac{\mathbb{P}(dN_i(t) = 1|S_i(t,\gamma,\sigma^2), t_i(t), Z_i, Y_i(t))}{dt} \tag{4.4}$$

$$= h_0(t)\exp\{\gamma S_i(t,\gamma,\sigma^2) - \gamma^2\sigma^2\theta_i(t)/2 + \eta^T Z_i\}Y_i(t). \tag{4.5}$$

The superscript c here is to show that this intensity process is conditional on the sufficient statistic and also to reflect that this is not the same intensity process as Equation (3.11) for the survival counting process. The proof that Equations (4.4) and (4.5) are equal is found in the appendix of Tsiatis and Davidian (2001) where the authors derive this conditional intensity process.

For convenience we shall define

$$E_{0i}(t,\gamma,\eta,\sigma^2) = \exp\{\gamma S_i(t,\gamma,\sigma^2) - \gamma^2\sigma^2\theta_i(t)/2 + \eta^T Z_i\} \tag{4.6}$$

which mirrors the exponential function in the intensity process for the survival data model. In Section 3.2.4 we saw how the Doob Meyer decomposition theorem is used to create a martingale that is the intensity process subtracted from its submartingale counting process. Although for the joint model, the compensated counting process will not be a martingale, because of the unknown nature of the random effects, it is still useful to define:

$$M_i(t) = N_i(t) - \int_0^t h_0(u)E_{0i}(u,\gamma,\eta,\sigma^2)Y_i(u)du \tag{4.7}$$

$$dM_i(t) = dN_i(t) - h_0(t)E_{0i}(t,\gamma,\eta,\sigma^2)Y_i(t)dt. \tag{4.8}$$

The important property for the asymptotic distribution derivation, is that the compensated counting process has expectation zero conditional on the sufficient statistic. The following lemma proves this result which shall be used in the proof of the asymptotic distribution of the parameter estimates in the joint model.

Lemma 4.1. *The function $dM_i(t)$ which is the compensated counting process for the joint model defined in equation* (4.8) *has expectation zero conditional on the* *sufficient statistic, that is*

$$\mathbb{E}(dM_i(t)|S_i(t,\gamma,\sigma^2),t_i(t),Z_i,Y_i(t))=0.$$

Proof. First, by equations (4.4) and (4.5), the conditional expectation of the counting process is

$$\begin{aligned}\mathbb{E}(dN_i(t)|S_i(t,\gamma,\sigma^2),t_i(t),Z_i(t),Y_i(t)) &= \mathbb{P}(dN_i(t)=1|S_i(t,\gamma,\sigma^2),t_i(t),Z_i(t),Y_i(t))\\ &= h_0(t)E_{0i}(t,\gamma,\eta,\sigma^2)Y_i(t)dt.\end{aligned}$$

Then the conditional expectation of the compensated counting process is

$$\begin{aligned}&\mathbb{E}(dM_i(t)|S_i(t,\gamma,\sigma^2),t_i(t),Z_i(t),Y_i(t))\\ =&\mathbb{E}(dN_i(t)-h_0(t)E_{0i}(t,\gamma,\eta,\sigma^2)Y_i(t)dt|S_i(t,\gamma,\sigma^2),t_i(t),Z_i(t),Y_i(t))\\ =&\mathbb{E}(dN_i(t)|S_i(t,\gamma,\sigma^2),t_i(t),Z_i(t),Y_i(t))-h_0(t)E_{0i}(t,\gamma,\eta,\sigma^2)Y_i(t)dt\\ =&0.\end{aligned}$$

□

Following the work of Tsiatis and Davidian (2001), we have presented definitions for the functions $S_i(t,\gamma,\sigma^2),\lambda_i^C(t),E_{0i}(t,\gamma,\eta,\sigma^2)$ and $M_i(t)$. We have then shown that $M_i(t)$ conditional on $S_i(t,\gamma,\sigma^2),t_i(t),Z_i(t),Y_i(t)$, has expectation zero and this result will be used in the proof of the asymptotic distribution for the parameter estimates in the joint model.

4.1.5 | Conditional score

It remains to define the conditional score function. This is the function we are seeking with a root that defines unbiased and asymptotically normal parameter estimates for the joint model. We introduce the conditional score before proving the distributional results for the treatment effect estimates in the upcoming Section 4.2.1 and Section 4.2.2. The conditional score is an analogue to the score for partial likelihood in survival analysis. The score for partial likelihood is a function of counting processes and their intensity functions. The conditional score is different because the intensity process is conditional on the sufficient statistic.

In the score statistic for partial likelihood, the function $E(\theta,t)$ is the expectation of the covariates at time t weighted by the intensity process $\lambda_i(t)$. This function is the main aspect of the score for partial likelihood and this concept motivates the conditional score approach. The underlying idea for the conditional score is to create the expectation of the covariates in the joint model weighted by the conditional intensity process $\lambda_i^C(t)$. However, the joint model includes the true, longitudinal trajectory as a covariate, and so this is replaced by the sufficient statistic. Therefore, we construct a function that is the expectation of the vector $\{S_i(t,\gamma,\sigma^2), Z_i^T\}^T$ weighted with probability $\lambda_i^C(t)$. With $E_{0i}(t,\gamma,\eta,\sigma^2)$ given in Equation (4.6), we define functions

$$\begin{aligned}
S_c^{(0)}(t,\gamma,\eta,\sigma^2) &= \frac{1}{n}\sum_{i=1}^{n} Y_i(t)E_{0i}(t,\gamma,\eta,\sigma^2)\\
S_c^{(1)}(t,\gamma,\eta,\sigma^2) &= \frac{1}{n}\sum_{i=1}^{n}\left\{\begin{array}{c} S_i(t,\gamma,\sigma^2)\\ Z_i\end{array}\right\} Y_i(t)E_{0i}(t,\gamma,\eta,\sigma^2)\\
S_c^{(2)}(t,\gamma,\eta,\sigma^2) &= \frac{1}{n}\sum_{i=1}^{n}\left\{\begin{array}{c} S_i(t,\gamma,\sigma^2)\\ Z_i\end{array}\right\}\left\{\begin{array}{c} S_i(t,\gamma,\sigma^2)\\ Z_i\end{array}\right\}^T Y_i(t)E_{0i}(t,\gamma,\eta,\sigma^2)\\
E_c(t,\gamma,\eta,\sigma^2) &= \frac{S_c^{(1)}(t,\gamma,\eta,\sigma^2)}{S_c^{(0)}(t,\gamma,\eta,\sigma^2)}.
\end{aligned}$$

The notation for this section reflects that of Jennison and Turnbull (1997) and Section 3.2.5 and the analogy to partial likelihood can be seen, in each case the subscript c indicates that these functions are conditional on the sufficient statistic. For comparison with Tsiatis and Davidian (2001), the functions $S_c^{(0)}, S_c^{(1)}$ and E_c are equivalent to E_0, E_1 and $\bar{S}$ respectively. These functions are the basis of the conditional score function. The score statistic for partial likelihood is a sum over patients of covariate vectors minus the weighted expectation of the covariate vector. By this interpretation, it is clear the expectation must be zero. Similarly, we define the conditional score to be a sum over patients of the vector $\{S_i(t,\gamma,\sigma^2), Z_i^T\}^T$ minus its weighted expectation. Let p be the length of the vector Z_i for all $i = 1,\ldots,n$, then the conditional score statistic will be a $(p+1)\times 1$ column vector given by

$$\begin{aligned}
U_c(\gamma,\eta,\sigma^2) &= \int_0^\infty \sum_{i=1}^{n}\left(\{S_i(u,\gamma,\sigma^2), Z_i^T\}^T - E_c(u,\gamma,\eta,\sigma^2)\right) dN_i(u) && (4.9)\\
&= \int_0^\infty \sum_{i=1}^{n}\left(\{S_i(u,\gamma,\sigma^2), Z_i^T\}^T - E_c(u,\gamma,\eta,\sigma^2)\right) dM_i(u). && (4.10)
\end{aligned}$$

Our conditional score Equation (4.9) is equivalent to Equation (6) of Tsiatis and Davidian (2001) and Equation (4.10) is equivalent to Equation (7) of Tsiatis and Davidian (2001). The equivalence of these two equations holds by a similar result to Lemma 3.4, which states that the compensated counting process can replace the counting process in the score statistic.

Tsiatis and Davidian (2001) write the conditional score statistic $U_c(\gamma, \eta, \sigma^2)$ as

$$\int_0^\infty \sum_{i=1}^n \left(\{S_i(u, \gamma, \sigma^2), Z_i^T\}^T - e_c(u, \gamma, \eta, \sigma^2) \right) dM_i(u) \tag{4.11}$$

$$+ \int_0^\infty \sum_{i=1}^n \left(e_c(u, \gamma, \eta, \sigma^2) - E_c(u, \gamma, \eta, \sigma^2) \right) dM_i(u), \tag{4.12}$$

where $e_c(u, \gamma, \eta, \sigma^2)$ is the probabilistic limit of $E_c(u, \gamma, \eta, \sigma^2)$. We shall later present a set of regularity conditions which imply that this limit exists and that the function $E_c(u, \gamma, \eta, \sigma^2)$ converges pointwise to $e_c(u, \gamma, \eta, \sigma^2)$. This regularity condition is assumed to hold. Expressions (4.11) and (4.12) are equivalent to (8a) and (8b) of Tsiatis and Davidian (2001) respectively. The authors prove that n^{-1} times our Expression (4.12) converges in probability to zero in a neighbourhood of (γ_0, η_0) and deduce that the behaviour of the estimators which are solutions to the equation $U_c(\gamma, \eta, \sigma^2) = 0$ will be dictated by Expression (4.11). In the proofs to follow, we therefore focus on Expression (4.11) when determining the asymptotic distribution.

We shall now show that Expression (4.11) has expectation zero. Hence the conditional score statistic defines an estimating function and if we let the parameter estimate be the root of the equation where the conditional score is set equal to zero, then the parameter estimate will be asymptotically normally distributed.

Lemma 4.2. *Expression* (4.11) *has expectation zero. That is*

$$\mathbb{E}\left[\int_0^\infty \sum_{i=1}^n \left(\{S_i(u, \gamma, \sigma^2), Z_i^T\}^T - e_c(u, \gamma, \eta, \sigma^2) \right) dM_i(u) \right] = 0.$$

Proof. We define some notation that mimics the martingale notation in the proof of Theorem 3.5. For each $j = 1, \ldots, p+1$ where p is the length of the vector Z_i, define

$$W_j^{(n)}(t, \gamma, \eta, \sigma^2) = \int_0^t \sum_{i=1}^n H_{ij}^{(n)}(u) dM_i(u)$$

where

$$H_{ij}^{(n)} = n^{-\frac{1}{2}}(\{S_i(u, \gamma, \sigma^2), Z_i^T\}_j^T - e_c(u, \gamma, \eta, \sigma^2)_j).$$

Then H is a $(p+1) \times (p+1)$ matrix. Thus $n^{-\frac{1}{2}} U_c(\gamma, \eta, \sigma^2) = W^{(n)}(\infty, \gamma, \eta, \sigma^2)$. To determine the expectation of $W^{(n)}(\infty, \gamma, \eta, \sigma^2)$ we use the *i.i.d* nature of the vectors $\int_0^\infty H_1^{(n)}(u) dM_1(u), \ldots, \int_0^\infty H_n^{(n)}(u) dM_n(u)$ and restrict attention to $\int_0^\infty H_1^{(n)}(u) dM_1(u)$. The expectation of $\int_0^\infty H_1^{(n)}(u) dM_1(u)$ is shown below. The first equality follows since orders of integration can be exchanged under suitable regularity conditions, the second equality is an application of the law of total expectation, and the third equality is obtained by taking out known factors. Finally, Lemma 4.1 gives that the inner conditional expectation is zero, showing the result.

$$
\begin{aligned}
&\mathbb{E}\left(\int_0^\infty H_1^{(n)}(u)dM_1(u)\right)\\
=&\int_0^\infty \mathbb{E}(H_1^{(n)}(u)dM_1(u))\\
=&\int_0^\infty \mathbb{E}(\mathbb{E}[H_1^{(n)}(u)dM_1(u)|S_1(u,\gamma,\sigma^2),Z_1,t_1(u),Y_1(u)])\\
=&\int_0^\infty \mathbb{E}(H_1^{(n)}(u)\mathbb{E}[dM_1(u)|S_1(u,\gamma,\sigma^2),Z_1,t_1(u),Y_1(u)])\\
=&0.
\end{aligned}
$$

□

Combining Lemma 4.2 with the fact that Expression (4.12) converges in probability to zero, we have that $U_c(\gamma,\eta,\sigma^2)$ is an estimating function.

4.1.6 | Differentiation of the conditional score

Another object of importance is the first derivative of the conditional score function, Equation (4.9), with respect to parameters γ and η. This matrix plays a key role in the definition of the covariance matrix for the estimates $\hat{\gamma}$ and $\hat{\eta}$ and has a likeness to the Fisher information matrix which is the derivative of the score statistic for general statistical models. Tsiatis and Davidian (2001) describe that the variance matrix can be found, however they do not present an equation for such an object. The derivation of the derivative of the conditional score function in this section is original work.

First we consider differentiating the function $E_{0i}(t,\gamma,\eta,\sigma^2)$ with respect to γ

and η separately. These are

$$\begin{aligned}
\frac{\partial}{\partial\gamma}E_{0i}(t,\gamma,\eta,\sigma^2) &= \frac{\partial}{\partial\gamma}\exp\{\gamma S_i(t,\gamma,\sigma^2) - \gamma^2\sigma^2\theta_i(t)/2 + \eta Z_i\} \\
&= \frac{\partial}{\partial\gamma}\exp\{\gamma^2\sigma^2\theta_i(t)dN_i(t) + \gamma\hat{X}_i(t) - \gamma^2\sigma^2\theta_i(t)/2 + \eta Z_i\} \\
&= (2\gamma\sigma^2\theta_i(t)dN_i(t) + \hat{X}_i(t) - \gamma\sigma^2\theta_i(t))E_{0i}(t,\gamma,\eta,\sigma^2) \\
&= (2S_i(t,\gamma,\sigma^2) - \hat{X}_i(t) - \gamma\sigma^2\theta_i(t))E_{0i}(t,\gamma,\eta,\sigma^2) \\
\frac{\partial}{\partial\eta}E_{0i}(t,\gamma,\eta,\sigma^2) &= \frac{\partial}{\partial\eta}\exp\{\gamma S_i(t,\gamma,\sigma^2) - \gamma^2\sigma^2\theta_i(t)/2 + \eta Z_i\} \\
&= Z_i E_{0i}(t,\gamma,\eta,\sigma^2).
\end{aligned}$$

Derivatives of $E_{0i}(t,\gamma,\eta,\sigma^2)$ are needed for differentiating the function $S_c^{(0)}(t,\gamma,\eta,\sigma^2)$. Similarly for differentiation of the function $S_c^{(1)}(t,\gamma,\eta,\sigma^2)$ we now calculate derivatives for $S_i(t,\gamma,\eta,\sigma^2)E_{0i}(t,\gamma,\eta,\sigma^2)$ and $Z_iE_{0i}(t,\gamma,\eta,\sigma^2)$. These are

$$\begin{aligned}
\frac{\partial}{\partial\gamma}S_i(t,\gamma,\eta,\sigma^2)E_{0i}(t,\gamma,\eta,\sigma^2) &= \Big[S_i(t,\gamma,\sigma^2)^2 + S_i(t,\gamma,\eta,\sigma^2)(S_i(t,\gamma,\sigma^2) - \hat{X}_i(t) \\
&\qquad - \gamma\sigma^2\theta_i(t)) + \sigma^2\theta_i(t)dN_i(t)\Big]E_{0i}(t,\gamma,\eta,\sigma^2) \\
\frac{\partial}{\partial\gamma}Z_iE_{0i}(t,\gamma,\eta,\sigma^2) &= Z_i(2S_i(t,\gamma,\sigma^2) - \hat{X}_i(t) - \gamma\sigma^2\theta_i(t))E_{0i}(t,\gamma,\eta,\sigma^2) \\
\frac{\partial}{\partial\eta}S_i(t,\gamma,\eta,\sigma^2)E_{0i}(t,\gamma,\eta,\sigma^2) &= S_i(t,\gamma,\eta,\sigma^2)Z_iE_{0i}(t,\gamma,\eta,\sigma^2) \\
\frac{\partial}{\partial\eta}Z_iE_{0i}(t,\gamma,\eta,\sigma^2) &= Z_i^TZ_iE_{0i}(t,\gamma,\eta,\sigma^2).
\end{aligned}$$

We will display the relationship between the matrices $S_c^{(0)}(t,\gamma,\eta,\sigma^2)$, $S_c^{(1)}(t,\gamma,\eta,\sigma^2)$ and $S_c^{(2)}(t,\gamma,\eta,\sigma^2)$ and their derivatives. In the simple survival model in Section 3.2, we have that $S^{(1)} = \partial S^{(0)}/\partial\theta$ and $S^{(2)} = \partial S^{(1)}/\partial\theta$. This is not true for the conditional score, but a similar relationship holds. To make this clear, it convenient to derive additional functions. Let

$$J_i(t,\gamma,\sigma^2) = \{S_i(t,\gamma,\sigma^2) - \hat{X}_i(t) - \gamma\sigma^2\theta_i(t), 0, \ldots, 0\}^T$$

be a $(p+1)\times 1$ column vector and define functions

$$\begin{aligned}
C^{(1)}(t,\gamma,\eta,\sigma^2) &= \frac{1}{n}\sum_{i=1}^{n} J_i(t,\gamma,\sigma^2)Y_i(t)E_{0i}(t,\gamma,\eta,\sigma^2)\\
C^{(2)}(t,\gamma,\eta,\sigma^2) &= \frac{1}{n}\sum_{i=1}^{n} \left\{\begin{array}{c} S_i(t,\gamma,\sigma^2)\\ Z_i\end{array}\right\} J_i(t,\gamma,\sigma^2)^T Y_i(t)E_{0i}(t,\gamma,\eta,\sigma^2)\\
C^{(3)}(t,\gamma,\eta,\sigma^2) &= \frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} \sigma^2\theta_i(u)dN_i(u) & 0 & \cdots & 0\\ 0 & 0 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 0\end{bmatrix} Y_i(t)E_{0i}(t,\gamma,\eta,\sigma^2)\\
V_c^{(1)}(t,\gamma,\eta,\sigma^2) &= \frac{S_c^{(2)}(t,\gamma,\eta,\sigma^2)}{S_c^{(0)}(t,\gamma,\eta,\sigma^2)} - \frac{S_c^{(1)}(t,\gamma,\eta,\sigma^2)S_c^{(1)}(t,\gamma,\eta,\sigma^2)^T}{[S_c^{(0)}(t,\gamma,\eta,\sigma^2)]^2}\\
V_c^{(2)}(t,\gamma,\eta,\sigma^2) &= \frac{C^{(2)}(t,\gamma,\eta,\sigma^2)}{S_c^{(0)}(t,\gamma,\eta,\sigma^2)} - \frac{S_c^{(1)}(t,\gamma,\eta,\sigma^2)C^{(1)}(t,\gamma,\eta,\sigma^2)^T}{[S_c^{(0)}(t,\gamma,\eta,\sigma^2)]^2}\\
V_c^{(3)}(t,\gamma,\eta,\sigma^2) &= \frac{C^{(3)}(t,\gamma,\eta,\sigma^2)}{S_c^{(0)}(t,\gamma,\eta,\sigma^2)}.
\end{aligned}$$

The function $S_c^{(0)}$ is a scalar and its derivative will be a $(p+1)\times 1$ column vector. The matrices $S_c^{(1)}$ and $C^{(1)}$ are both $(p+1)\times 1$ column vectors. For differentiation, entry $(\partial S_c^{(1)}/\partial(\gamma,\eta)^T)_{i,j}$ represents the i^{th} element of $S_c^{(1)}$ differentiated with respect to the j^{th} element of the vector $(\gamma,\eta)^T$. The functions $S_c^{(2)}, C^{(2)}, C^{(3)}, V_c^{(1)}, V_c^{(2)}$ and $V_c^{(3)}$ are all $(p+1)\times(p+1)$ matrices. Using the above calculation, we find the following relationships hold

$$\begin{aligned}
&\frac{\partial}{\partial(\gamma,\eta)^T}S_c^{(0)}(t,\gamma,\eta,\sigma^2) = \frac{\partial}{\partial(\gamma,\eta)^T}\left[\frac{1}{n}\sum_{i=1}^{n}E_{0i}(t,\gamma,\eta,\sigma^2)Y_i(t)\right]\\
&\quad=\frac{1}{n}\sum_{i=1}^{n}\left\{\begin{array}{c} 2S_i(t,\gamma,\sigma^2)-\hat{X}_i(t)-\gamma\sigma^2\theta_i(t)\\ Z_i\end{array}\right\}E_{0i}(t,\gamma,\eta,\sigma^2)Y_i(t)\\
&\quad=\frac{1}{n}\sum_{i=1}^{n}\left\{\begin{array}{c} S_i(t,\gamma,\sigma^2)\\ Z_i\end{array}\right\}E_{0i}(t,\gamma,\eta,\sigma^2)Y_i(t)+\frac{1}{n}\sum_{i=1}^{n}J_i(t,\gamma,\sigma^2)E_{0i}(t,\gamma,\eta,\sigma^2)Y_i(t)\\
&\quad=S_c^{(1)}(t,\gamma,\eta,\sigma^2)+C^{(1)}(t,\gamma,\eta,\sigma^2)
\end{aligned}$$

A similar calculation is performed for the differentiation of $S_c^{(1)}(t,\gamma,\eta,\sigma^2)$. Temporarily removing dependence of all functions on parameters t,γ,η and σ^2, we have

$$\begin{aligned}
\frac{\partial}{\partial(\gamma,\eta)^T}S_c^{(1)}(t,\gamma,\eta,\sigma^2) &= \frac{\partial}{\partial(\gamma,\eta)^T}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\begin{array}{c} S_i(t,\gamma,\sigma^2) \\ Z_i \end{array}\right\} E_{0i}(t,\gamma,\eta,\sigma^2)Y_i(t)\right] \\
=&\frac{1}{n}\sum_{i=1}^{n}\left\{\begin{array}{cc} S_i^2 + S_i(S_i-\hat{X}_i-\gamma\sigma^2\theta_i)+\sigma^2\theta_i dN_i & S_iZ_i \\ Z_i(2S_i-\hat{X}_i-\gamma\sigma^2\theta_i) & Z_i^TZ_i \end{array}\right\} E_{0i}Y_i \\
=&\frac{1}{n}\sum_{i=1}^{n}\left\{\begin{array}{c} S_i \\ Z_i \end{array}\right\}\left\{\begin{array}{c} S_i \\ Z_i \end{array}\right\}^T E_{0i}Y_i + \frac{1}{n}\sum_{i=1}^{n}\left\{\begin{array}{c} S_i \\ Z_i \end{array}\right\}\left\{\begin{array}{c} S_i-\hat{X}-\gamma\sigma^2\theta_i \\ 0 \end{array}\right\}^T E_{0i}Y_i \\
+&\frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} \sigma^2\theta_i(u)dN_i(u) & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} E_{0i}Y_i \\
=&S_c^{(2)}(t,\gamma,\eta,\sigma^2)+C^{(2)}(t,\gamma,\eta,\sigma^2)+C^{(3)}(t,\gamma,\eta.\sigma^2).
\end{aligned}$$

We are now able to differentiate the function $E_c(t,\gamma,\eta,\sigma^2)$. By the quotient rule and the relationships derived above, we have

$$\begin{aligned}
&\frac{\partial}{\partial(\gamma,\eta)^T}E_c(t,\gamma,\eta,\sigma^2) = \frac{\partial}{\partial(\gamma,\eta)^T}\left[\frac{S_c^{(1)}(t,\gamma,\eta,\sigma^2)}{S_c^{(0)}(t,\gamma,\eta,\sigma^2)}\right] \\
&= \frac{\frac{\partial}{\partial(\gamma,\eta)^T}S_c^{(1)}(t,\gamma,\eta,\sigma^2)}{S_c^{(0)}(t,\gamma,\eta,\sigma^2)} - \frac{S_c^{(1)}(t,\gamma,\eta,\sigma^2)\frac{\partial}{\partial(\gamma,\eta)^T}S_c^{(0)}(t,\gamma,\eta,\sigma^2)}{\left[S_c^{(0)}(t,\gamma,\eta,\sigma^2)\right]^2} \\
&= \frac{S_c^{(2)}(t,\gamma,\eta,\sigma^2)}{S_c^{(0)}(t,\gamma,\eta,\sigma^2)} + \frac{C^{(2)}(t,\gamma,\eta,\sigma^2)}{S_c^{(0)}(t,\gamma,\eta,\sigma^2)} + \frac{C^{(3)}(t,\gamma,\eta,\sigma^2)}{S_c^{(0)}(t,\gamma,\eta,\sigma^2)} \\
&- \frac{S_c^{(1)}(t,\gamma,\eta,\sigma^2)S_c^{(1)}(t,\gamma,\eta,\sigma^2)^T}{\left[S_c^{(0)}(t,\gamma,\eta,\sigma^2)\right]^2} - \frac{S_c^{(1)}(t,\gamma,\eta,\sigma^2)C^{(1)}(t,\gamma,\eta,\sigma^2)^T}{\left[S_c^{(0)}(t,\gamma,\eta,\sigma^2)\right]^2} \\
&= V_c^{(1)}(t,\gamma,\eta,\sigma^2)+V_c^{(2)}(t,\gamma,\eta,\sigma^2)+V_c^{(3)}(t,\gamma,\eta,\sigma^2).
\end{aligned}$$

Finally, we can differentiate the conditional score of Equation (4.9). This is

$$
\begin{aligned}
&\frac{\partial}{\partial(\gamma,\eta)^T}U_c(\gamma,\eta,\sigma^2)\\
=&\frac{\partial}{\partial(\gamma,\eta)^T}\left[\int_0^\infty\sum_{i=1}^n\left(\{S_i(u,\gamma,\sigma^2),Z_i^T\}^T-E_c(u,\gamma,\eta,\sigma^2)\right)dN_i(u)\right]\\
=&\sum_{i=1}^n\int_0^\infty\begin{bmatrix}\sigma^2\theta_i(u) & 0 & \cdots & 0\\ 0 & 0 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 0\end{bmatrix}dN_i(u)\\
-&\sum_{i=1}^n\int_0^\infty\left[V_c^{(1)}(u,\gamma,\eta,\sigma^2)+V_c^{(2)}(u,\gamma,\eta,\sigma^2)+V_c^{(3)}(u,\gamma,\eta,\sigma^2)\right]dN_i(u). \qquad (4.13)
\end{aligned}
$$

4.2 | Asymptotic theory for the joint model

4.2.1 | Fixed sample results

In this section, we show that the estimates which are the root of the conditional score function are asymptotically normally distributed. The conditional score function is

$$
U_c(\gamma,\eta,\sigma^2)=\int_0^\infty\sum_{i=1}^n\left(\{S_i(u,\gamma,\sigma^2),Z_i^T\}^T-E_c(u,\gamma,\eta,\sigma^2)\right)dN_i(u)
$$

and we have so far shown that the conditional score function is an estimating function. Therefore, we shall make use of Theorem 3.1 to derive the asymptotic distribution of the resulting parameter estimates. Further, we then describe how to perform a hypothesis test making use of this distributional result.

In what follows we shall treat σ^2 as known and later we discuss consistent estimation of σ^2. Let $\hat{\gamma}_n$ and $\hat{\eta}_n$ be the values of γ and η respectively which are the solution to the equation $U_c^{(n)}(\gamma,\eta,\sigma^2)=0$. Dependence on the number of patients n is to clarify that we assess asymptotic results as $n\to\infty$. We follow the proof given by Tsiatis and Davidian (2001), who show that the estimates $\hat{\gamma}_n$ and $\hat{\eta}_n$ converge in distribution to a Gaussian random variable and we go beyond this proof by finding a specific form for the variance matrices and their probabilistic limits.

For the proof of the asymptotic distribution of the parameter estimates in the joint model, we require asymptotic limits of elements of the score statistic and variance matrices to exist. The definitions of these limits are specified through the following regularity conditions:

Conditions 4.1.

1. *There exist neighbourhoods* Γ *of* γ_0 *and* N *of* η_0 *and functions* $s_c^{(1)}(t,\gamma,\eta,\sigma^2), s_c^{(2)}(t,\gamma,\eta,\sigma^2), s_c^{(3)}(t,\gamma,\eta,\sigma^2), c^{(1)}(t,\gamma,\eta,\sigma^2), c^{(2)}(t,\gamma,\eta,\sigma^2)$ *and* $c^{(3)}(t,\gamma,\eta,\sigma^2)$ *defined on* $[0,\infty)\times\Gamma\times N$ *such that*

$$\sup_{t\in[0,\infty),\gamma\in\Gamma,\eta\in N}\left\|S_c^{(j)}(t,\gamma,\eta,\sigma^2)-s_c^{(j)}(t,\gamma,\eta,\sigma^2)\right\|\xrightarrow{p}0 \textit{ for } j=0,1,2$$
$$\sup_{t\in[0,\infty),\gamma\in\Gamma,\eta\in N}\left\|C^{(j)}(t,\gamma,\eta,\sigma^2)-c^{(j)}(t,\gamma,\eta,\sigma^2)\right\|\xrightarrow{p}0 \textit{ for } j=0,1,2.$$

2. *For each* $j=0,1,2$, $s_c^{(j)}(t,\gamma,\eta,\sigma^2)$ *and* $c^{(j)}(t,\gamma,\eta,\sigma^2)$ *are continuous functions of* $\gamma\in\Gamma$ *and* $\eta\in N$ *uniformly in* $t\in[0,\infty)$, *and bounded on* $[0,\infty)\times\Gamma\times N$. *Also,* $s_c^{(0)}(t,\gamma,\eta,\sigma^2)$ *and* $c^{(0)}(t,\gamma,\eta,\sigma^2)$ *are bounded away from zero on* $[0,\infty)\times\Gamma\times N$.

It is clear that the following relationships hold: $s_c^{(1)}$ and $s_c^{(2)}$ are the first and second derivatives of $s_c^{(0)}$ with respect to the vector $(\gamma,\eta^T)^T$; $c^{(1)}$ and $c^{(2)}$ are the first and second derivatives of $c^{(0)}$.

The probabilistic limits $e_c(t,\gamma,\eta,\sigma^2)$ of $E_c(t,\gamma,\eta,\sigma^2)$, $v_c^{(1)}(t,\gamma,\eta,\sigma^2)$ of $V_c^{(1)}(t,\gamma,\eta,\sigma^2)$ and $v_c^{(2)}(t,\gamma,\eta,\sigma^2)$ of $V_c^{(2)}(t,\gamma,\eta,\sigma^2)$ are define by the following:

$$\begin{aligned}
e_c(t,\gamma,\eta,\sigma^2) &= \frac{s_c^{(1)}(t,\gamma,\eta,\sigma^2)}{s_c^{(0)}(t,\gamma,\eta,\sigma^2)}\\
v_c^{(1)}(t,\gamma,\eta,\sigma^2) &= \frac{s_c^{(2)}(t,\gamma,\eta,\sigma^2)}{s_c^{(0)}(t,\gamma,\eta,\sigma^2)}-\frac{s_c^{(1)}(t,\gamma,\eta,\sigma^2)s_c^{(1)}(t,\gamma,\eta,\sigma^2)^T}{[s_c^{(0)}(t,\gamma,\eta,\sigma^2)]^2}\\
v_c^{(2)}(t,\gamma,\eta,\sigma^2) &= \frac{c^{(2)}(t,\gamma,\eta,\sigma^2)}{s_c^{(0)}(t,\gamma,\eta,\sigma^2)}-\frac{s_c^{(1)}(t,\gamma,\eta,\sigma^2)c^{(1)}(t,\gamma,\eta,\sigma^2)^T}{[s_c^{(0)}(t,\gamma,\eta,\sigma^2)]^2}.
\end{aligned}$$

We now prove that Theorem 3.1 applies to the conditional score estimating equation by proving that 1–3 of Conditions 3.1 hold. The remaining conditions are assumed to hold to avoid technical distractions.

Theorem 4.3. *Suppose that* γ_0,η_0 *and* σ_0^2 *are the true values of the parameters* γ,η,σ^2 *respectively. Let* $\hat{\gamma}_n$ *and* $\hat{\eta}_n$ *be the values of* γ *and* η *which are the solution to the equation* $U_c^{(n)}(\gamma,\eta,\sigma_0^2)=0$ *and suppose that the regularity conditions of 4.1 hold.*

Then the vector $(\hat{\gamma}_n, \hat{\eta}_n)^T$ *converges in distribution to Gaussian random variable, specifically*

$$\sqrt{n}\left[\begin{pmatrix}\hat{\gamma}_n\\ \hat{\eta}_n\end{pmatrix} - \begin{pmatrix}\gamma_0\\ \eta_0\end{pmatrix}\right] \xrightarrow{d} N(\mathbf{0}, \Sigma)$$

where

$$\Sigma = A^{-1}B(A^{-1})^T$$

and the matrices A *and* B *are defined by*

$$A = \int_0^\infty \begin{bmatrix}\sigma_0^2\mathbb{E}(\theta_i(u)) & 0 & \cdots & 0\\ 0 & 0 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 0\end{bmatrix} s_c^{(0)}(u, \gamma_0, \eta_0, \sigma_0^2)h_0(u)du$$

$$- \int_0^\infty \left(v_c^{(1)}(u, \gamma_0, \eta_0, \sigma_0^2) + v_c^{(2)}(u, \gamma_0, \eta_0, \sigma_0^2)\right) s_c^{(0)}(u, \gamma_0, \eta_0, \sigma_0^2)h_0(u)du \qquad (4.14)$$

$$B = \int_0^\infty v_c^{(1)}(u, \gamma_0, \eta_0, \sigma_0^2)s_c^{(0)}(u, \gamma_0, \eta_0, \sigma_0^2)h_0(u)du. \qquad (4.15)$$

Proof. In this theorem, the conditional score function, $U_c^{(n)}(\gamma, \eta, \sigma^2)$, plays the role of the estimating function $\mathbf{G}_n(\boldsymbol{\theta}, \mathbf{x}_n)$ and the vector of parameters $(\gamma, \eta, \sigma^2)^T$ is represented by $\boldsymbol{\theta}$ in Theorem 3.1. In applying Theorem 3.1, we show the following conditions are satisfied:

A1 $\hat{\gamma}_n \xrightarrow{p} \gamma_0$ and $\hat{\eta}_n \xrightarrow{p} \eta_0$.

A2 $n^{-\frac{1}{2}}U_c^{(n)}(\gamma_0, \eta_0, \sigma_0^2) \xrightarrow{d} N(0, B)$.

A3 For all γ_n^* and η_n^* such that $\gamma_n^* \xrightarrow{p} \gamma_0$ and $\eta_n^* \xrightarrow{p} \eta_0$,

$$\frac{1}{n}\frac{\partial}{\partial(\gamma, \eta^T)^T}U_c^{(n)}(\gamma, \eta, \sigma_0^2)|_{\gamma=\gamma^*, \eta=\eta^*} \xrightarrow{p} A.$$

We shall focus on the proof that conditions A2 and A3 are satisfied since we shall later build upon this proof in the group sequential case. The condition A1, the proof of consistency is given by Van der Vaart (2000) in their Section 5.2, who prove that consistency holds for any estimator which is the root of an estimating equation.

For the remainder of the proof, we follow the argument by Tsiatis and Davidian (2001), that the asymptotic distribution of the conditional score function is dictated by the following function:

$$U_c(\gamma, \eta, \sigma^2) = \int_0^\infty \sum_{i=1}^n \left(\{S_i(u, \gamma, \sigma^2), Z_i^T\}^T - e_c(u, \gamma, \eta, \sigma^2)\right) dM_i(u).$$

Further, we shall use the same set-up as Lemma 4.2. As a reminder, for each $j = 1, \dots, p+1$, let

$$W_j^{(n)}(t, \gamma, \eta, \sigma^2) = \int_0^t \sum_{i=1}^n H_{ij}^{(n)}(u, \gamma, \eta, \sigma^2) dM_i(u)$$

where

$$H_{ij}^{(n)}(u, \gamma, \eta, \sigma^2) = n^{-\frac{1}{2}}(\{S_i(u, \gamma, \sigma^2), Z_i^T\}_j^T - e_c(u, \gamma, \eta, \sigma^2)_j).$$

Thus $n^{-\frac{1}{2}} U_c(\gamma, \eta, \sigma^2) = W^{(n)}(\infty, \gamma, \eta, \sigma^2)$.

For condition A2, we have shown that $\mathbb{E}\left(n^{-\frac{1}{2}} U_c^{(n)}(\gamma_0, \eta_0, \sigma_0^2)\right) = 0$ by Lemma 4.2. The vectors

$$\int_0^\infty H_1^{(n)}(u, \gamma_0, \eta_0, \sigma_0^2) dM_1(u), \dots, \int_0^\infty H_n^{(n)}(u, \gamma_0, \eta_0, \sigma_0^2) dM_n(u)$$

are independent and identically distributed ($i.i.d$). This is because patients are independent and there is an underlying population wide distribution for the covariates, meaning that each vector has the same probability density function. Therefore, by the central limit theorem, $W^{(n)}(\infty, \gamma_0, \eta_0, \sigma_0^2)$ converges in distribution to a normal random variable.

The covariance matrix of the limiting distribution of the function $W^{(n)}(\infty, \gamma_0, \eta_0, \sigma_0^2)$ is now determined. For all $j = 1, \dots, p+1$ we have that $\mathbb{E}\left(W_j^{(n)}(\infty, \gamma_0, \eta_0, \sigma_0^2)\right) = 0$ by Lemma 4.2. Further, because patients are independent the covariance between elements j_1 and j_2 is

$$\begin{aligned} &Cov\left(W_{j_1}^{(n)}(\infty, \gamma_0, \eta_0, \sigma_0^2), W_{j_2}^{(n)}(\infty, \gamma_0, \eta_0, \sigma_0^2)\right) \\ =&\mathbb{E}\left(\sum_{i=1}^n \int_0^\infty H_{ij_1}^{(n)}(u, \gamma_0, \eta_0, \sigma_0^2) dN_i(u) \int_0^\infty H_{ij_2}^{(n)}(u, \gamma_0, \eta_0, \sigma_0^2) dN_i(u)\right). \end{aligned}$$

In the calculation for the covariance below, the first equality uses the fact that $dN_i(u)$ is an indicator function so that $dN_i(u)^2 = dN_i(u)$. Then, using a similar method to the calculation of the expectation, we use the law of total expectation to derive the following expression for covariance

$$\begin{aligned}
&\mathbb{E}\left(\sum_{i=1}^{n}\int_0^\infty H_{ij_1}^{(n)}(u,\gamma_0,\eta_0,\sigma_0^2)dN_i(u)\int_0^\infty H_{ij_2}^{(n)}(u,\gamma_0,\eta_0,\sigma_0^2)dN_i(u)\right)\\
=&\mathbb{E}\left(\sum_{i=1}^{n}\int_0^\infty H_{ij_1}^{(n)}(u,\gamma_0,\eta_0,\sigma_0^2)H_{ij_2}^{(n)}(u,\gamma_0,\eta_0,\sigma_0^2)dN_i(u)\right)\\
=&\mathbb{E}\left(\sum_{i=1}^{n}\int_0^\infty \mathbb{E}\left[H_{ij_1}^{(n)}(u,\gamma_0,\eta_0,\sigma_0^2)H_{ij_2}^{(n)}(u,\gamma_0,\eta_0,\sigma_0^2)dN_i(u)|S_i(u,\gamma_0,\sigma_0^2),Z_i,t_i(u),Y_i(u)\right]\right)\\
=&\mathbb{E}\left(\sum_{i=1}^{n}\int_0^\infty H_{ij_1}^{(n)}(u,\gamma_0,\eta_0,\sigma_0^2)H_{ij_2}^{(n)}(u,\gamma_0,\eta_0,\sigma_0^2)\mathbb{E}\left[dN_i(u)|S_i(u,\gamma_0,\sigma_0^2),Z_i,t_i(u),Y_i(u)\right]\right)\\
=&\mathbb{E}\left(\sum_{i=1}^{n}\int_0^\infty H_{ij_1}^{(n)}(u,\gamma_0,\eta_0,\sigma_0^2)H_{ij_2}^{(n)}(u,\gamma_0,\eta_0,\sigma_0^2)h_0(u)E_{0i}(u,\gamma_0,\eta_0,\sigma_0^2)Y_i(u)du\right).
\end{aligned}$$

For the following calculation, dependency on parameters γ_0,η_0 and σ_0^2 is removed from the functions $S_c^{(0)}(t),S_c^{(1)}(t)$ and $S_c^{(2)}(t)$ for notational purposes. Further, the dependency on parameters γ_0,η_0 and σ_0^2 is also removed from the probabilistic limits $s_c^{(0)}(t),s_c^{(1)}(t),s_c^{(2)}(t)$ and $e_c(t)$. Then following calculation holds

$$\begin{aligned}
&\sum_{i=1}^{n}H_{ij_1}^{(n)}(u,\gamma_0,\eta_0,\sigma_0^2)H_{ij_2}^{(n)}(u,\gamma_0,\eta_0,\sigma_0^2)h_0(u)E_{0i}(u,\gamma_0,\eta_0,\sigma_0^2)Y_i(u)du\\
=&\sum_{i=1}^{n}n^{-1}\Big(\{S_i(u,\gamma_0,\sigma_0^2),Z_i^T\}_{j_1}^T-e_c(u)_{j_1}\Big)\\
&\qquad\times\Big(\{S_i(u,\gamma_0,\sigma_0^2),Z_i^T\}_{j_2}^T-e_c(u)_{j_2}\Big)h_0(u)E_{0i}(u,\gamma_0,\eta_0,\sigma_0^2)Y_i(u)\\
=&\Bigg([S_c^{(2)}(u)]_{j_1j_2}-\frac{[s_c^{(1)}(u)]_{j_1}}{s_c^{(0)}(u)}[S_c^{(1)}(u)]_{j_2}-\frac{[s_c^{(1)}(u)]_{j_2}}{s_c^{(0)}(u)}[S_c^{(1)}(u)]_{j_1}\\
&\qquad+\frac{[s_c^{(1)}(u)]_{j_1}}{s_c^{(0)}(u)}\frac{[s_c^{(1)}(u)]_{j_2}}{s_c^{(0)}(u)}S_c^{(0)}(u)\Bigg)h_0(u)\\
\xrightarrow{p}&\left(\frac{[s_c^{(2)}(u)]_{j_1j_2}}{s_c^{(0)}(u)}-\frac{[s_c^{(1)}(u)]_{j_1}[s_c^{(1)}(u)]_{j_2}}{s_c^{(0)\otimes 2}(u)}\right)s_c^{(0)}(u)h_0(u)\\
=&[v_c^{(1)}(u)]_{j_1j_2}s_c^{(0)}(u)h_0(u).
\end{aligned}$$

Therefore, combining the above results, it can be seen that

$$\begin{aligned}
&Cov\left(W_{j_1}^{(n)}(\infty,\gamma_0,\eta_0,\sigma_0^2), W_{j_2}^{(n)}(\infty,\gamma_0,\eta_0,\sigma_0^2)\right)\\
&\xrightarrow{p} \int_0^\infty [v_c^{(1)}(u,\gamma_0,\eta_0,\sigma_0^2)]_{j_1 j_2} s_c^{(0)}(u,\gamma_0,\eta_0,\sigma_0^2)h_0(u)du\\
&\xrightarrow{p} \left(\int_0^\infty [v_c^{(1)}(u,\gamma_0,\eta_0,\sigma_0^2)] s_c^{(0)}(u,\gamma_0,\eta_0,\sigma_0^2)h_0(u)du\right)_{j_1 j_2}\\
&=B_{j_1 j_2}
\end{aligned}$$

and by the central limit theorem we have

$$n^{-\frac{1}{2}}U_c^{(n)}(\gamma_0,\eta_0,\sigma_0^2) = W^{(n)}(\infty,\gamma_0,\eta_0,\sigma_0^2) \xrightarrow{d} N(0,B).$$

To prove condition A3, we can write

$$\frac{1}{n}\frac{\partial}{\partial(\gamma,\eta^T)^T}U_c^{(n)}(\gamma,\eta,\sigma_0^2)|_{\gamma=\gamma^*,\eta=\eta^*} - A = -D_0 + D_1 + D_2 + D3$$

where

$$\begin{aligned}
D_0 =& \frac{1}{n}\sum_{i=1}^n \int_0^\infty \begin{bmatrix} \sigma_0^2\theta_i(u) & 0 & \cdots & 0\\ 0 & 0 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 0 \end{bmatrix} dN_i(u)\\
&- \int_0^\infty \begin{bmatrix} \sigma_0^2\mathbb{E}(\theta_i(u)) & 0 & \cdots & 0\\ 0 & 0 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 0 \end{bmatrix} s_c^{(0)}(u,\gamma_0,\eta_0,\sigma_0^2)h_0(u)du\\
D_1 =& \frac{1}{n}\sum_{i=1}^n \int_0^\infty V_c^{(1)}(u,\gamma^*,\eta^*,\sigma_0^2)dN_i(u) - \int_0^\infty v_c^{(1)}(u,\gamma_0,\eta_0,\sigma_0^2)s_c^{(0)}(u,\gamma_0,\eta_0,\sigma_0^2)h_0(u)du\\
D_2 =& \frac{1}{n}\sum_{i=1}^n \int_0^\infty V_c^{(2)}(u,\gamma^*,\eta^*,\sigma_0^2)dN_i(u) - \int_0^\infty v_c^{(2)}(u,\gamma_0,\eta_0,\sigma_0^2)s_c^{(0)}(u,\gamma_0,\eta_0,\sigma_0^2)h_0(u)du\\
D_3 =& \frac{1}{n}\sum_{i=1}^n \int_0^\infty V_c^{(3)}(u,\gamma^*,\eta^*,\sigma_0^2)dN_i(u),
\end{aligned}$$

and it remains to show that each of the terms D_0, D_1, D_2 and D_3 converge in probability to zero. For the terms D_0, D_1 and D_2, analogous results are presented by Andersen and Gill (1982) for survival data and the heuristic sketch for this was shown in our Theorem 3.5. For the term D_3, we make use of the relationship

$dN_i(u)dN_j(u) = 0$ for $i \neq j$ and $dN_i(u)^2 = 1$ since $dN_i(u)$ is an indicator function. The function $V^{(3)}$ is a $(p+1) \times (p+1)$ matrix where all entries are zero apart from the top left. Therefore we shall only consider entry $(1,1)$. We have that

$$\begin{aligned}
[D_3]_{11} =& \frac{1}{n}\sum_{i=1}^{n}\int_0^{\infty}\left[V^{(3)}(u,\gamma,\eta,\sigma^2)\right]_{11} dN_i(u)\\
=& \frac{1}{n}\sum_{i=1}^{n}\int_0^{\infty}\frac{\left[C^{(3)}(u,\gamma,\eta,\sigma^2)\right]_{11}}{\sum_{j=1}^{n}E_{0j}(u,\gamma,\eta,\sigma^2)Y_j(u)}dN_i(u)\\
=& \frac{1}{n}\sum_{i=1}^{n}\int_0^{\infty}\frac{\sum_{j=1}^{n}\sigma^2\theta_j(u)dN_j(u)E_{0j}(u,\gamma,\eta,\sigma^2)Y_j(u)}{\sum_{j=1}^{n}E_{0j}(u,\gamma,\eta,\sigma^2)Y_j(u)}dN_i(u)\\
=& \frac{1}{n}\sum_{i=1}^{n}\int_0^{\infty}\frac{\sigma^2\theta_i(u)E_{0i}(u,\gamma,\eta,\sigma^2)Y_i(u)}{\sum_{j=1}^{n}E_{0j}(u,\gamma,\eta,\sigma^2)Y_j(u)}dN_i(u). \qquad (4.16)
\end{aligned}$$

A single element in the summand in Expression (4.16) can be written

$$\int_0^{\infty}\frac{\frac{1}{n}\sigma^2\theta_i(u)E_{0i}(u,\gamma,\eta,\sigma^2)Y_i(u)dN_i(u)}{n\frac{1}{n}\sum_{j=1}^{n}E_{0j}(u,\gamma,\eta,\sigma^2)Y_j(u)}$$

and it is clear that $[D_3]_{11} \xrightarrow{p} 0$ as $n \to \infty$. Therefore, we have the result

$$\frac{1}{n}\frac{\partial}{\partial(\gamma,\eta^T)^T}U_c^{(n)}(\gamma,\eta,\sigma_0^2)|_{\gamma=\gamma^*,\eta=\eta^*} \xrightarrow{p} A.$$

□

In the above derivation of the distribution of the estimates $\hat{\gamma}_n$ and $\hat{\eta}_n$, we have assumed that σ_0^2, the variance of measurement error, is known. However, this is rarely the case and we proceed to find an estimator for σ^2. Suppose we are interested in the referral example where the longitudinal model takes the form $X_i(t) = b_{i0}+b_{i1}t$, then Tsiatis and Davidian (2001) suggest replacing σ^2 with the pooled estimator

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\mathbb{I}\{m_i > 2\}R_i}{\sum_{i=1}^{n}\mathbb{I}\{m_i > 2\}(m_i-2)}, \qquad (4.17)$$

where R_i is the residual sum of squares for the least squares fit to all m_i observations for patient i. The inclusion requirement for more than two longitudinal observations is due to design of the longitudinal model. Tsiatis and Davidian (2001) prove that $\hat{\sigma}^2$ is consistent for σ^2 and by arguments in Carroll et al. (2006) Section A.3.3, this estimator can replace σ_0^2 in the conditional score function.

In the joint model Equation (4.1)–(4.3), the treatment effect is the parameter

η and for this model, η is a scalar. Let γ_0, η_0 and σ_0^2 be the true values of the parameters γ, η and σ^2 respectively, then the hypothesis test in this case is

$$H_0 : \eta_0 \leq 0, \qquad H_A : \eta_0 > 0.$$

To make inferences about the individual parameter η, we consider the vector of parameters $(\gamma, \eta)^T$. Using the conditional score method, we can find estimates $\hat{\gamma}$ and $\hat{\eta}$ and using Theorem 4.3 we can determine the asymptotic joint distribution for these estimates. Further, Equation (4.17) is used to find an estimate $\hat{\sigma}^2$. The marginal asymptotic distribution for the parameter $\hat{\eta}$ is therefore

$$\sqrt{n}(\hat{\eta} - \eta_0) \xrightarrow{d} N(0, \Sigma_{22})$$

where

$$\Sigma = A^{-1}B(A^{-1})^T$$

and the matrices A and B are defined in Equations (4.14) and (4.15). The subscript on the covariance matrix Σ represents that η is the second parameter in the vector $(\gamma, \eta)^T$. The matrices A and B are estimated using

$$\hat{A} = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\infty}\left(\begin{bmatrix}\hat{\sigma}^2\theta_i(u) & 0 & \cdots & 0\\ 0 & 0 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 0\end{bmatrix} - V_c^{(1)}(u,\hat{\gamma},\hat{\eta},\hat{\sigma}^2) - V_c^{(2)}(u,\hat{\gamma},\hat{\eta},\hat{\sigma}^2)\right) dN_i(u) \tag{4.18}$$

$$\hat{B} = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\infty}\left[V_c^{(1)}(u,\hat{\gamma},\hat{\eta},\hat{\sigma}^2)\right] dN_i(u). \tag{4.19}$$

The information matrix for $\hat{\eta}$ in the fixed sample trial is therefore

$$\mathcal{I} = \frac{1}{n}\left[\hat{A}^{-1}\hat{B}(\hat{A}^{-1})^T\right]_{22}^{-1}$$

and a standardised statistic is given by

$$Z = \hat{\eta}\sqrt{\mathcal{I}}.$$

4.2.2 | Group sequential results

It is now desirable to extend the fixed sample theory for the joint model to group sequential trials. To perform a group sequential trial we need to know the joint distribution of the sequence of treatment effect estimates that will be obtained at each analysis. To determine this distribution, we shall define the conditional score and information matrix for each analysis which are calculated using data obtained at that analysis. The theoretical work in this section and the proof of asymptotic normality in the upcoming Theorem 4.4 builds upon the fixed sample joint modelling results of Section 4.2.1 and the group sequential survival results of Section 3.3.2. All the work presented in this section is original work, including Theorem 4.4 and its proof.

For the conditional score at each analysis, we shall define group sequential versions of all objects included in the fixed sample conditional score. Similarly to the group sequential version of the score for partial likelihood, the censoring mechanism is used to keep patients in the at-risk set who have yet to experience an event. For patient i with time-to-failure random variable F_i, let $C_i(k)$ be the time-to-censoring random variable at analysis k. This censoring event includes "end of study" censoring for the total follow-up time of patient i at analysis k, then at analysis k the event time random variable is $T_i(k) = \min\{F_i, C_i(k)\}$. The observed event time is $t_i(k)$ and the observed censoring indicator is $\delta_i(k) = \mathbb{I}\{F_i \leq C_i(k)\}$.

In the conditional score approach, to be included in the at-risk set at time t the patient must have at least two longitudinal observations to fit the longitudinal regression model. The at-risk process at analysis k is an indicator for not yet observing the event, not yet censored, or having enough longitudinal observations. Therefore for patient i at analysis k the at-risk process and counting process for the joint model are

$$
\begin{aligned}
Y_i(k,t) &= \mathbb{I}\{t_i(k) \geq t, t_{i2} \leq t\} \\
N_i(k,t) &= \mathbb{I}\{t_i(k) \leq t, \delta_i(k) = 1, t_{i2} \leq t\}.
\end{aligned}
$$

The corresponding conditional intensity process and compensated counting process are defined by

$$\begin{aligned} \lambda_i^C(k,t) &= lim_{dt\downarrow 0}\frac{\mathbb{P}(dN_i(k,t)=1|S_i(t,\gamma,\sigma^2),t_i(t),Z_i,Y_i(k,t))}{dt} \qquad (4.20)\\ &= h_0(t)\exp\{\gamma S_i(t,\gamma,\sigma^2)-\gamma^2\sigma^2\theta_i(t)/2+\eta^T Z_i\}Y_i(k,t)\\ &= h_0(t)E_{0i}(t,\gamma,\eta,\sigma^2)Y_i(k,t)\\ M_i(k,t) &= N_i(k,t)-\int_0^t h_0(u)E_{0i}(u,\gamma,\eta,\sigma^2)Y_i(k,u)du. \end{aligned}$$

The following functions are needed to define the group sequential conditional score at analysis k. These functions are

$$\begin{aligned} S_c^{(0)}(k,t,\gamma,\eta,\sigma^2) &= \frac{1}{n}\sum_{i=1}^n Y_i(k,t)E_{0i}(t,\gamma,\eta,\sigma^2)\\ S_c^{(1)}(k,t,\gamma,\eta,\sigma^2) &= \frac{1}{n}\sum_{i=1}^n \left\{\begin{array}{c} S_i(t,\gamma,\sigma^2)\\ Z_i\end{array}\right\} Y_i(k,t)E_{0i}(t,\gamma,\eta,\sigma^2)\\ S_c^{(2)}(k,t,\gamma,\eta,\sigma^2) &= \frac{1}{n}\sum_{i=1}^n \left\{\begin{array}{c} S_i(t,\gamma,\sigma^2)\\ Z_i\end{array}\right\}\left\{\begin{array}{c} S_i(t,\gamma,\sigma^2)\\ Z_i\end{array}\right\}^T Y_i(k,t)E_{0i}(t,\gamma,\eta,\sigma^2)\\ E_c(k,t,\gamma,\eta,\sigma^2) &= \frac{S_c^{(1)}(k,t,\gamma,\eta,\sigma^2)}{S_c^{(0)}(k,t,\gamma,\eta,\sigma^2)}. \end{aligned}$$

The function $E_c(k,t,\gamma,\eta,\sigma^2)$ has the interpretation of the expectation of the vector $\{S_i(t,\gamma,\sigma^2),Z_i^T\}^T$ at analysis k weighted by the conditional intensity process. Let τ_k be the maximum follow-up time at analysis k, then the conditional score for analysis k is

$$\begin{aligned} U_c(k,\gamma,\eta,\sigma^2) &= \int_0^{\tau_k}\sum_{i=1}^n \left(\{S_i(u,\gamma,\sigma^2),Z_i^T\}^T-E_c(k,u,\gamma,\eta,\sigma^2)\right)dN_i(k,u) \qquad (4.21)\\ &= \int_0^{\tau_k}\sum_{i=1}^n \left(\{S_i(u,\gamma,\sigma^2),Z_i^T\}^T-E_c(k,u,\gamma,\eta,\sigma^2)\right)dM_i(k,u). \end{aligned}$$

This equality follows by the same reasoning as Lemma 3.4, which states that the compensated counting process can replace the counting process.

We now follow a similar structure to the partial likelihood function for survival data in Section 3.3.2 and we create a new counting process that allows the conditional score statistic to be written as the sum of distinct increments. This counting process

is

$$
\begin{aligned}
&DN_i(0,t) = 0\\
&DN_i(k,t) = N_i(k,t) - N_i(k-1,t) \quad \text{for } k = 1, \dots, K.
\end{aligned}
$$

The corresponding compensated counting process is therefore given by

$$
\begin{aligned}
DM_i(0,t) =&0\\
DM_i(k,t) =&DN_i(k,t) - \int_0^t h_0(u)E_{0i}(t,\gamma,\eta,\sigma^2)(Y_i(k,u) - Y_i(k-1,u))du\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for } k = 1, \dots, K.
\end{aligned}
$$

The event for patient i can only occur in one interval, and therefore we have $N_i(k,t) = \sum_{l=1}^{k} DN_i(l,t)$ and the conditional score statistic at analysis k is

$$
\begin{aligned}
U_c(k,\gamma,\eta,\sigma^2) &= \int_0^{\tau_k} \sum_{i=1}^{n}\sum_{l=1}^{k} \left(\{S_i(u,\gamma,\sigma^2), Z_i^T\}^T - E_c(l,u,\gamma,\eta,\sigma^2)\right) dDN_i(k,u)\\
&= \int_0^{\tau_k} \sum_{i=1}^{n}\sum_{l=1}^{k} \left(\{S_i(u,\gamma,\sigma^2), Z_i^T\}^T - E_c(l,u,\gamma,\eta,\sigma^2)\right) dDM_i(k,u).
\end{aligned}
$$

By the same argument as for the fixed sample case, we shall write the conditional score at analysis k, $U_c(k,\gamma,\eta,\sigma^2)$, as

$$
\int_0^{\tau_k} \sum_{i=1}^{n}\sum_{l=1}^{k} \left(\{S_i(u,\gamma,\sigma^2), Z_i^T\}^T - e_c(l,u,\gamma,\eta,\sigma^2)\right) dDM_i(l,u) \tag{4.22}
$$

$$
+\int_0^{\tau_k} \sum_{i=1}^{n}\sum_{l=1}^{k} \left(e_c(l,u,\gamma,\eta,\sigma^2) - E_c(l,u,\gamma,\eta,\sigma^2)\right) dDM_i(l,u) \tag{4.23}
$$

where $e_c(l,u,\gamma,\eta,\sigma^2)$ denotes the probabilistic limit of $E_c(l,u,\gamma,\eta,\sigma^2)$. In Conditions 4.2, we shall assume that this limit exists and that $E_c(l,u,\gamma,\eta,\sigma^2)$ converges pointwise to $e_c(l,u,\gamma,\eta,\sigma^2)$. Further, by the same argument as the fixed sample case, we see that n^{-1} times Expression (4.23) converges in probability to zero in a neighbourhood of (γ_0,η_0) and deduce that the behaviour of the estimators of γ and η, which are solutions to the equation $U_c(k,\gamma,\eta,\sigma^2) = 0$, will be dictated by Expression (4.22). Therefore, we now restrict our attention to Expression (4.22).

By a similar argument to Lemma 4.2 we have the expectation of Expression (4.22)

is

$$\mathbb{E}\left(\int_0^{\tau_k}\sum_{i=1}^n\sum_{l=1}^k\left(\{S_i(u,\gamma,\sigma^2),Z_i^T\}^T-e_c(l,u,\gamma,\eta,\sigma^2)\right)dDM_i(l,u)\right)=0 \quad (4.24)$$

for each $k=1,\dots,K$. To see this, we substitute $e_c(l,u,\gamma,\eta,\sigma^2)$ for $e_c(u,\gamma,\eta,\sigma^2)$ in Lemma 4.2 and the proof follows. Therefore combined with the fact that Expression (4.23) converges in probability to zero, we have that setting the group sequential conditional score function equal to zero defines a set of estimating equations. For each $k=1,\dots,K$, let $\hat{\gamma}_n^{(k)}$ and $\hat{\eta}_n^{(k)}$ be the values of γ and η respectively such that $U_c(k,\gamma,\eta,\sigma^2)=0$, then the estimates $\hat{\gamma}_n^{(k)}$ and $\hat{\eta}_n^{(k)}$ will be asymptotically multivariate normal.

We derived the first derivative of the fixed sample conditional score with respect to the vector $(\gamma,\eta)^T$. This derivative function was seen to play a similar role to the Fisher information matrix in a general statistical model using maximum likelihood theory. Previously, we defined the $(p+1)$-dimensional vector

$$J_i(t,\gamma,\sigma^2)=\{S_i(t,\gamma,\sigma^2)-\hat{X}_i(t)-\gamma\sigma^2\theta_i(t),0,\dots,0\}^T$$

. Then, functions that are needed for the variance matrix of the group sequential conditional score are

$$\begin{aligned}
C^{(1)}(k,t,\gamma,\eta,\sigma^2)&=\frac{1}{n}\sum_{i=1}^n J_i(t,\gamma,\sigma^2)Y_i(k,t)E_{0i}(t,\gamma,\eta,\sigma^2)\\
C^{(2)}(k,t,\gamma,\eta,\sigma^2)&=\frac{1}{n}\sum_{i=1}^n J_i(t,\gamma,\sigma^2)\left\{\begin{array}{c}S_i(t,\gamma,\sigma^2)\\ Z_i\end{array}\right\}^T Y_i(k,t)E_{0i}(t,\gamma,\eta,\sigma^2)\\
C^{(3)}(k,t,\gamma,\eta,\sigma^2)&=\frac{1}{n}\sum_{i=1}^n\begin{bmatrix}\sigma^2\theta_i(u)&0&\cdots&0\\0&0&\cdots&0\\\vdots&\vdots&\ddots&\vdots\\0&0&\cdots&0\end{bmatrix}Y_i(k,t)E_{0i}(t,\gamma,\eta,\sigma^2)\\
V_c^{(1)}(k,t,\gamma,\eta,\sigma^2)&=\frac{S_c^{(2)}(k,t,\gamma,\eta,\sigma^2)}{S_c^{(0)}(k,t,\gamma,\eta,\sigma^2)}-\frac{S_c^{(1)}(k,t,\gamma,\eta,\sigma^2)S_c^{(1)}(k,t,\gamma,\eta,\sigma^2)^T}{[S_c^{(0)}(k,t,\gamma,\eta,\sigma^2)]^2}\\
V_c^{(2)}(k,t,\gamma,\eta,\sigma^2)&=\frac{C^{(2)}(k,t,\gamma,\eta,\sigma^2)}{S_c^{(0)}(k,t,\gamma,\eta,\sigma^2)}-\frac{S_c^{(1)}(k,t,\gamma,\eta,\sigma^2)C^{(1)}(k,t,\gamma,\eta,\sigma^2)^T}{[S_c^{(0)}(k,t,\gamma,\eta,\sigma^2)]^2}\\
V_c^{(3)}(k,t,\gamma,\eta,\sigma^2)&=\frac{C^{(3)}(k,t,\gamma,\eta,\sigma^2)}{S_c^{(0)}(k,t,\gamma,\eta,\sigma^2)}.
\end{aligned}$$

The first derivative of the group sequential conditional score function with respect to $(\gamma, \eta)^T$ at analysis k is

$$
\begin{aligned}
&\frac{\partial}{\partial(\gamma,\eta)^T} U_c(k,\gamma,\eta,\sigma^2) \\
=&\sum_{i=1}^{n}\int_0^\infty \begin{bmatrix} \sigma^2\theta_i(u) & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} dN_i(k,u) \\
-&\sum_{i=1}^{n}\int_0^\infty \left[V_c^{(1)}(k,u,\gamma,\eta,\sigma^2) + V_c^{(2)}(k,u,\gamma,\eta,\sigma^2) + V_c^{(3)}(k,u,\gamma,\eta,\sigma^2)\right] dN_i(k,u).
\end{aligned}
\tag{4.25}
$$

To ensure the existence of the asymptotic covariance matrix, we require the probabilistic limits of $S_c^{(j)}(k)$ and $C^{(j)}(k)$ to exist. The limits are defined through the following conditions.

Conditions 4.2.

1. *There exist neighbourhoods Γ of γ_0 and N of η_0 and for each $k = 1, \ldots, K$ there are functions $s_c^{(0)}(k,t,\gamma,\eta,\sigma^2)$, $s_c^{(1)}(k,t,\gamma,\eta,\sigma^2)$, $s_c^{(2)}(k,t,\gamma,\eta,\sigma^2)$, $c^{(1)}(k,t,\gamma,\eta,\sigma^2)$ and $c^{(2)}(k,t,\gamma,\eta,\sigma^2)$ defined on $[0,\infty) \times \Gamma \times N$ such that*

$$\sup_{t\in[0,\infty),\gamma\in\Gamma,\eta\in N} \left\| S_c^{(j)}(k,t,\gamma,\eta,\sigma^2) - s_c^{(j)}(k,t,\gamma,\eta,\sigma^2) \right\| \xrightarrow{p} 0 \text{ for } j = 0,1,2$$

$$\sup_{t\in[0,\infty),\gamma\in\Gamma,\eta\in N} \left\| C^{(j)}(k,t,\gamma,\eta,\sigma^2) - c^{(j)}(k,t,\gamma,\eta,\sigma^2) \right\| \xrightarrow{p} 0 \text{ for } j = 1,2.$$

Each $s_c^{(j)}(k,t,\gamma,\eta,\sigma^2)$ and $c^{(j)}(k,t,\gamma,\eta,\sigma^2)$ is a continuous function of $\gamma \in \Gamma$ and $\eta \in N$ uniformly in $t \in [0,\infty)$, and bounded on $[0,\infty) \times \Gamma \times N$. For each $k = 1, \ldots, K$ $s_c^{(0)}$ and $c^{(0)}$ are bounded away from zero on $[0,\infty) \times \Gamma \times N$.

It is clear that the probabilistic limits $e_c(k,t,\gamma,\eta,\sigma^2)$ of $E_c(k,t,\gamma,\eta,\sigma^2)$, $v_c^{(1)}(k,t,\gamma,\eta,\sigma^2)$ of $V_c^{(1)}(k,t,\gamma,\eta,\sigma^2)$ and $v_c^{(2)}(k,t,\gamma,\eta,\sigma^2)$ of $V_c^{(2)}(k,t,\gamma,\eta,\sigma^2)$ exist and can expressed in terms of $s_c^{(j)}(k,t,\gamma,\eta,\sigma^2)$ and $c^{(j)}(k,t,\gamma,\eta,\sigma^2)$ for $j = 0,1,2$

and these are

$$\begin{aligned}
e_c(k,t,\gamma,\eta,\sigma^2) &= \frac{s_c^{(1)}(k,t,\gamma,\eta,\sigma^2)}{s_c^{(0)}(k,t,\gamma,\eta,\sigma^2)}\\
v_c^{(1)}(k,t,\gamma,\eta,\sigma_0^2) &= \frac{s_c^{(2)}(k,t,\gamma,\eta,\sigma_0^2)}{s_c^{(0)}(k,t,\gamma,\eta,\sigma_0^2)} - \frac{s_c^{(1)}(k,t,\gamma,\eta,\sigma^2)s_c^{(1)}(k,t,\gamma,\eta,\sigma^2)^T}{[s_c^{(0)}(k,t,\gamma,\eta,\sigma^2)]^2}\\
v_c^{(2)}(k,t,\gamma,\eta,\sigma^2) &= \frac{c^{(2)}(k,t,\gamma,\eta,\sigma^2)}{s_c^{(0)}(k,t,\gamma,\eta,\sigma^2)} - \frac{s_c^{(1)}(k,t,\gamma,\eta,\sigma^2)c^{(1)}(k,t,\gamma,\eta,\sigma^2)^T}{[s_c^{(0)}(k,t,\gamma,\eta,\sigma^2)]^2}.
\end{aligned}$$

We shall now prove that the estimates $\hat{\gamma}_n^{(1)}, \hat{\eta}_n^{(1)}, \ldots, \hat{\gamma}_n^{(K)}, \hat{\eta}_n^{(K)}$ are asymptotically multivariate normally distributed and we shall derive an explicit form for the covariance matrix of this vector of parameters. To do so, we shall prove that 1–4 of Conditions 3.2 hold and hence apply Theorem 3.2. The remaining conditions and the additional Conditions 4.2 are assumed to hold to avoid technical distractions.

Theorem 4.4. *Suppose that γ_0, η_0 and σ_0^2 are the true values of the parameters γ, η and σ^2 respectively. For each $k = 1, \ldots, K$, let $\hat{\gamma}_n^{(k)}$ and $\hat{\eta}_n^{(k)}$ be the values of γ and η which are the solution to the equation $(U_c^{(n)}(k,\gamma,\eta,\sigma_0) = 0$ and suppose that Conditions 4.2 hold. Then the vector $(\hat{\gamma}_n^{(1)}, \hat{\eta}_n^{(1)}, \ldots, \hat{\gamma}_n^{(K)}, \hat{\eta}_n^{(K)})^T$ converges in distribution to a Gaussian random variable, specifically*

$$n^{\frac{1}{2}} \begin{pmatrix} \hat{\gamma}_n^{(1)} - \gamma_0 \\ \hat{\eta}_n^{(1)} - \eta_0 \\ \hat{\gamma}_n^{(2)} - \gamma_0 \\ \hat{\eta}_n^{(2)} - \eta_0 \\ \vdots \\ \hat{\gamma}_n^{(K)} - \gamma_0 \\ \hat{\eta}_n^{(K)} - \eta_0 \end{pmatrix} \xrightarrow{d} N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1K} \\ \Sigma_{12} & \Sigma_{22} & \cdots & \Sigma_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{1K} & \Sigma_{2K} & \cdots & \Sigma_{KK} \end{bmatrix} \right)$$

where

$$\Sigma_{k_1 k_2} = (A^{(k1)})^{-1} B^{(k1)} ((A^{(k2)})^{-1})^T$$

and the matrices $A^{(k)}$ and $B^{(k)}$ are defined by

$$A^{(k)} = \int_0^\infty \begin{bmatrix} \sigma_0^2 \mathbb{E}(\theta_i(u)) & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} s_c^{(0)}(k, u, \gamma_0, \eta_0, \sigma_0^2) h_0(u) du \tag{4.26}$$

$$- \int_0^\infty \left(v_c^{(1)}(k, u, \gamma_0, \eta_0, \sigma_0^2) + v_c^{(2)}(k, u, \gamma_0, \eta_0, \sigma_0^2) \right) s_c^{(0)}(k, u, \gamma_0, \eta_0, \sigma_0^2) h_0(u) du \tag{4.27}$$

$$B^{(k)} = \int_0^\infty v_c^{(1)}(k, u, \gamma_0, \eta_0, \sigma_0^2) s_c^{(0)}(k, u, \gamma_0, \eta_0, \sigma_0^2) h_0(u) du. \tag{4.28}$$

Proof. In Theorem 3.2, we proved that the sequence of estimates which are the solutions to estimating equations, are asymptotically normally distributed. In applying Theorem 3.2, the group sequential conditional score function $U_c(k, \gamma, \eta, \sigma^2)$ plays the role of the estimating function $\mathbf{G}_n(\boldsymbol{\theta}, \mathbf{x}_n^{(k)})$. Therefore, we show that the following conditions are satisfied:

B1 For each $k = 1, \ldots, K$, $\hat{\gamma}_n^{(k)} \xrightarrow{p} \gamma_0$ and $\hat{\eta}_n^{(k)} \xrightarrow{p} \eta_0$.

B2 For each $k = 1, \ldots, K$, $n^{-\frac{1}{2}} U_c^{(n)}(k, \gamma_0, \eta_0, \sigma_0^2) \xrightarrow{d} N(0, B^{(k)})$.

B3 For each $k = 1, \ldots, K$, and for all γ_n^*, η_n^* such that $\gamma_n^* \xrightarrow{p} \gamma_0, \eta_n^* \xrightarrow{p} \eta_0$,

$$n^{-1} \frac{\partial}{\partial(\gamma, \eta^T)^T} U_c^{(n)}(k, \gamma, \eta, \sigma_0^2) \bigg|_{\gamma=\gamma_n^*, \eta=\eta_n^*} = \xrightarrow{p} A^{(k)}$$

.

B4 For $1 \leq k_1 \leq k_2 \leq K$, we require

$$n^{-\frac{1}{2}} Cov(U_c^{(n)}(k_1, \gamma_0, \eta_0, \sigma_0^2), U_c^{(n)}(k_2, \gamma_0, \eta_0, \sigma_0^2)) \xrightarrow{p} B^{(k_1)}.$$

The proof that conditions B1–B3 hold, follow directly from the fixed sample case and we shall focus on the proof that condition B4 holds. For the remainder of this proof, we shall use the argument that the asymptotic behaviour of the estimates $\hat{\gamma}_n^{(k)}$ and $\hat{\eta}_n^{(k)}$ is dictated by Expression (4.22) and we shall therefore focus on the following form for the group sequential conditional score

$$U_c(k, \gamma, \eta, \sigma^2) = \int_0^{\tau_k} \sum_{i=1}^n \sum_{l=1}^k \left(\{S_i(u, \gamma, \sigma^2), Z_i^T\}^T - e_c(l, u, \gamma, \eta, \sigma^2) \right) dDM_i(l, u).$$

For consistency, we shall use the martingale notation. For each $j = 1, \ldots, p+1$ where p is the length of the vector Z_i, let

$$\begin{aligned} W_j^{(n)}(k, t, \gamma, \eta, \sigma^2) &= \int_0^t \sum_{i=1}^n \sum_{l=1}^k H_{ij}^{(n)}(l, u, \gamma, \eta, \sigma^2) dDN_i(l, u) \\ &= \int_0^t \sum_{i=1}^n \sum_{l=1}^k H_{ij}^{(n)}(l, u, \gamma, \eta, \sigma^2) dDM_i(l, u) \end{aligned}$$

where

$$H_{ij}^{(n)}(l, u, \gamma, \eta, \sigma^2) = n^{-\frac{1}{2}}(\{S_i(u, \gamma, \sigma^2), Z_i\}_j^T - e_c(l, u, \gamma, \eta, \sigma^2)_j).$$

Thus $n^{-\frac{1}{2}} U_c^{(n)}(k, \gamma, \eta, \sigma^2) = W^{(n)}(k, \tau_k, \gamma, \eta, \sigma^2)$. We have previously shown in Equation (4.24) that $U_c(k, \gamma, \eta, \sigma^2)$ has expectation zero for all $k = 1, \ldots, K$. Hence we deduce that for for all $j = 1, \ldots, p$,

$$\mathbb{E}\left(W_j^{(n)}(k, \tau_k, \gamma, \eta, \sigma^2)\right) = 0.$$

In the following, we shall drop the dependency of the function $H^{(n)}$ on the parameters γ, η and σ^2. This is for notational simplicity. The covariance is therefore given by

$$\begin{aligned} &Cov\Big(W_{j_1}^{(n)}(k_1, \tau_{k_1}, \gamma_0, \eta_0, \sigma_0^2), W_{j_2}^{(n)}(k_2, \tau_{k_2}, \gamma_0, \eta_0, \sigma_0^2)\Big) \\ &= \mathbb{E}\Big(\int_0^{\tau_{k_1}} \sum_{i=1}^n \sum_{l_1=1}^{k_1} H_{ij_1}^{(n)}(l_1, u) dDN_i(l_1, u) \int_0^{\tau_{k_2}} \sum_{i=1}^n \sum_{l_2=1}^{k_2} H_{ij_2}^{(n)}(l_2, u) dDN_i(l_2, u)\Big) \\ &= \mathbb{E}\Big(\sum_{i=1}^n \Big[\sum_{l_1=1}^{k_1} \int_0^{\tau_{k_1}} H_{ij_1}^{(n)}(l_1, u) dDN_i(l_1, u) \sum_{l_2=1}^{k_2} \int_0^{\tau_{k_2}} H_{ij_2}^{(n)}(l_2, u) dDN_i(l_2, u)\Big]\Big). \end{aligned}$$

This second equality holds because patients are independent and the orders of summation and integration can be interchanged.

Further, the event for each patient can only happen in one analysis so that if $l_1 \neq l_2$ then $dDN_i(l_1, u) dDN_i(l_2, u) = 0$ and because $dDN_i(l, u)$ is an indicator function, we have that $dDNi(l,u)^2 = dDN_i(l,u)$. Thus for $k_1 \leq k_2$,

$$
\begin{aligned}
&\mathbb{E}\bigg(\sum_{i=1}^{n}\bigg[\sum_{l_1=1}^{k_1}\int_0^{\tau_{k_1}} H_{ij_1}^{(n)}(l_1,u)dDN_i(l_1,u)\sum_{l_2=1}^{k_2}\int_0^{\tau_{k_2}} H_{ij_2}^{(n)}(l_2,u)dDN_i(l_2,u)\bigg]\bigg)\\
&=\mathbb{E}\bigg(\sum_{i=1}^{n}\sum_{l=1}^{k_1}\int_0^{\tau_1} H_{ij_1}^{(n)}(l,u)H_{ij_2}^{(n)}(l,u)dDN_i(l,u)\bigg).
\end{aligned}
$$

For the following, note that we have

$$
\mathbb{E}(dN_i(k,u)|S_i(t,\gamma,\sigma^2),t_i(t),Z_i,Y_i(k,u) = h_0(u)E_{0i}(u,\gamma,\eta,\sigma^2)Y_i(k,u)du
$$

which follows by definition of the conditional intensity process, $\lambda_i^C(k,t)$, in Equation (4.20). The following calculations use the law of total expectation and the expectation of the counting process $dN_i(k,u)$ in a similar way to the fixed sample case.

$$
\begin{aligned}
&\mathbb{E}\bigg(\sum_{i=1}^{n}\sum_{l=1}^{k_1}\int_0^{\tau_1} H_{ij_1}^{(n)}(l,u)H_{ij_2}^{(n)}(l,u)dDN_i(l,u)\bigg)\\
&=\mathbb{E}\bigg(\sum_{i=1}^{n}\sum_{l=1}^{k_1}\int_0^{\tau_1} \mathbb{E}[H_{ij_1}^{(n)}(l,u)H_{ij_2}^{(n)}(l,u)dDN_i(l,u)|S_i(u,\gamma_0,\sigma_0^2),Z_i,t_i(u),Y_i(l,u)]\bigg)\\
&=\mathbb{E}\bigg(\sum_{i=1}^{n}\sum_{l=1}^{k_1}\int_0^{\tau_1} H_{ij_1}^{(n)}(l,u)H_{ij_2}^{(n)}(l,u)\mathbb{E}[dN_i(l,u)-dN_i(l-1,u)|S_i(u,\gamma_0,\sigma_0^2),Z_i,t_i(u),Y_i(l,u)]\bigg)\\
&=\mathbb{E}\bigg(\sum_{i=1}^{n}\sum_{l=1}^{k_1}\int_0^{\tau_1} H_{ij_1}^{(n)}(l,u)H_{ij_2}^{(n)}(l,u)h_0(u)E_{0i}(u,\gamma_0,\eta_0,\sigma_0^2)(Y_i(l,u)-Y_i(l-1,u))du\bigg)\\
&=\mathbb{E}\bigg(\sum_{i=1}^{n}\int_0^{\tau_{k_1}} H_{ij_1}^{(n)}(k_1,u)H_{ij_2}^{(n)}(k_1,u)h_0(u)E_{0i}(u,\gamma_0,\eta_0,\sigma_0^2)Y_i(k_1,u)du\bigg).
\end{aligned}
$$

For the following calculation, dependency on parameters γ_0, η_0 and σ_0^2 is removed from the functions $S_c^{(0)}(k,t), S_c^{(1)}(k,t)$ and $S_c^{(2)}(k,t)$ for notational purposes. Further, the dependency on parameters γ_0, η_0 and σ_0^2 is also removed from the probabilistic limits $s_c^{(0)}(k,t), s_c^{(1)}(k,t), s_c^{(2)}(k,t)$ and $e_c(k,t)$. Then following calculation holds

$$\begin{aligned}
&\sum_{i=1}^{n} H^{(n)}_{ij_1}(k,u)H^{(n)}_{ij_2}(k,u)h_0(u)E_{0i}(u,\gamma_0,\eta_0,\sigma_0^2)Y_i(k,u)du\\
&=\sum_{i=1}^{n} n^{-1}\big(\{S_i(u,\gamma_0,\sigma_0^2),Z_i^T\}^T_{j_1}-e_c(k,u)_{j_1}\big)\\
&\qquad\qquad \times\big(\{S_i(u,\gamma_0,\sigma_0^2),Z_i^T\}^T_{j_2}-e_c(k,u)_{j_2}\big)h_0(u)E_{0i}(u,\gamma_0,\eta_0,\sigma_0^2)Y_i(k,u)\\
&=\Bigg([S_c^{(2)}(k,u)]_{j_1j_2}-\frac{[s_c^{(1)}(k,u)]_{j_1}}{s_c^{(0)}(u)}[S_c^{(1)}(k,u)]_{j_2}\\
&\qquad -\frac{[s_c^{(1)}(k,u)]_{j_2}}{s_c^{(0)}(k,u)}[S_c^{(1)}(k,u)]_{j_1}+\frac{[s_c^{(1)}(k,u)]_{j_1}}{s_c^{(0)}(k,u)}\frac{[s_c^{(1)}(k,u)]_{j_2}}{s_c^{(0)}(k,u)}S_c^{(0)}(k,u)\Bigg)h_0(u)\\
&\xrightarrow{p}\Bigg(\frac{[s_c^{(2)}(k,u)]_{j_1j_2}}{s_c^{(0)}(k,u)}-\frac{[s_c^{(1)}(k,u)]_{j_1}[s_c^{(1)}(k,u)]_{j_2}}{s_c^{(0)\otimes 2}(k,u)}\Bigg)s_c^{(0)}(k,u)h_0(u)\\
&=[v_c^{(1)}(k,u)]_{j_1j_2}s_c^{(0)}(k,u)h_0(u).
\end{aligned}$$

Therefore, combining the above, it can be seen that for $1 \leq k_1 \leq k_2 \leq K$,

$$\begin{aligned}
&Cov\left(W^{(n)}_{j_1}(k_1,\tau_{k_1},\gamma_0,\eta_0,\sigma_0^2),W^{(n)}_{j_2}(k_2,\tau_{k_2},\gamma_0,\eta_0,\sigma_0^2)\right)\\
&\xrightarrow{p}\int_0^\infty [v_c^{(1)}(k_1,u,\gamma_0,\eta_0,\sigma_0^2)]_{j_1j_2}s_c^{(0)}(k_1,u,\gamma_0,\eta_0,\sigma_0^2)h_0(u)du\\
&\xrightarrow{p}\left(\int_0^\infty v_c^{(1)}(k_1,u,\gamma_0,\eta_0,\sigma_0^2)s_c^{(0)}(k_1,u,\gamma_0,\eta_0,\sigma_0^2)h_0(u)du\right)_{j_1j_2}\\
&=B^{(k_1)}_{j_1j_2}
\end{aligned}$$

We have the result, for $1 \leq k_1 \leq k_2 \leq K$

$$n^{-\frac{1}{2}}Cov(U_c^{(n)}(k_1,\gamma_0,\eta_0,\sigma_0^2),U_c^{(n)}(k_2,\gamma_0,\eta_0,\sigma_0^2)) \xrightarrow{p} B^{(k_1)}.$$

□

Similarly to the fixed sample case, we have assumed that σ_0^2 is known in the derivation of the distribution of $\hat{\gamma}_n^{(k)}$ and $\hat{\eta}_n^{(k)}$. This is not generally the case but by arguments in Carroll et al. (2006) Section A.3.3, we can find a consistent estimate to replace σ_0^2 with in the group sequential conditional score function. At analysis k this estimate is given by

$$\hat{\sigma}^{(k)2}=\frac{\sum_{i=1}^n \mathbb{I}\{m_i(k)>2\}R_i(k)}{\sum_{i=1}^n \mathbb{I}\{m_i(k)>2\}(m_i(k)-2)}, \tag{4.29}$$

where $R_i(k)$ is the residual sum of squares for the least squares fit to all $m_i(k)$ observations for patient i available at analysis k.

We shall use Theorem 4.4 to create a group sequential trial based on the joint model. Let γ_0, η_0 and σ_0^2 be the true values of the parameters γ, η and σ^2 respectively in the joint model Equation (4.1)–(4.3). We shall test the hypothesis

$$H_0 : \eta_0 \leq 0, \qquad H_A : \eta_0 > 0.$$

Using the group sequential conditional score method, let $\hat{\gamma}^{(k)}, \hat{\eta}^{(k)}$ be the values of the parameters γ and η such that $U_c(k, \gamma, \eta, \sigma^2) = 0$ where the conditional score function is calculated using Equation (4.21). Further, let $\hat{\sigma}^{(k)2}$ be the estimate for σ_0^2 given in Equation (4.29). By Theorem 4.4, for each $k = 1, \ldots, K$, the marginal distribution of the parameter $\hat{\eta}^{(k)}$ is

$$\sqrt{n}(\hat{\eta}^{(k)} - \eta_0) \xrightarrow{d} N(0, \Sigma_{22}^{(k)})$$

where

$$\Sigma^{(k)} = (A^{(k)})^{-1} B^{(k)} ((A^{(k)})^{-1})^T$$

and the matrices $A^{(k)}$ and $B^{(k)}$ are defined by Equations (4.26)–(4.28). Note that the subscript notation in the covariance matrix represents that the parameter η is the second parameter in the vector (γ, η). The matrices $A^{(k)}$ and $B^{(k)}$ are estimated using

$$\begin{aligned} \hat{A}^{(k)} = & \frac{1}{n} \sum_{i=1}^{n} \int_0^{\infty} \begin{bmatrix} \hat{\sigma}^{(k)2}\theta_i(u) & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} dN_i(k, u) \\ & - \frac{1}{n} \sum_{i=1}^{n} \int_0^{\infty} \left[V_c^{(1)}(k, u, \hat{\gamma}^{(k)}, \hat{\eta}^{(k)}, \hat{\sigma}^{(k)2}) + V_c^{(2)}(k, u, \hat{\gamma}^{(k)}, \hat{\eta}^{(k)}, \hat{\sigma}^{(k)2}) \right] dN_i(k, u) \end{aligned} \tag{4.30}$$

$$\hat{B}^{(k)} = \frac{1}{n} \sum_{i=1}^{n} \int_0^{\infty} \left[V_c^{(1)}(k, u, \hat{\gamma}^{(k)}, \hat{\eta}^{(k)}, \hat{\sigma}^{(k)2}) \right] dN_i(k, u). \tag{4.31}$$

The information matrix at analysis k of the group sequential trial is given by

$$\mathcal{I}_k = \frac{1}{n} \left[(\hat{A}^{(k)})^{-1} \hat{B}^{(k)} ((\hat{A}^{(k)})^{-1})^T \right]_{22}^{-1}.$$

Further, a standardised test statistic at analysis k is given by

$$Z_k = \hat{\eta}^{(k)} \sqrt{\mathcal{I}_k}.$$

In Section 4.4.1 we shall investigate the covariance structure for the joint distribution of the estimates $\hat{\eta}^{(1)}, \ldots, \hat{\eta}^{(K)}$.

4.3 | Simulation study of the parameter estimates

4.3.1 | Parameter values for simulation studies

We have so far provided a distribution for the treatment effect estimate in the joint model for both a fixed sample clinical trial and a group sequential trial. These results are proved theoretically and in the asymptotic setting. In this section, we shall perform some simulation studies to confirm these distributional results. Further, we assess the impact of having a small sample size when following an asymptotic assumption.

For convenience, the joint model, Equations (4.1)-(4.3) of Section 4.1 is presented again below. Longitudinal observations, $W_i(t)$ for patients $i = 1, \ldots, n$ follow the random effects model

$$W_i(t) = b_{i0} + b_{i1}t + \epsilon_i(t)$$

where

$$\begin{bmatrix} b_{i0} \\ b_{i1} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_0 \\ \mu_1 \end{bmatrix}, \begin{bmatrix} \phi_0^2 & 0 \\ 0 & \phi_1^2 \end{bmatrix} \right)$$
$$\epsilon_i(t)|\mathbf{b}_i \sim N(0, \sigma^2).$$

The model for the hazard function $h_i(t)$ for the survival endpoint is given below with baseline hazard function $h_0(t)$. In this model, for simplicity, the only covariate included is the treatment indicator $Z_i = \mathbb{I}\{\text{patient } i \text{ receives the new treatment}\}$. The hazard function for patient i is given by

$$h_i(t) = h_0(t) \exp\{\gamma(b_{i0} + b_{i1}t) + \eta Z_i\}.$$

Finally, we are assuming non-informative censoring and the distribution of the censoring random variable for patient i is given by $C_i \sim \exp\{\lambda\}$. In all cases, we shall simulate data to reflect that roughly 10% of patients will be censored for reasons other than end-of-study censoring. To do so, we have used the value $\lambda = 0.022$ and chosen this by trial and error.

We shall simulate data in the case

$$(\mu_0, \mu_1) = (6, 3), \phi_0 = 3.5, \phi_1 = 2.5, \sigma^2 = 10, \qquad (4.32)$$
$$h_0(t) = 5.5, \gamma = 0.03, \lambda = 0.022 \text{ and } \eta = -0.5.$$

Figure 4.1 shows the biomarker trajectory of four randomly generated patients. The value $\sigma^2 = 10$ produces trajectories with a clear trend that is still subject to measurement error. In later sections we shall consider how properties of the model are affected by a change in σ^2. For example, when the longitudinal data is extremely noisy, there may not be any gain from including it in the model.

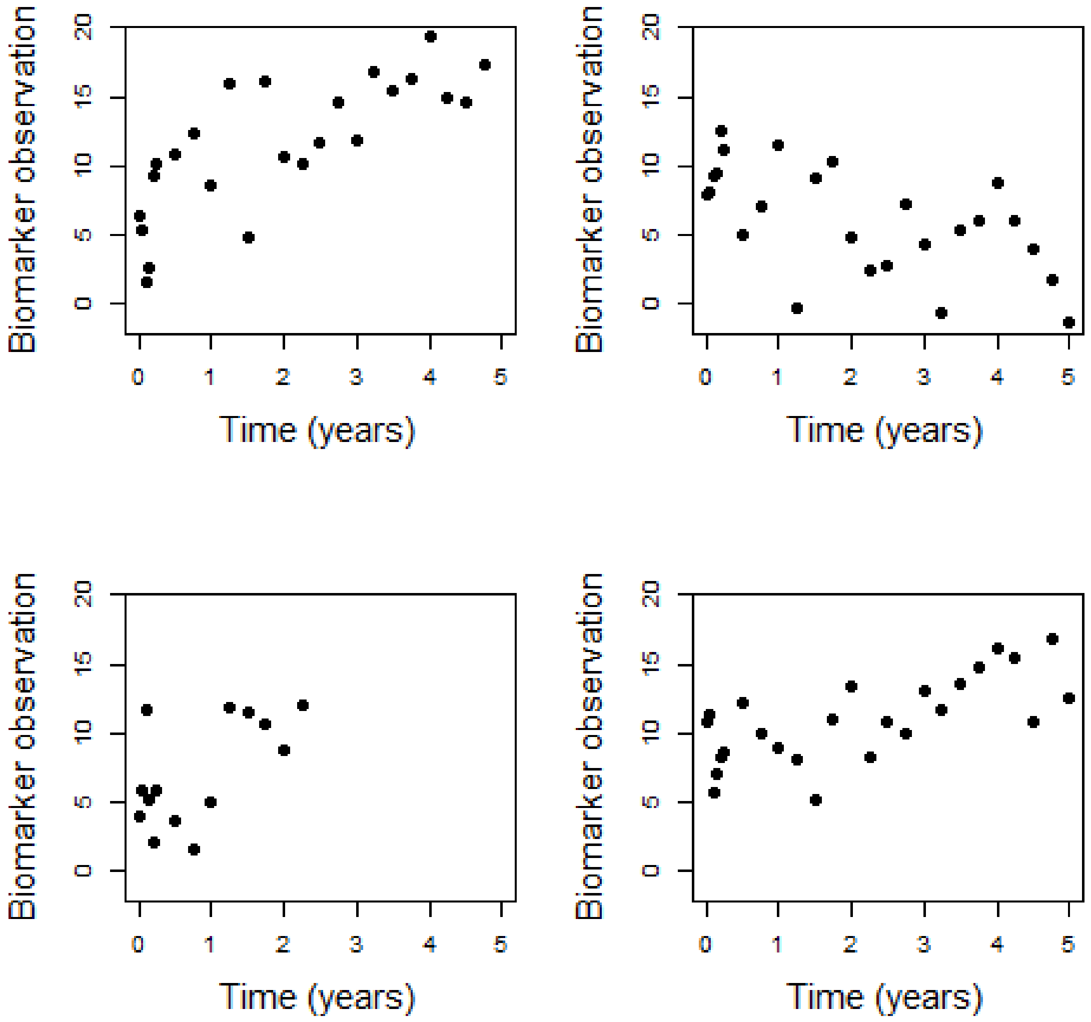


Figure 4.1: Longitudinal observations of four randomly selected patients with parameter values (4.32).

The variance-covariance matrix for the random effects $\mathbf{b}_1, \ldots, \mathbf{b}_n$ has been chosen so that the longitudinal data and the treatment effect have a roughly equal impact on the hazard function. This is demonstrated by Figure 5.3.2 below, which shows the survival function at the mean of each random effect for each treatment arm, and also at one standard deviation of each random effect above and below the mean. For example, the upper dashed blue line is for patient i with $b_{0i} = \mu_0 + \phi_0 = 9.5$ and $b_{1i} = \mu_1 + \phi_1 = 5.5$. At the median survival time $t \approx 3.25$ years, differences in the survival function between treatment arms are roughly equal to differences in the survival function between at the mean and at the mean plus or minus one standard deviation of random effects on the same arm. For comparison, we have also included a similar plot for the AIDS data set found in the JM R package written by Rizopoulos (2010). This plot shows that the biomarker observations dominate the hazard function and that treatment has little effect on survival. Therefore, the parameter values that we have chosen are conservative with respect to the amount of information that comes through the biomarker.

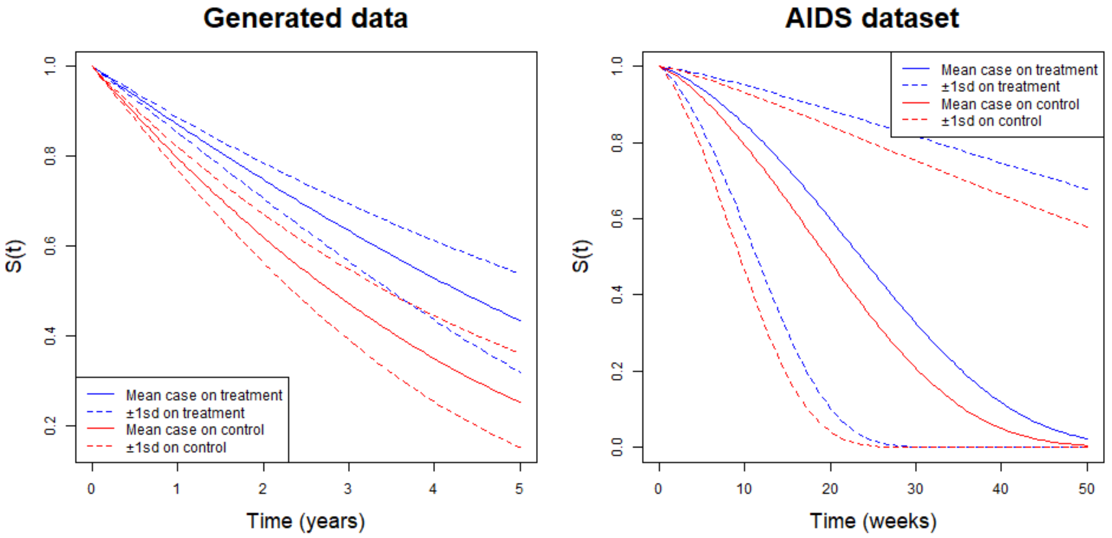


Figure 4.2: Survival function for the simulated data and from AIDS data set with dashed lines showing survival function at 1 standard deviation of random effects above and below mean.

In the hazard rate formula, the coefficient of the longitudinal variable, γ, affects the contribution of the longitudinal data to the hazard ratio. Therefore, in later sections we shall assess the effect of varying γ. This is because we would like to see how properties of the clinical trial vary when the longitudinal data has differing levels of influence.

For completeness, Figure 4.3 gives a histogram of 1000 randomly generated survival times under the joint model. The histogram shows that at the time of the final analysis, 5 years, not all of the events have occurred. The trial is designed with to observe 60% of events at the time of an analysis after 5 years.

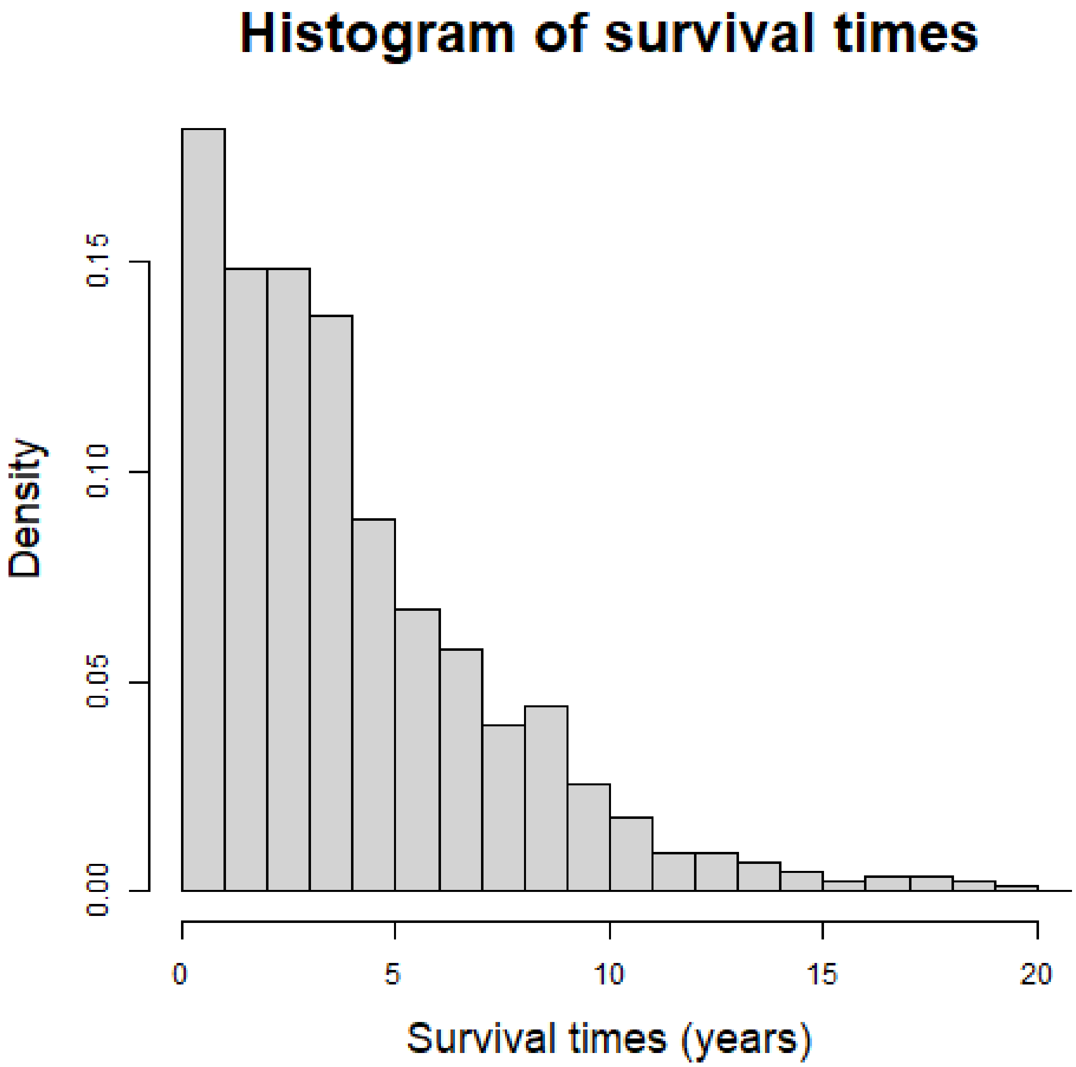


Figure 4.3: Histogram of survival times generated using the parameter values (4.32).

4.3.2 | Fixed sample simulations

Using the parameter values (4.32) for the joint model, we can now simulate clinical trials to find the distribution of the treatment effect estimate, $\hat{\eta}$, using a Monte Carlo method and compare this to the asymptotic theoretical result. For completeness we shall also check the distribution of $\hat{\gamma}$ which is the estimate of the longitudinal data coefficient. In Theorem 4.3 we proved that for a fixed sample, the parameter estimates $\hat{\gamma}$ and $\hat{\eta}$ of the joint model found using the conditional score method had the following distribution:

$$\sqrt{n}\begin{bmatrix}\hat{\gamma} & \gamma_0\\ \hat{\eta} & \eta_0\end{bmatrix} \xrightarrow{d} N\left(\begin{bmatrix}0\\0\end{bmatrix}, \Sigma\right)$$

where

$$\Sigma = A^{-1}B(A^{-1})^T$$

and the matrices A and B of the covariance matrix are defined by equations (4.14) and (4.15). The covariance matrix Σ found using the sandwich method is estimated in practice using estimates $\hat{A}$ and $\hat{B}$ given by Equations (4.18)–(4.19), and is also data dependent. Therefore, for a true distributional comparison, we shall simulate standardised parameter estimates. These are

$$Z_\gamma = \frac{\sqrt{n}(\hat{\gamma}-\gamma_0)}{\sqrt{\Sigma_{11}}}, \qquad Z_\eta = \frac{\sqrt{n}(\hat{\eta}-\eta_0)}{\sqrt{\Sigma_{22}}}. \tag{4.33}$$

Then, it is clear that theory implies

$$Z_\gamma \sim N(0,1) \quad \text{and} \quad Z_\eta \sim N(0,1).$$

Suppose that we consider an alternative approach. For example, let $Z_\gamma = \sqrt{n}\hat{\gamma}/\sqrt{\Sigma_{11}}$, then $Z_\gamma \sim N(\sqrt{n}\gamma_0/\sqrt{\Sigma_{11}}, 1)$. The true value of Σ_{11} is unknown and it is therefore difficult to evaluate the true theoretical mean of Z_γ. For this reason, we have chosen to use the formulation of Z_γ and Z_η given in Equation (4.33), which center these statistics on zero.

To choose a suitable sample size n, we follow the method outlined in Section 2.1.1 and calculate n based on a power requirement. For this trial type 1 error is chosen to be $\alpha = 0.025$ and power is required to be $1-\beta = 0.9$ when $\eta = -0.5$. Therefore, by Equation (2.1), the information needed in the fixed sample study is

$$\mathcal{I}_f = \left(\frac{\Phi^{-1}(0.975)+\Phi^{-1}(0.9)}{-0.5}\right)^2 = 42.03.$$

For survival data Jennison and Turnbull (2000) show, in their Chapter 13, that information is approximately proportional to the number of events divided by 4. We use this assumption to calculate the required sample size. The fixed sample trial is designed with 2 years recruitment and 3 years follow up, where the trial is designed to achieve 60% of events upon termination of the study at time 5 years. In addition, we expect to see 10% of patients leave the study due to censoring other than end-of-study-censoring, and we simulate censoring observations from the distribution $C_i \sim Exp(\lambda)$ where λ is chosen to achieve this 10%. We therefore choose the sample size

$$n = \frac{4\,\mathcal{I}_f}{0.9 \times 0.6} \approx 311.$$

Figures 4.4 and 4.5 show the outcome of this simulation study for a fixed sample trial with 10^4 Monte Carlo replicates. Figure 4.4 shows the outcome when simulating under H_0 with $\eta = 0$ and Figure 4.5 shows the corresponding plot when simulating under H_A using $\eta = -0.5$. The Q-Q plots for the standardised parameter estimates calculated by Equation (4.33) show that $N(0,1)$ is a good fit for Z_γ and Z_η. Further, the histograms show that $Z_\gamma \sim N(0,1)$ and $Z_\eta \sim N(0,1)$ since they closely follow the red line which is the true probability density function of a $N(0,1)$ distribution.

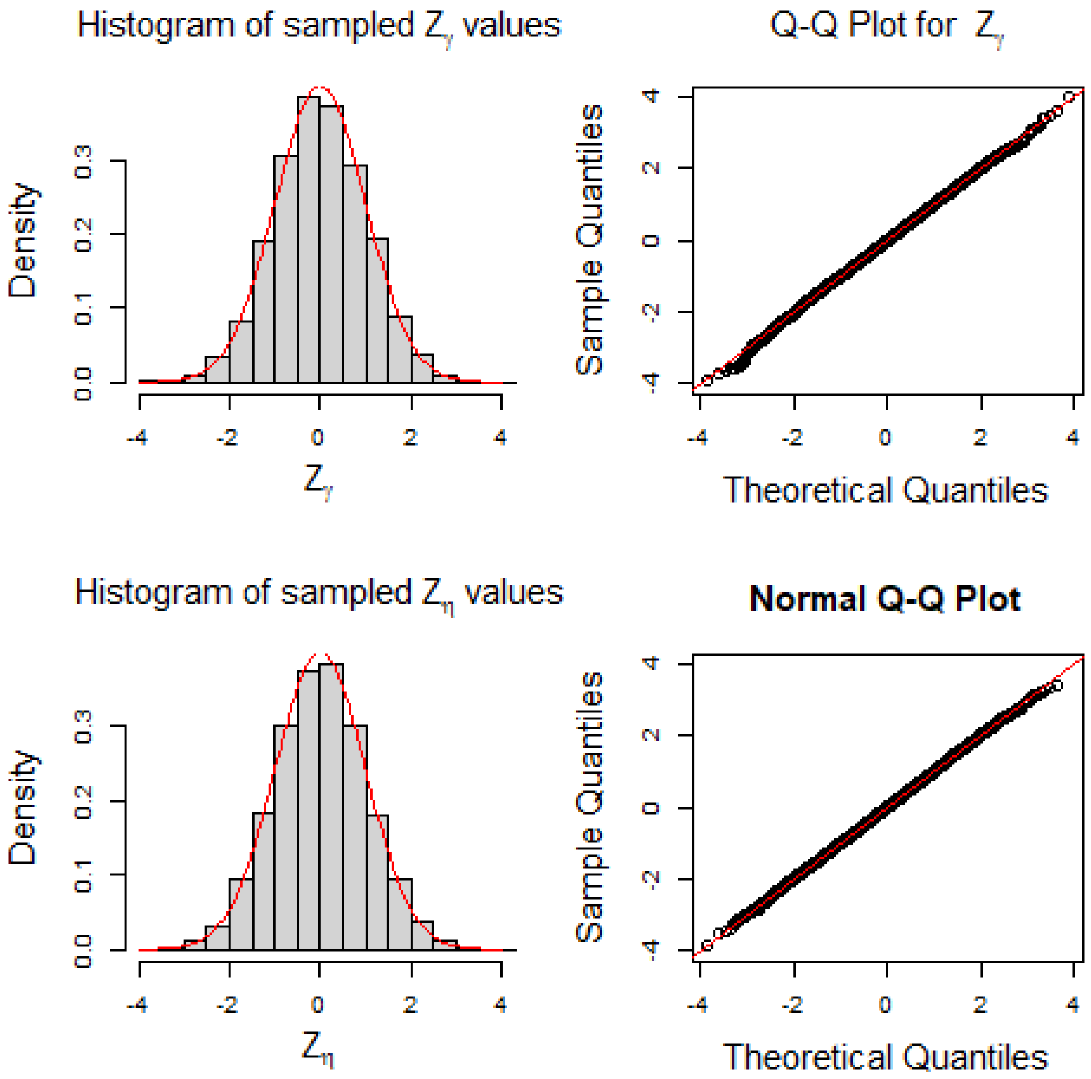


Figure 4.4: Histogram and QQ plots for simulated Z_γ and Z_η in the fixed sample joint model with parameter values (4.32), $\eta = 0$ and 10^4 replicates.

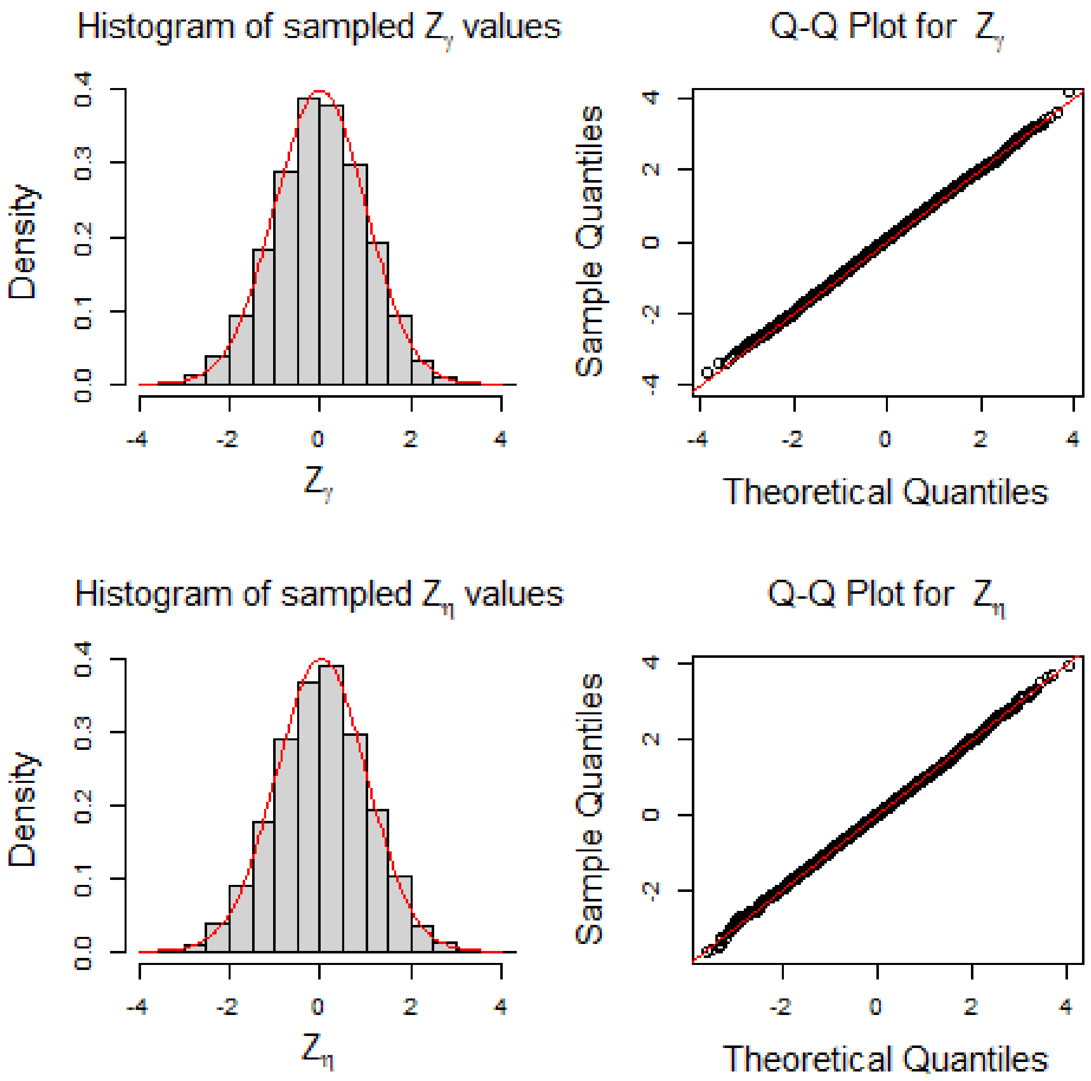


Figure 4.5: Histogram and QQ plots for simulated Z_γ and Z_η in the fixed sample joint model with parameter values (4.32), $\eta = -0.5$ and 10^4 replicates.

4.3.3 | Group sequential simulations

In a similar manner to the fixed sample simulations, we would like to asses the distribution of the sequence of treatment effect estimates in the group sequential trial. We shall use a Monte Carlo method in which for each replicate we simulate a clinical trial and calculate a sequence of treatment effect estimates. For each analysis we can then compare the Monte Carlo distribution of the standardised parameter estimates to the theoretical distribution. The trial design that we have chosen has the first analysis occurring during recruitment and so the first analysis is likely to have a much smaller observed number of events than other analyses. We shall give particular focus to the first analysis to check that the asymptotic results still hold for small sample sizes.

For the design of the group sequential trial, our example will have $K = 5$ analyses. All of the parameter values in the joint model are chosen as in Section 4.3.1 and we shall update the sample size calculation to adjust from fixed sample to group sequential. The calendar analysis times are chosen to be $19, 28, 37, 47$ and 60 months. These analysis times are chosen so that we have roughly evenly spaced information levels and roughly 60% of subjects have events observed by the final analysis, which occurs at 5 years. Similarly to the fixed sample trial, we choose type 1 error $\alpha = 0.025$ and power $1 - \beta = 0.9$. For the group sequential trial we shall use an error-spending test as described in Section 2.1.3 with error spending functions

$$f(t) = \min\{\alpha t^2, \alpha\} \quad \text{and} \quad g(t) = \min\{\beta t^2, \beta\}.$$

Then, by the method described in Section 2.1.3, assuming 5 equally spaced information levels, we calculate that at the final analysis we require information

$$\mathcal{I}_{max} = 46.36.$$

Given that we expect to see 10% of patients leave the study due to censoring other than end-of-study or interim analysis censoring, and 60% of events happen by the final analysis, we therefore choose the following sample size:

$$n = \frac{4\mathcal{I}_{max}}{0.6 \times 0.9} \approx 343.$$

In Theorem 4.4 we proved that for each $k = 1, \dots, 5$, the parameter estimates of the joint model at analysis k are distributed such that

$$\sqrt{n} \begin{bmatrix} \hat{\gamma}^{(k)} & - & \gamma_0 \\ \hat{\eta}^{(k)} & - & \eta_0 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma^{(k)} \right)$$

where

$$\Sigma^{(k)} = (A^{(k)})^{-1} B^{(k)} (A^{(k)})^{-1}$$

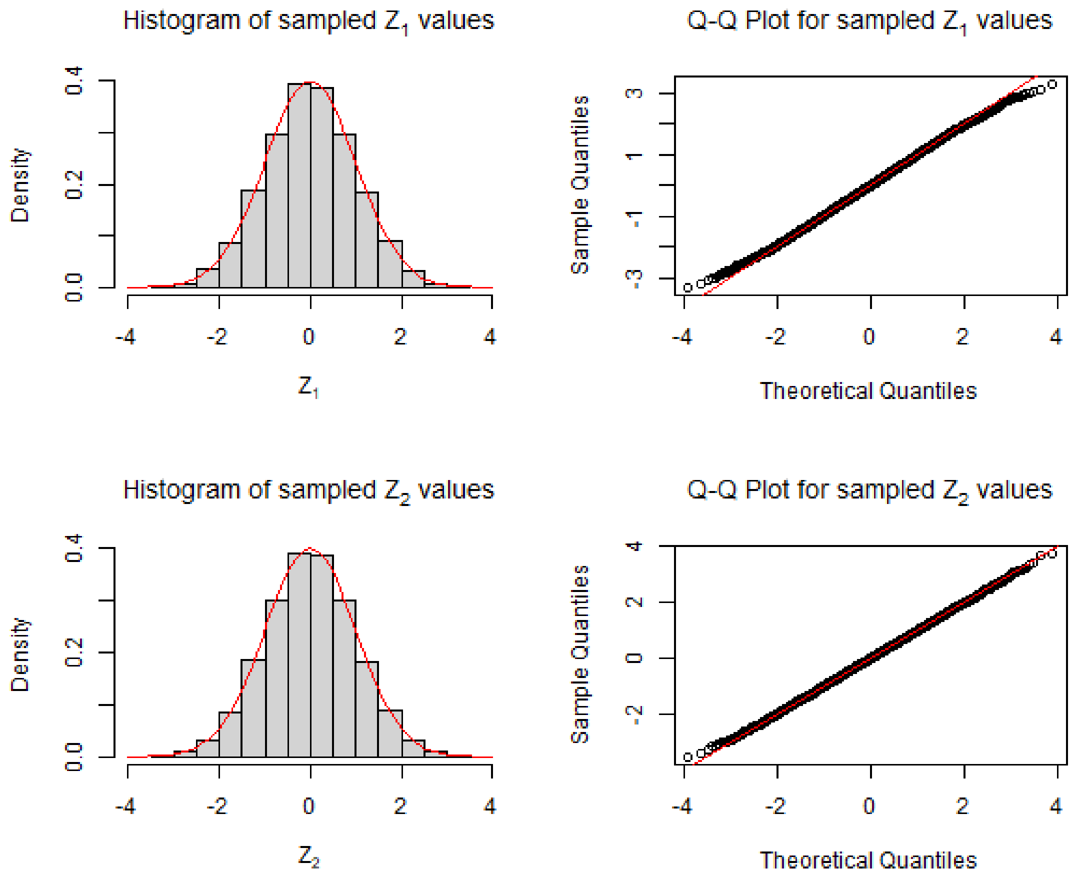
and the matrices $A^{(k)}$ and $B^{(k)}$ are given by equations (4.26) and (4.28) respectively. To check this distribution for the $\hat{\eta}^{(k)}$s, we shall consider the standardised treatment effect estimates at each analysis

$$Z_k = \frac{\sqrt{n}(\hat{\eta}^{(k)} - \eta_0)}{\sqrt{\Sigma_{22}^{(k)}}} \qquad \text{for } k = 1, \ldots, 5.$$

Then we can compare these estimates to the theoretical distribution

$$Z_k \sim N(0,1).$$

Figures 4.6 and 4.7 show the result of this group sequential simulation study with 10^4 Monte Carlo replicates. Figure 4.6 was simulated under H_0 with $\eta = 0$ and Figure 4.7 was simulated under H_A with $\eta = -0.5$.



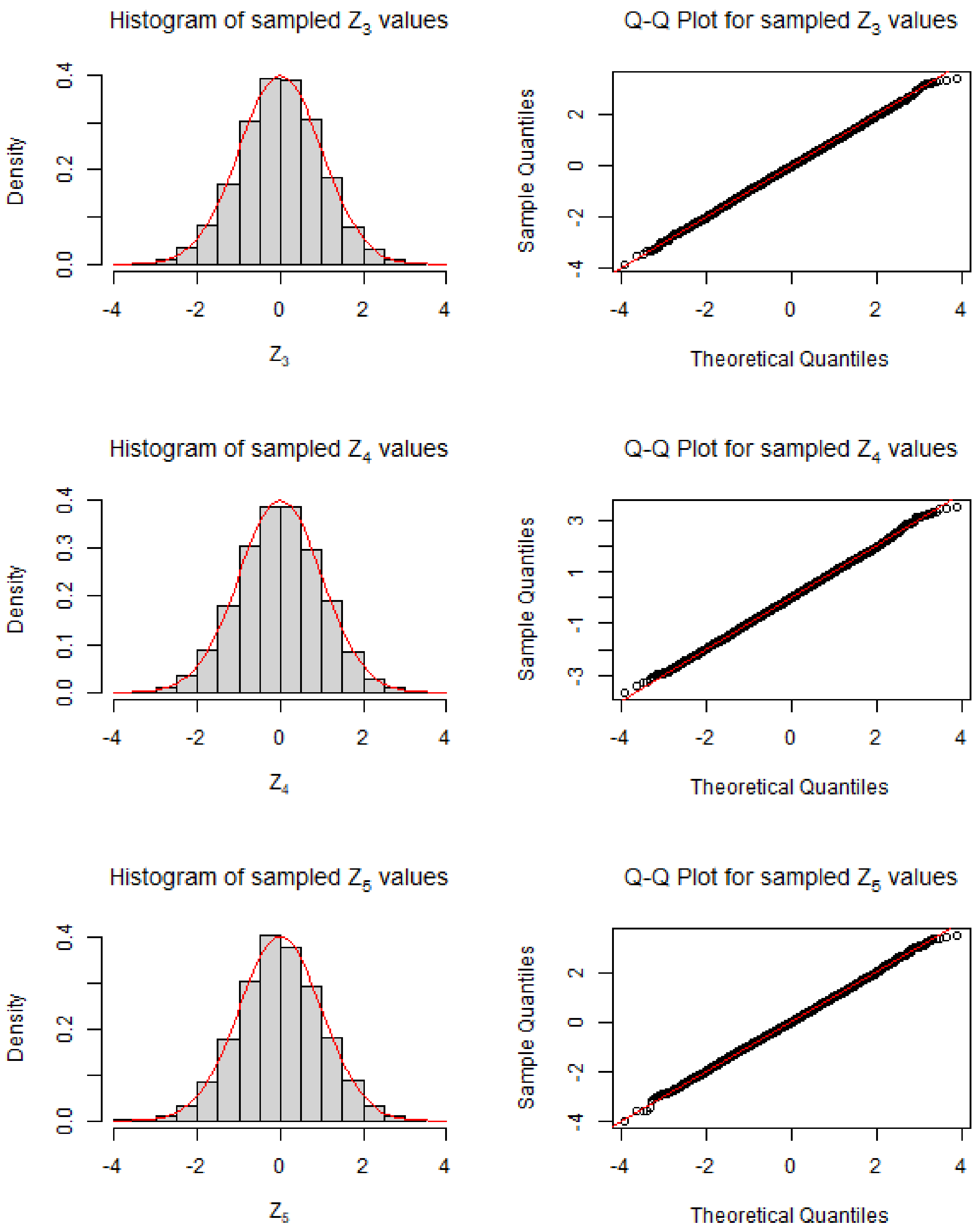
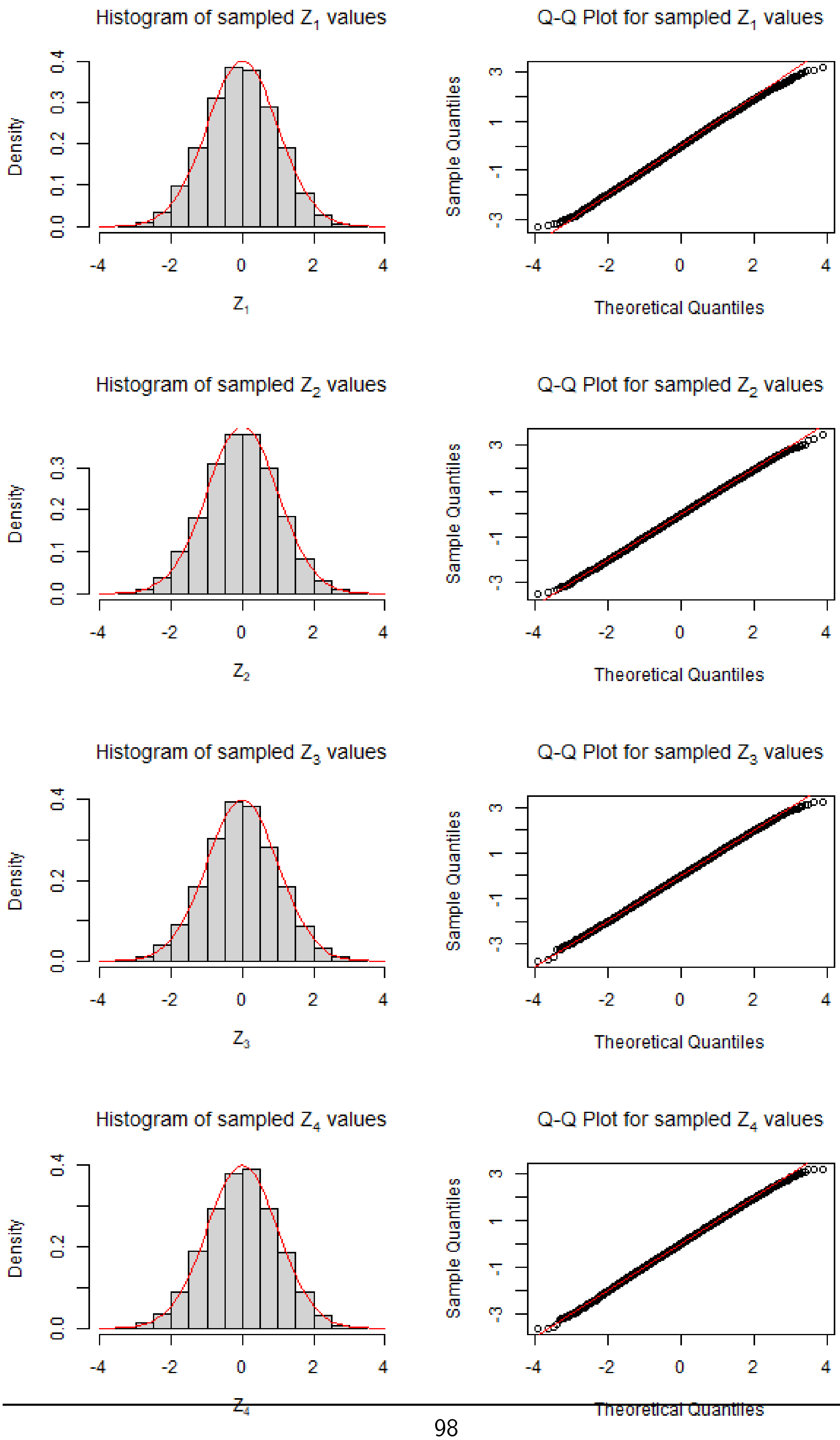


Figure 4.6: Histogram and QQ plots for simulated parameter estimates in the joint model under H_0 using $\eta = 0$ for each analysis of a group sequential trial using 10^4 replicates.


Histogram of sampled Z_1 values
Density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
Z_1
Q-Q Plot for sampled Z_1 values
Sample Quantiles
-3
-1
1
3
-4
-2
0
2
4
Theoretical Quantiles
Histogram of sampled Z_2 values
Density
0.0
0.1
0.2
0.3
-4
-2
0
2
4
Z_2
Q-Q Plot for sampled Z_2 values
Sample Quantiles
-3
-1
1
3
-4
-2
0
2
4
Theoretical Quantiles
Histogram of sampled Z_3 values
Density
0.0
0.2
0.4
-4
-2
0
2
4
Z_3
Q-Q Plot for sampled Z_3 values
Sample Quantiles
-3
-1
1
3
-4
-2
0
2
4
Theoretical Quantiles
Histogram of sampled Z_4 values
Density
0.0
0.2
0.4
-4
-2
0
2
4
Z_4
Q-Q Plot for sampled Z_4 values
Sample Quantiles
-3
-1
1
3
-4
-2
0
2
4
Theoretical Quantiles

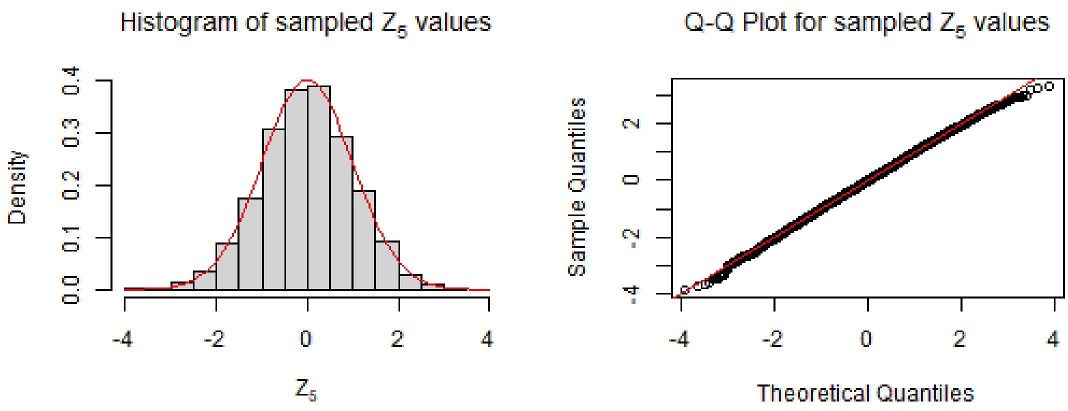


Figure 4.7: Histogram and QQ plots for simulated parameter estimates in the joint model under H_A using $\eta = -0.5$ for each analysis of a group sequential trial using 10^4 replicates.

The Q-Q plots and histograms confirm that the marginal distributions of the standardised treatment effect estimates are such that $Z_k \sim N(0,1)$ for each $k = 1, \dots, 5$. The simulation results for the first analysis do not match the theoretical distribution as closely as for the other analyses, which is seen at the tails of the QQ-plot. This problem is common for both cases $\eta = 0$ and $\eta = -0.5$. The first analysis happens during recruitment and has a small sample size and, with limited follow-up, a small number of events. The mean number of events at the first analysis was 37.6 when $\eta = 0$ and 37.2 when $\eta = -0.5$

At this stage, we have considered the marginal distributions of the treatment effect estimate at each analysis. We shall check that the sequence of treatment effect estimates has the covariance structure given in Theorem 4.4. For $1 \leq k_1 \leq k_2 \leq K$ we have that

$$Cov(\hat{\eta}^{(k_1)}, \hat{\eta}^{(k_2)}) = (A^{k_1})^{-1} B^{(k_1)} (A^{(k_2)})^{-1}$$

which implies that

$$\begin{aligned} Cov(Z_{k_1}, Z_{k_2}) &= Cov\left(\frac{\sqrt{n}(\hat{\eta}^{(k_1)} - \eta_0)}{\sqrt{\Sigma_{22}^{(k_1)}}}, \frac{\sqrt{n}(\hat{\eta}^{(k_2)} - \eta_0)}{\sqrt{\Sigma_{22}^{(k_2)}}} \right) \\ &= \frac{n(A^{k_1})^{-1} B^{(k_1)} (A^{(k_2)})^{-1}}{\sqrt{\Sigma_{22}^{(k_1)} \Sigma_{22}^{(k_2)}}}. \end{aligned}$$

For each $k = 1, \dots, K$, the matrices $A^{(k)}, B^{(k)}$ and $\Sigma^{(k)}$ can be calculated for each clinical trial. Hence, a value of $Cov(Z_{k_1}, Z_{k_2})$ can be obtained for each clinical trial. We can find the value of $\mathbb{E}(Cov(Z_{k_1}, Z_{k_2}))$ using Monte Carlo methods, and compare this to the empirical value of $\widehat{Cov}(Z_{k_1}, Z_{k_2})$ which is found by taking the covariance of all the Z_{k_1} and Z_{k_2} from the simulations. Tables 4.3 and 4.6 show these results.

$$\begin{bmatrix} 1.000 & 0.774 & 0.627 & 0.543 & 0.487 \\ 0.774 & 1.000 & 0.811 & 0.703 & 0.631 \\ 0.627 & 0.811 & 1.000 & 0.867 & 0.779 \\ 0.543 & 0.703 & 0.867 & 1.000 & 0.899 \\ 0.487 & 0.631 & 0.779 & 0.899 & 1.000 \end{bmatrix}$$

Table 4.1: Under H_0, $\eta = 0$

$$\begin{bmatrix} 1.000 & 0.767 & 0.615 & 0.527 & 0.467 \\ 0.767 & 1.000 & 0.804 & 0.689 & 0.611 \\ 0.615 & 0.804 & 1.000 & 0.858 & 0.761 \\ 0.527 & 0.689 & 0.858 & 1.000 & 0.887 \\ 0.467 & 0.611 & 0.761 & 0.887 & 1.000 \end{bmatrix}$$

Table 4.2: Under H_A, $\eta = -0.5$

Table 4.3: Matrix of $\mathbb{E}(Cov(Z_{k1}, Z_{k_2}))$ for group sequential trial with $K = 5$ analyses with 10^4 replicates.

$$\begin{bmatrix} 0.957 & 0.663 & 0.540 & 0.479 & 0.430 \\ 0.663 & 0.975 & 0.783 & 0.686 & 0.617 \\ 0.540 & 0.783 & 0.967 & 0.843 & 0.760 \\ 0.479 & 0.686 & 0.843 & 0.984 & 0.886 \\ 0.430 & 0.617 & 0.760 & 0.886 & 0.989 \end{bmatrix}$$

Table 4.4: Under H_0, $\eta = 0$

$$\begin{bmatrix} 0.967 & 0.682 & 0.549 & 0.467 & 0.415 \\ 0.682 & 0.991 & 0.799 & 0.692 & 0.611 \\ 0.549 & 0.799 & 0.994 & 0.861 & 0.760 \\ 0.467 & 0.692 & 0.861 & 1.004 & 0.887 \\ 0.415 & 0.611 & 0.760 & 0.887 & 0.996 \end{bmatrix}$$

Table 4.5: Under H_A, $\eta = -0.5$

Table 4.6: Matrix of $\widehat{Cov}(Z_{k_1}, Z_{k_2})$ for group sequential trial with $K = 5$ analyses with 10^4 replicates.

We can see that, in general, there is little difference in these matrices and the small deviations are consistent with sampling error. There are small differences between $\mathbb{E}(Cov(Z_{k_1}, Z_{k_2}))$ and $\widehat{Cov}(Z_{k_1}, Z_{k_2})$ whenever $k_1 = 1$ which may be explained by the small number of events at the first analysis. The simulation studies for both the fixed sample and group sequential trials confirm the asymptotic distributional results of the parameter estimates in the joint model. Therefore, we may have confidence to perform a clinical trial based on the joint model using the conditional score method. Care should be taken to avoid small numbers of observed events, which may occur due to early first interim analyses in group sequential trials. This could be adjusted for by making the first analysis later in the recruitment stage or by increasing the total sample size.

4.4 | Designing group sequential trials when the canonical joint distribution does not hold

4.4.1 | Deviation of the parameter estimates from the canonical joint distribution

In Section 4.1 we saw that asymptotically the sequence of treatment effect estimates in a group sequential trial obtained from the joint model is multivariate normally distributed. Further, each of these estimates is asymptotically unbiased. The first two conditions of Definition 2.2 for the canonical distribution of the sequence of test statistics are satisfied. However, for the joint model, we have shown that

$$Var(\hat{\eta}^{(k)}) = \left[(A^{(k)})^{-1}B^{(k)}(A^{(k)})^{-1}\right]_{22} \text{ for } k = 1, \ldots, K \tag{4.34}$$

and

$$Cov(\hat{\eta}^{(k_1)}, \hat{\eta}^{(k_2)}) = \left[A^{(k_1)})^{-1}B^{(k_1)}(A^{(k_2)})^{-1}\right]_{22} \text{ for } k_1 < k_2. \tag{4.35}$$

This implies that the third condition for the canonical joint distribution is not satisfied. In this section, we discuss the implications of performing a group sequential trial when the assumption of a canonical joint distribution fails. We first show that there are some small differences between the matrices $(A^{(k)})^{-1}B^{(k)}$ and the identity matrix I and describe why this difference is important. Then, we present some alternative methods which aim to correct for this violation of the canonical joint distribution. In method 1, the trial is performed acting as though the canonical joint distribution holds, and we present some theory that this method controls the type 1 error rate conservatively when a non-binding futility boundary is used. This theoretical result acts as good evidence that the trial will be conservative with respect to type 1 error when binding futility boundaries are used. For method 2, we create a new estimate which is a linear combination of the treatment effect estimates at previous analyses and the current analysis and we show how this estimate has the canonical joint distribution asymptotically. For method 3, the estimated covariance structure is used to calculate the boundaries of the group sequential trial. For each of these 3 methods, we display properties of the trial through simulation studies.

We saw in Section 4.2.1, that

$$A^{(k)} = \int_0^\infty \begin{bmatrix} \sigma^2 \mathbb{E}(\theta_i(u)) & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} s_c^{(0)}(k,u,\gamma,\eta,\sigma^2)h_0(u)du$$

$$- \int_0^\infty \left(v_c^{(1)}(k,u,\gamma,\eta,\sigma^2) + v_c^{(2)}(k,u,\gamma,\eta,\sigma^2)\right) s_c^{(0)}(k,u,\gamma,\eta,\sigma^2)h_0(u)du \quad (4.36)$$

$$B^{(k)} = \int_0^\infty v_c^{(1)}(k,u,\gamma,\eta,\sigma^2) s_c^{(0)}(k,u,\gamma,\eta,\sigma^2)h_0(u)du. \quad (4.37)$$

and these can be estimated by

$$\hat{A}^{(k)} = \sum_{i=1}^n \int_0^\infty \begin{bmatrix} \hat{\sigma}^{(k)2}\theta_i(u) & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} dN_i(k,u)$$

$$- \sum_{i=1}^n \int_0^\infty \left[V_c^{(1)}(k,u,\hat{\gamma}^{(k)},\hat{\eta}^{(k)},\hat{\sigma}^{(k)2}) + V_c^{(2)}(k,u,\hat{\gamma}^{(k)},\hat{\eta}^{(k)},\hat{\sigma}^{(k)2})\right] dN_i(k,u) \quad (4.38)$$

$$\hat{B}^{(k)} = \sum_{i=1}^n \int_0^\infty \left[V_c^{(1)}(k,u,\hat{\gamma},\hat{\eta},\hat{\sigma}^2)\right] dN_i(k,u). \quad (4.39)$$

If the relationship $A^{(k)} = B^{(k)}$ holds for each $k = 1,\ldots,K$, then

$$Cov(\hat{\eta}^{(k_1)},\hat{\eta}^{(k_2)}) = \left[(A^{(k_2)})^{-1}\right]_{p+1,p+1} = Var(\hat{\eta}^{(k_2)})$$

and the third condition of Definition 2.2 holds. Therefore, we shall assess the magnitude of the problem by considering the matrix $(A^{(k)})^{-1}B^{(k)}$ and to what extent it differs from the identity matrix.

We can find estimates $\hat{A}^{(k)}$ and $\hat{B}^{(k)}$ from simulated data. We have done this using a large sample size of 4800 patients to reduce noise in these estimates. This is appropriate because, although both matrices depend on the sample size n, they can each be written in the form $\hat{A}^{(k)} = (1/n)\sum_{i=1}^n X_i(k,\gamma,\eta)$ and $\hat{B}^{(k)} = (1/n)\sum_{i=1}^n Y_i(k,\gamma,\eta)$ for some functions $X_i(k,\gamma,\eta)$ and $Y_i(k,\gamma,\eta)$. Therefore, in the formula $(\hat{A}^{(k)})^{-1}\hat{B}^{(k)}$, the value of n cancels out, and we are left with a function that converges in distribution to $(A^{(k)})^{-1}B^{(k)}$ as $n \to \infty$. Further, to reduce simulation error, the true values of γ and η are used in this calculation, which is appropriate because of consistency of the estimates $\hat{\gamma}$ and $\hat{\eta}$.

	$\gamma = 0$	$\gamma = 0.03$	$\gamma = 0.06$	$\gamma = 0.09$
$\sigma^2 = 0$	$\begin{bmatrix} 1.00 & 0.00 \\ 0.00 & 1.00 \end{bmatrix}$	$\begin{bmatrix} 1.00 & 0.00 \\ 0.00 & 1.00 \end{bmatrix}$	$\begin{bmatrix} 1.00 & 0.00 \\ 0.00 & 1.00 \end{bmatrix}$	$\begin{bmatrix} 1.00 & 0.00 \\ 0.00 & 1.00 \end{bmatrix}$
$\sigma^2 = 1$	$\begin{bmatrix} 1.00 & 0.00 \\ 0.00 & 1.00 \end{bmatrix}$	$\begin{bmatrix} 1.01 & 0.00 \\ 0.00 & 1.00 \end{bmatrix}$	$\begin{bmatrix} 1.01 & 0.00 \\ 0.00 & 1.00 \end{bmatrix}$	$\begin{bmatrix} 1.02 & 0.00 \\ 0.00 & 1.00 \end{bmatrix}$
$\sigma^2 = 10$	$\begin{bmatrix} 1.03 & 0.00 \\ 0.01 & 1.00 \end{bmatrix}$	$\begin{bmatrix} 1.06 & 0.00 \\ 0.00 & 1.00 \end{bmatrix}$	$\begin{bmatrix} 1.12 & 0.00 \\ -0.02 & 1.00 \end{bmatrix}$	$\begin{bmatrix} 1.22 & 0.00 \\ 0.00 & 1.00 \end{bmatrix}$
$\sigma^2 = 100$	$\begin{bmatrix} 1.28 & 0.00 \\ -0.10 & 1.00 \end{bmatrix}$	$\begin{bmatrix} 1.63 & 0.00 \\ -0.47 & 1.00 \end{bmatrix}$	$\begin{bmatrix} 2.32 & 0.00 \\ -0.01 & 1.00 \end{bmatrix}$	$\begin{bmatrix} 3.49 & 0.00 \\ 1.00 & 1.00 \end{bmatrix}$

Table 4.7: Matrix $(\hat{A}^{(1)})^{-1}\hat{B}^{(1)}$ for parameter values $\gamma = 0, 0.03, 0.06, 0.09$ and $\sigma^2 = 0, 1, 10, 100$ of the joint model for the null hypothesis $\eta = 0$ simulated with 4800 patients.

In Section 4.3 we discussed how varying the parameters γ and σ^2 in the joint model may alter the properties of the trial, and in a similar way, altering the magnitude of these parameters affects the value of the matrix $(A^{(k)})^{-1}B^{(k)}$. This is also clear from Equations (4.36) and (4.37), which give formulas for $A^{(k)}$ and $B^{(k)}$ that are dependent on γ and σ^2.

Table 4.7 below shows the matrix $(\hat{A}^{(1)})^{-1}\hat{B}^{(1)}$ for $\eta = 0$ and different values of γ and σ^2. We have chosen to investigate the properties of this matrix at the first analysis because we see that the majority of problems occur at early interim analyses. We simulated a data set of 4800 patients with parameter values $\gamma = 0, 0.03, 0.06, 0.09, \sigma^2 = 0, 1, 10, 100$ and $\eta = 0$.

The matrices $A^{(k)}, \hat{A}^{(k)}, B^{(k)}$ and $\hat{B}^{(k)}$ are each of dimension 2×2. The function $V^{(2)}(k, u, \gamma, \eta)$ is such that $\left[V^{(2)}(k, u, \gamma, \eta)\right]_{12} = \left[V^{(2)}(k, u, \gamma, \eta)\right]_{22} = 0$, and hence by Equations (4.38) and (4.39) it can be shown that $[\hat{A}^{(k)}]_{12} = [\hat{B}^{(k)}]_{12}$ and $[\hat{A}^{(k)}]_{22} = [\hat{B}^{(k)}]_{22}$. Further simple algebraic manipulation gives $[(\hat{A}^{(k)})^{-1}\hat{B}^{(k)}]_{12} = 1$ and $[(\hat{A}^{(k)})^{-1}\hat{B}^{(k)}]_{22} = 0$ exactly, which is shown in Table 4.7. By definition, the variance of the estimate $\hat{\eta}^{(k)}$ is found in the bottom right entry of the matrix $(\hat{A}^{(k)})^{-1}\hat{B}^{(k)}(\hat{A}^{(k)})^{-1}$ and hence, the bottom row of $(\hat{A}^{(k)})^{-1}\hat{B}^{(k)}$ is of interest here. The fact that $[(\hat{A}^{(1)})^{-1}\hat{B}^{(1)}]_{11}$ is a long way from 1 for $\sigma^2 = 10$ and $\sigma^2 = 100$ is therefore not a problem. As σ^2 increases, the absolute value of $[(\hat{A}^{(1)})^{-1}\hat{B}^{(1)}]_{21}$ increases, but the value of γ has a small impact on the value of $[(\hat{A}^{(1)})^{-1}\hat{B}^{(1)}]_{21}$.

Thus, we may expect large values of σ^2 to affect the achieved type 1 error rate.

Table 4.8 provides a rough estimate for the standard deviation of $[(\hat{A}^{(1)})^{-1}\hat{B}^{(1)}]_{21}$ in Table 4.7 above. Calculation of a single value of $[(\hat{A}^{(1)})^{-1}\hat{B}^{(1)}]_{21}$ with 10^4 patients is computationally expensive and hence we cannot calculate the standard deviation by Monte Carlo. The alternative approach, which we have employed, is to estimate the standard deviation of $[(\hat{A}^{(1)})^{-1}\hat{B}^{(1)}]_{21}$ for 1600 patients, using 10 Monte Carlo replicates each, and use the relationship $\sqrt{Var(\hat{\eta}_n^{(1)})} \propto 1/\sqrt{n}$ to calculate the standard deviation of $\hat{\eta}_n^{(1)}$ for a sample size of $n = 4800$. Table 4.8 shows that the none of the values of $[(\hat{A}^{(1)})^{-1}\hat{B}^{(1)}]_{21}$ are significantly different from zero.

	$\gamma = 0$	$\gamma = 0.03$	$\gamma = 0.06$	$\gamma = 0.09$
$\sigma^2 = 0$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
$\sigma^2 = 1$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
$\sigma^2 = 10$	0.01 (0.02)	0.00 (0.02)	-0.02 (0.04)	0.00 (0.05)
$\sigma^2 = 100$	-0.10 (0.25)	-0.47 (0.37)	-0.01 (0.54)	1.00 (0.56)

Table 4.8: Mean and standard deviation of the estimate for $[(\hat{A}^{(1)})^{-1}\hat{B}^{(1)}]_{21}$ to 2 decimal places, for parameter values $\gamma = 0, 0.03, 0.06, 0.09$ and $\sigma^2 = 0, 1, 10, 100$ of the joint model under the null hypothesis $\eta = 0$.

Similarly, Table 4.9 below shows the matrix $(\hat{A}^{(1)})^{-1}\hat{B}^{(1)}$ under the alternative hypothesis when $\eta = -0.5$. The absolute value of $[(\hat{A}^{(1)})^{-1}\hat{B}^{(1)}]_{21}$ increases as γ increases and also increases as σ^2 increases and hence, we expect the power calculation to be affected by large values of γ and σ^2.

	$\gamma = 0$	$\gamma = 0.03$	$\gamma = 0.06$	$\gamma = 0.09$
$\sigma^2 = 0$	$\begin{bmatrix} 1.00 & 0.00 \\ 0.00 & 1.00 \end{bmatrix}$	$\begin{bmatrix} 1.00 & 0.00 \\ 0.00 & 1.00 \end{bmatrix}$	$\begin{bmatrix} 1.00 & 0.00 \\ 0.00 & 1.00 \end{bmatrix}$	$\begin{bmatrix} 1.00 & 0.00 \\ 0.00 & 1.00 \end{bmatrix}$
$\sigma^2 = 1$	$\begin{bmatrix} 1.00 & 0.00 \\ 0.00 & 1.00 \end{bmatrix}$	$\begin{bmatrix} 1.01 & 0.00 \\ 0.00 & 1.00 \end{bmatrix}$	$\begin{bmatrix} 1.01 & 0.00 \\ -0.01 & 1.00 \end{bmatrix}$	$\begin{bmatrix} 1.02 & 0.00 \\ -0.03 & 1.00 \end{bmatrix}$
$\sigma^2 = 10$	$\begin{bmatrix} 1.03 & 0.00 \\ -0.01 & 1.00 \end{bmatrix}$	$\begin{bmatrix} 1.05 & 0.00 \\ -0.11 & 1.00 \end{bmatrix}$	$\begin{bmatrix} 1.10 & 0.00 \\ -0.19 & 1.00 \end{bmatrix}$	$\begin{bmatrix} 1.17 & 0.00 \\ -0.25 & 1.00 \end{bmatrix}$
$\sigma^2 = 100$	$\begin{bmatrix} 1.22 & 0.00 \\ -0.38 & 1.00 \end{bmatrix}$	$\begin{bmatrix} 1.48 & 0.00 \\ -1.08 & 1.00 \end{bmatrix}$	$\begin{bmatrix} 1.97 & 0.00 \\ -1.84 & 1.00 \end{bmatrix}$	$\begin{bmatrix} 2.92 & 0.00 \\ -2.94 & 1.00 \end{bmatrix}$

Table 4.9: Matrix $(\hat{A}^{(1)})^{-1}\hat{B}^{(1)}$ for parameter values $\gamma = 0, 0.03, 0.06, 0.09$ and $\sigma^2 = 0, 1, 10, 100$ of the joint model for the alternative hypothesis $\eta = -0.5$ simulated with 4800 patients.

Table 4.10 shows the standard deviations of the estimates $[(\hat{A}^{(1)})^{-1}\hat{B}^{(1)}]_{21}$ of Table 4.9. These standard deviation estimates are calculated in the same way as those in Table 4.8, by fitting standard deviation estimates 1600 patients and using the relationship $\sqrt{Var(\hat{\eta}_n^{(1)})} \propto 1/\sqrt{n}$ to find the standard deviation for $\hat{\eta}_n^{(1)}$ when the sample size is $n = 4800$ patients. The values of $[(\hat{A}^{(1)})^{-1}\hat{B}^{(1)}]_{21}$ are significantly different from zero whenever $\sigma^2 = 100$. Hence, the matrix $(\hat{A}^{(k)})^{-1}\hat{B}^{(k)}$ is significantly different from the identity matrix when $\sigma^2 = 100$.

	$\gamma = 0$	$\gamma = 0.03$	$\gamma = 0.06$	$\gamma = 0.09$
$\sigma^2 = 0$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
$\sigma^2 = 1$	0.00 (0.00)	0.00 (0.01)	-0.01 (0.03)	-0.03 (0.01)
$\sigma^2 = 10$	-0.01 (0.03)	-0.11 (0.06)	-0.19 (0.10)	-0.25 (0.13)
$\sigma^2 = 100$	-0.38 (0.29)	-1.08 (0.31)	-1.84 (0.71)	-2.94 (0.86)

Table 4.10: Mean and standard deviation of the estimate for $[(\hat{A}^{(1)})^{-1}\hat{B}^{(1)}]_{21}$ to 2 decimal places, for parameter values $\gamma = 0, 0.03, 0.06, 0.09$ and $\sigma^2 = 0, 1, 10, 100$ of the joint model for the alternative hypothesis $\eta = -0.5$.

The value $\sigma^2 = 100$ implies incredibly noisy data and this value is so high that we would not expect to use the longitudinal data in a clinical trial. In Figure 4.1 we showed a sample of longitudinal data simulated using $\sigma^2 = 10$ and this data was already quite noisy. Hence, the fact that $[(\hat{A}^{(1)})^{-1}\hat{B}^{(1)}]_{21}$ is significantly different from zero for $\sigma^2 = 100$ is not a concern.

4.4.2 | Method 1: Canonical distribution assumed

We consider three methods for creating a group sequential trial when the canonical joint distribution of Definition 2.1 does not hold. In the first method, we construct the group sequential test by estimating $Var(\hat{\eta}^{(k)}), k = 1, \ldots, K$ from the data and supposing $Cov(\hat{\eta}^{(k_1)}, \hat{\eta}^{(k_2)})$ for $k_1 < k_2$ are as specified in the canonical joint distribution. We show, through simulation, that this method performs satisfactorily in practice using binding futility boundaries, with error rates diverging minimally from planned significance and power. As further evidence, we shall later show that this method is conservative with respect to type 1 error when we use non-binding futility boundaries. This method is simple to perform and computationally efficient.

We shall follow two approaches to find type 1 and type 2 error rates. In the first approach, a single large sample is generated and theoretical probabilities for accepting or rejecting H_0 are calculated. Calculating theoretical probabilities for accepting or rejecting H_0 requires knowing the true variance-covariance matrix Σ, and hence, using a (very) large sample of patients is a convenient way to estimate the entries of Σ with minimal noise. During the large sample approach, we scale the information levels to match a clinical trial which recruits 343 patients and has planned $\mathcal{I}_{max} = 46.36$ as described in Section 4.3.3. This is to ensure that deviations from planned power $1-\beta = 0.9$ are not due to sample size. The steps below describe a method for calculating error rates for a clinical trial with a large sample size.

1. Choose parameter values γ and σ^2 and set $\eta = 0$. Simulate a clinical trial with 4800 patients.

2. Using this data set, calculate the covariance matrix

$$V = Cov((\hat{\eta}^{(1)}, \dots, \hat{\eta}^{(K)})^T, (\hat{\eta}^{(1)}, \dots, \hat{\eta}^{(K)})^T)$$

   for the sequence of treatment effect estimates using equations (4.30) and (4.31) and Theorem 4.4. The true values of the parameters η and γ, which are used to simulate the data, can be used in the calculation of the matrices $\hat{A}(\eta, \gamma)$ and $\hat{B}(\eta, \gamma)$ because of the consistency of the estimates $\hat{\gamma}$ and $\hat{\eta}$.

3. To scale the covariance matrix to the correct sample size, let

$$\Sigma = \frac{V\mathcal{I}_{max}^{-1}}{V_{kk}}.$$

4. Let $\tilde{\Sigma}$ be the covariance matrix under the assumption that the canonical joint distribution holds. That is, let

$$\begin{aligned} \tilde{\Sigma}_{kk} &= \Sigma_{kk} \qquad \text{for } k = 1, \dots, K \\ \tilde{\Sigma}_{k_1 k_2} &= \Sigma_{k_1 k_1} \qquad \text{for } k_1 < k_2. \end{aligned}$$

5. Calculate the boundaries of the group sequential trial assuming that the canonical joint distribution holds. This process was described in Section 2.1.3 for the Z-statistic. For an error spending test with error spending functions $f(t) = \min\{\alpha t^2, \alpha\}$ and $g(t) = \min\{\beta t^2, \beta\}$, information levels $\mathcal{I}_k = \tilde{\Sigma}_{kk}^{-1}$ and vinding futility boundary, let the boundaries points $\tilde{a}_1, \dots, \tilde{a}_K$ and $\tilde{b}_1, \dots, \tilde{b}_K$

be defined as the solutions to the equations

$$
\begin{aligned}
&\mathbb{P}_{\eta=0}\{\hat{\eta}^{(1)} > \tilde{b}_1\} = f(\mathcal{I}_1/\mathcal{I}_{max})\\
&\mathbb{P}_{\eta=\delta}\{\hat{\eta}^{(1)} < \tilde{a}_1\} = g(\mathcal{I}_1/\mathcal{I}_{max})\\
&\mathbb{P}_{\eta=0}\{\tilde{a}_1 < \hat{\eta}^{(1)} < \tilde{b}_1, \dots, \tilde{a}_{k-1} < \hat{\eta}^{(k-1)} < \tilde{b}_{k-1}, \hat{\eta}^{(k)} > \tilde{b}_k\}\\
&\qquad\qquad = f(\mathcal{I}_k/\mathcal{I}_{max}) - f(\mathcal{I}_{k-1}/\mathcal{I}_{max}) \qquad \text{for } k = 2, \dots, K\\
&\mathbb{P}_{\eta=\delta}\{\tilde{a}_1 < \hat{\eta}^{(1)} < \tilde{b}_1, \dots, \tilde{a}_{k-1} < \hat{\eta}^{(k-1)} < \tilde{b}_{k-1}, \hat{\eta}^{(k)} < \tilde{a}_k\}\\
&\qquad\qquad = g(\mathcal{I}_k/\mathcal{I}_{max}) - g(\mathcal{I}_{k-1}/\mathcal{I}_{max}) \qquad \text{for } k = 2, \dots, K.
\end{aligned}
$$

6. Find the type 1 error as the probability of crossing any of the upper boundaries $\tilde{b}_1, \dots, \tilde{b}_K$ before crossing a lower boundary using the *true* distribution of $\hat{\eta}^{(1)}, \dots, \hat{\eta}^{(K)}$ according to the covariance matrix Σ. This calculation can be performed in R using the package "mvtnorm" by Genz et al. (2020).

7. Repeats steps 1–5 but with $\eta = -0.5$ used to simulate the data, obtaining new estimates for V, Σ, $\tilde{\Sigma}$ and boundary points $\tilde{a}_1, \dots, \tilde{a}_K$ and $\tilde{b}_1, \dots, \tilde{b}_K$.

8. Calculate the type 2 error as the probability of crossing any of the lower boundaries $\tilde{a}_1, \dots, \tilde{a}_K$ before crossing an upper boundary using the *true* distribution of $\hat{\eta}^{(1)}, \dots, \hat{\eta}^{(K)}$ according to the covariance matrix Σ.

As an example, let the parameters be chosen as $\gamma = 0.03, \sigma^2 = 10$ and $\eta = 0$. The matrices Σ and $\tilde{\Sigma}$ calculated in steps 3 and 4 above, are shown below.

$$
\begin{bmatrix}
0.110 & 0.059 & 0.040 & 0.029 & 0.024\\
0.059 & 0.053 & 0.035 & 0.026 & 0.021\\
0.040 & 0.035 & 0.035 & 0.026 & 0.022\\
0.029 & 0.026 & 0.026 & 0.026 & 0.022\\
0.024 & 0.021 & 0.022 & 0.022 & 0.022
\end{bmatrix}
$$

Table 4.11: Matrix Σ.

$$
\begin{bmatrix}
0.110 & 0.053 & 0.035 & 0.026 & 0.022\\
0.053 & 0.053 & 0.035 & 0.026 & 0.022\\
0.035 & 0.035 & 0.035 & 0.026 & 0.022\\
0.026 & 0.026 & 0.026 & 0.026 & 0.022\\
0.022 & 0.022 & 0.022 & 0.022 & 0.022
\end{bmatrix}
$$

Table 4.12: Matrix $\tilde{\Sigma}$.

Table 4.13: Covariance matrices Σ and $\tilde{\Sigma}$ for group sequential trial with $K = 5$ analyses and parameter values $\gamma = 0.03, \sigma^2 = 10$ and $\eta = 0$.

The columns of Table 4.14 headed "Large sample" show the computational results for the error rates, calculated using the steps outlined above. In each case, the type 1 error is very close to 0.025 and is always conservative. The difference between the large sample type 1 error and the planned significance level α is nominally small. Further, the large sample power is always close to 0.9. This trial was designed with equally spaced information levels however in practice, these information levels are not exactly equally spaced and this is the reason why power is not exactly equal to 0.9. The fact that asymptotic theory does not lead to the canonical joint distribution is not a problem because the distribution that does arise is so close to the canonical joint distribution that the impact on type 1 error is negligible. Therefore, the effect is a small degree of conservatism.

γ	σ^2	**Type 1 error**		**Power**	
		Large sample	**Simulation**	**Large sample**	**Simulation**
0	0	0.0249	0.0224	0.9026	0.8918
0	1	0.0249	0.0240	0.9024	0.8926
0	10	0.0249	0.0229	0.9035	0.8897
0	100	0.0249	0.0228	0.9030	0.8887
0.03	0	0.0249	0.0221	0.9015	0.8911
0.03	1	0.0249	0.0247	0.9019	0.8912
0.03	10	0.0249	0.0236	0.9019	0.8949
0.03	100	0.0249	0.0277	0.9019	0.8914
0.06	0	0.0249	0.0221	0.9006	0.8937
0.06	1	0.0249	0.0259	0.9006	0.8913
0.06	10	0.0249	0.0228	0.9009	0.8941
0.06	100	0.0249	0.0265	0.9009	0.8911
0.09	0	0.0249	0.0230	0.8995	0.8931
0.09	1	0.0250	0.0229	0.8995	0.8837
0.09	10	0.0249	0.0250	0.8996	0.8898
0.09	100	0.0250	0.0242	0.8996	0.8725

Table 4.14: Method 1: Type 1 and error and power calculated using large sample theory and a simulation study with 343 patients and 10^4 replicates for parameter values $\gamma = 0, 0.03, 0.06, 0.09$ and $\sigma^2 = 0, 1, 10, 100$.

The columns headed “Simulation” of Table 4.14 also show estimates obtained by our second approach for calculating type 1 and 2 error rates. This was by simulating 10^4 data sets, each with a sample size $n = 343$. For each data set, we calculate the estimates $\hat{\eta}^{(1)}, \ldots, \hat{\eta}^{(K)}$ and estimates of the covariance matrices $\Sigma^{(1)}, \ldots, \Sigma^{(K)}$ using estimates $\hat{A}^{(1)}, \ldots, \hat{A}^{(K)}, \hat{B}^{(1)}, \ldots, \hat{B}^{(K)}$ given by Equations (4.38) and (4.39). The boundary points $\tilde{a}_1, \ldots, \tilde{a}_K$ and $\tilde{b}_1, \ldots, \tilde{b}_K$ are then calculated under the assumption that the canonical joint distribution holds and the error is calculated as the proportion of replicates that accept or reject the null hypothesis. The choice and calculation of n is discussed in Section 4.3 and is chosen to attain power of 0.9. This sample size calculation is approximate and is one of the reasons why the power estimate is not 0.9 exactly. This simulation study is particularly computationally expensive and with 10^4 replicates, there is noise in the simulation results. Taking this into account, the simulation results support the relevance of asymptotic theory since all empirical type 1 error rates are within 2 standard deviations of 0.025.

In summary, method 1 is convenient to use when the canonical joint distribution does not hold because type 1 error is conservative. There may be a slight loss of power relative to the specified target, however if the parameter values γ and σ^2 are not extreme, then this loss of power is minimal.

4.4.3 | Proof that type 1 error rates are conservative under the canonical joint distribution assumption

We now consider how a clinical trial is affected when the canonical joint distribution does not hold and we use a non-binding futility boundary. We shall discuss the case $Cov(\hat{\eta}^{(k_1)}, \hat{\eta}^{(k_2)}) \geq Var(\hat{\eta}^{(k_2)})$ since this is the case suggested by simulation results. We consider performing the group sequential trial proceeding as if the canonical joint distribution does hold and assess how this affects the error rates. In particular, we shall prove that we have type 1 error less than α, where α is the planned type 1 error. This means that the trial is conservative with respect to type 1 error. In Section 4.4.2, we showed by simulation, that the magnitude of the deviances from planned type 1 error α and planned power $1 - \beta$ were very small when a binding futility boundary is used.

We are considering the weaker case where the futility boundary is non-binding. This is where stopping for futility at an interim analysis is not mandatory. The calculation of the type 1 error therefore only depends on the upper boundary $b_1, \ldots, b_K$. Alternatively we can think of setting $a_1 = \cdots = a_K = -\infty$. This limitation ensures that the theoretical result holds, however we believe that this scenario is a popular design choice and is therefore useful to present. Further, this result is good evidence that the canonical joint distribution holds for a trial which uses a binding futility function and we have shown motivating numerical evidence for such a trial in Section 4.4.1.

To prove that the type 1 error rates are conservative, we shall compare the probabilities of crossing the boundaries of a group sequential trial for two sequences of treatment effect estimates; one where the canonical joint distribution does not hold and one where this assumption does hold. Suppose that $\hat{\eta}_1, \ldots, \hat{\eta}_K$ are the sequence of treatment effect estimates in a group sequential trial, with K analyses, that are calculated using the conditional score method. Let the true variance-covariance matrix for this sequence of estimates be Σ. Under H_0 we have

$$\begin{bmatrix} \hat{\eta}_1 \\ \hat{\eta}_2 \\ \vdots \\ \hat{\eta}_K \end{bmatrix} \sim N_K \left[\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \ldots & \Sigma_{1K} \\ \Sigma_{12} & \Sigma_{22} & & \\ \vdots & & \ddots & \vdots \\ \Sigma_{1K} & & \ldots & \Sigma_{KK} \end{pmatrix} \right]. \tag{4.40}$$

Proceeding using method 1, we let the information levels be calculated as $\mathcal{I}_k = (\Sigma_{kk})^{-1}$ and the $Z-$statistic is given by $Z_k = \hat{\eta}_k \sqrt{\mathcal{I}_k}$ for $k = 1, \ldots, K$. Under H_0 the sequence of $Z-$statistics therefore has the following distribution

$$\begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_K \end{bmatrix} \sim N_K \left[\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{12} & \ldots & \rho_{1K} \\ \rho_{12} & 1 & & \\ \vdots & & \ddots & \vdots \\ \rho_{1K} & & \ldots & 1 \end{pmatrix} \right] \tag{4.41}$$

where entries of the covariance matrix are given by

$$\rho_{k_1 k_2} = Cov(Z_{k_1}, Z_{k_2}) = \frac{\Sigma_{k_1 k_2}}{\sqrt{\Sigma_{k_1 k_1} \Sigma_{k_2 k_2}}} \qquad \text{for } 1 \leq k_1 < k_2 \leq K. \tag{4.42}$$

Suppose instead, that we have a different sequence of treatment effect estimates $\hat{\eta}_1^*, \ldots, \hat{\eta}_K^*$ with the following distribution:

$$\begin{bmatrix} \hat{\eta}_1^* \\ \hat{\eta}_2^* \\ \vdots \\ \hat{\eta}_K^* \end{bmatrix} \sim N_K \left[\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{22} & \ldots & \Sigma_{KK} \\ \Sigma_{22} & \Sigma_{22} & & \\ \vdots & & \ddots & \vdots \\ \Sigma_{KK} & & \ldots & \Sigma_{KK} \end{pmatrix} \right]$$

where $\Sigma_{11}, \ldots, \Sigma_{KK}$ are the same as in (4.40). Using the same information levels given by $\mathcal{I}_k = (\Sigma_{kk})^{-1}$, we define the standardised statistics $Z_k^* = \hat{\eta}_k^* \sqrt{\mathcal{I}_k}$ for

$k = 1, \ldots, K$. The distribution of these $Z-$statistics is therefore given by

$$\begin{bmatrix} Z_1^* \\ Z_2^* \\ \vdots \\ Z_K^* \end{bmatrix} \sim N_K \left[\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{12}^* & \cdots & \rho_{1K}^* \\ \rho_{12}^* & 1 & & \\ \vdots & & \ddots & \vdots \\ \rho_{1K}^* & & \cdots & 1 \end{pmatrix} \right] \tag{4.43}$$

where entries of the covariance matrix are given by

$$\rho_{k_1 k_2}^* = Cov(Z_{k_1}^*, Z_{k_2}^*) = \sqrt{\frac{\Sigma_{k_2 k_2}}{\Sigma_{k_1 k_1}}} \quad \text{for } 1 \leq k_1 < k_2 \leq K. \tag{4.44}$$

This sequence of treatment effect estimates $\hat{\eta}_1^*, \ldots, \hat{\eta}_K^*$ therefore has the canonical joint distribution given by Definition 2.1 and $Z_1^*, \ldots, Z_K^*$ has the canonical joint distribution for a sequence of Z-statistics, with information levels $\mathcal{I}_k = 1/\Sigma_{kk}$ for $k = 1, \ldots, K$, given by Definition 2.1.

The upper boundary points $b_1, \ldots, b_K$ are calculated under the assumption that the canonical joint distribution holds and as to give a group sequential test with the correct type 1 error rate α. Full details of this calculation are given in Section 2.1. Hence, for planned type 1 error α and using the fact that the canonical joint distribution holds for $Z_1^*, \ldots, Z_K^*$, we have that

$$\mathbb{P}(Z_1^* > b_1 \cup Z_2^* > b_2 \cup \cdots \cup Z_K^* > b_K) = \alpha. \tag{4.45}$$

We consider the probability of rejecting H_0 when we apply this boundary to the sequence $Z_1, \ldots, Z_K$. We aim to prove that

$$\mathbb{P}(Z_1 > b_1 \cup Z_2 > b_2 \cup \cdots \cup Z_K > b_K) \leq \alpha.$$

The following conditions are needed for the proof that type 1 error is conservative.

Conditions 4.3.

1. *The upper boundary of a group sequential trial, on the $Z-$scale, given by $b_1, \ldots, b_K$ is such that*

$$b_1 \geq b_2 \geq \cdots \geq b_K \geq 0.$$

2. *For all $k_1 < k_2$,*

$$\Sigma_{k_1 k_2} \geq \Sigma_{k_2 k_2}.$$

3. The futility boundary is non-binding. We have

$$a_1 = a_2 = \cdots = a_K = -\infty.$$

Condition 1 of Conditions 4.3 appears to be common in practice, for example, the O'Brien and Fleming (1979), Haybittle (1971) and Pocock (1977) boundaries all satisfy this condition. Condition 2 of Conditions 4.3 should be checked by simulation before proceeding with the analysis. To do so, the investigator would choose sensible values for all the parameters in the joint model, simulate a large dataset of 4800 patients using these parameter values, and calculate an estimate for the variance-covariance matrix Σ for the sequence of estimates $\hat{\eta}^{(1)}, \dots, \hat{\eta}^{(K)}$. This process was described in steps 1–2 for calculating the large sample sample type 1 error when applying method 1. We have found that calculations for various examples has always lead to condition 2 being satisfied. Further, the scenarios that we have checked span a 3-dimensional grid of η, γ and σ^2 values each ranging from small to large and hence, we believe that the scenarios we have checked span a suitable range of the parameter values. In the rare event that this condition is checked and does not appear to hold, a solution is to employ method 2, which will be described in Section 4.4.4, and this ensures that type 1 error will not be inflated. It can also be seen by simple algebraic manipulation that condition 2 implies

$$\rho_{k_1 k_2} \geq \rho^*_{k_1 k_2} \quad \text{for } k_1 < k_2.$$

The following theorem shows that type 1 error is lower than the required significance level α.

Proposition 4.5. *Let $Z_1, \dots, Z_K$ be the standardised statistics of a group sequential trial with distribution given by* (4.41) *and let $Z^*_1, \dots, Z^*_K$ be the statistics with distribution given by* (4.43)*. Let α be the planned type 1 error and suppose that $b_1, \dots, b_K$ are the upper boundary points on the Z-scale such that*

$$\mathbb{P}(Z^*_1 > b_1 \cup Z^*_2 > b_2 \cup \cdots \cup Z^*_K > b_K) = \alpha.$$

Suppose that Conditions 4.3 hold. Then the Type 1 error when applying the boundary for $Z^*_1, \dots, Z^*_K$ to $Z_1, \dots, Z_K$ is

$$\mathbb{P}(Z_1 > b_1 \cup Z_2 > b_2 \cup \cdots \cup Z_K > b_K) \leq \alpha.$$

We shall prove that Proposition 4.5 holds for the case $K = 2$ and we present a heuristic argument for $K = 3$. Proving Proposition 4.5 is equivalent to proving that

$$\mathbb{P}(Z_1 > b_1 \cup Z_2 > b_2 \cup \cdots \cup Z_K > b_K) \leq \mathbb{P}(Z_1^* > b_1 \cup Z_2^* > b_2 \cup \cdots \cup Z_K^* > b_K).$$

First, note another representation for the above probabilities. For example, the probability on the left hand side can be written as

$$\begin{aligned}
&\mathbb{P}(Z_1 > b_1) + \cdots + \mathbb{P}(Z_K > b_K) \\
&- \mathbb{P}(Z_1 > b_1 \cap X_2 > b_2) - \cdots - \mathbb{P}(Z_{K_1} > b_{K-1} \cap Z_K > b_K) \\
&\vdots \\
&+ \mathbb{P}(Z_1 > b_1 \cap \cdots \cap Z_K > b_K).
\end{aligned} \tag{4.46}$$

Theorem 4.6. *Proposition 4.5 holds for the case K=2.*

Proof. By the formulation in equation (4.46), the problem is equivalent to proving that

$$\begin{aligned}
&\mathbb{P}(Z_1 > b_1) + \mathbb{P}(Z_1 > b_2) - \mathbb{P}(Z_1 > b_1 \cap Z_2 > b_2) \leq \mathbb{P}(Z_1 > b_1 \cap Z_2 > b_2) \\
\leq &\mathbb{P}(Z_1^* > b_1) + \mathbb{P}(Z_1^* > b_2) - \mathbb{P}(Z_1^* > b_1 \cap Z_2^* > b_2) \leq \mathbb{P}(Z_1 > b_1 \cap Z_2 > b_2).
\end{aligned}$$

For simplicity, the subscripts of ρ_{12} and ρ_{12}^* are dropped so $\rho = Corr(Z_1, Z_2)$ and $\rho^* = Corr(Z_1^*, Z_2^*)$. The marginal distributions of Z_1, Z_2, Z_1^* and Z_2^* are equivalent and are all $N(0, 1)$ random variables and hence the probabilities are such that $\mathbb{P}(Z_k > b_k) = \mathbb{P}(Z_k^* > b_k)$ for each $k = 1, 2$. Therefore, the problem is reduced to showing that

$$\mathbb{P}(Z_1^* > b_1 \cap Z_2^* > b_2) \leq \mathbb{P}(Z_1 > b_1 \cap Z_2 > b_2). \tag{4.47}$$

when $\rho^* \leq \rho$.

In the below calculations, we appeal to the fact that for two random variables which are bivariate normally distributed, the conditional distribution of one normal random variable on the other normal random variable is also normal. Specifically, we have that $Z_2|Z_1 = z_1 \sim N(\rho z_1, 1 - \rho^2)$. Let $\phi(\cdot)$ and $\Phi(\cdot)$ denote the probability density function and cumulative distribution function of a standard normal random variable respectively, then the probability on the left hand side in Equation (4.47)

is

$$\begin{aligned}\mathbb{P}(Z_1 > b_1 \cap Z_2 > b_2) &= \mathbb{P}(Z_2 > b_2 | Z_1 > b_1)\mathbb{P}(Z_1 > b_2)\\ &= \int_{b_1}^{\infty} \mathbb{P}(Z_2 > b_2 | Z_1 = z_1)\phi(z_1)dz_1\\ &= \int_{b_1}^{\infty} \left[1 - \Phi\left(\frac{b_2 - \rho z_1}{\sqrt{1-\rho^2}}\right)\right]\phi(z_1)dz_1.\end{aligned}$$

The corresponding calculation where Z_1^* and Z_2^* replace Z_1 and Z_2 yields the following

$$\mathbb{P}(Z_1^* > b_1 \cap Z_2^* > b_2) = \int_{b_1}^{\infty} \left[1 - \Phi\left(\frac{b_2 - \rho^* z_1}{\sqrt{1-\rho^{*2}}}\right)\right]\phi(z_1)dz_1$$

and since $\Phi(\cdot)$ is strictly increasing, it suffices to show that whenever $z_1 > b_1$, then

$$\frac{b_2 - \rho z_1}{\sqrt{1-\rho^2}} \leq \frac{b_2 - \rho^* z_1}{\sqrt{1-\rho^{*2}}}. \tag{4.48}$$

We have by assumption that $\rho^* \leq \rho$. Further note that by definition $0 \leq \rho \leq 1$ and $0 \leq \rho^* \leq 1$. The following steps show some simple algebraic manipulation of this inequality, which gives

$$\begin{aligned}\rho^* \leq \rho &\iff \frac{1+\rho^*}{1-\rho^*} \leq \frac{1+\rho}{1-\rho}\\ &\iff \frac{\sqrt{1-\rho^{*2}}}{1-\rho^*} \leq \frac{\sqrt{1-\rho^2}}{1-\rho}\\ &\iff \frac{\sqrt{1-\rho^{*2}} - \sqrt{1-\rho^2}}{\rho\sqrt{1-\rho^{*2}} - \rho^*\sqrt{1-\rho^2}} \leq 1.\end{aligned}$$

Finally, using the above inequality and Conditions 4.3 that $0 \leq b_2 \leq b_1$, we have for $z_1 \geq b_1$ that

$$b_2\frac{\sqrt{1-\rho^{*2}} - \sqrt{1-\rho^2}}{\rho\sqrt{1-\rho^{*2}} - \rho^*\sqrt{1-\rho^2}} \leq b_2 \leq b_1 \leq z_1$$

and a simple rearrangement shows that Equation (4.48) is satisfied. □

The result for $K = 2$ covers many trials since just 1 interim analysis is common in practice. We now give a heuristic argument for the proof of this theorem for the case $K = 3$. We follow a similar approach to the proof of the case $K = 2$. The marginal distributions are equivalent, that is we have that $\mathbb{P}(Z_k > b_k) = \mathbb{P}(Z_k^* > b_k)$ for each $k = 1, 2, 3$. Let the remaining probabilities of Equation (4.46) be summarised by the functions $f(\cdot)$ and $g(\cdot)$, which are given by

$$\begin{aligned} f(\rho_{12},\rho_{13},\rho_{23}) &= f_{12}(\rho_{12})+f_{13}(\rho_{13})+f_{23}(\rho_{23}) \\ &= \mathbb{P}(Z_1>b_1\cap Z_2>b_2)+\mathbb{P}(Z_1>b_1\cap Z_3>b_3)+\mathbb{P}(Z_2>b_2\cap Z_3>b_3) \\ g(\rho_{12},\rho_{13},\rho_{23}) &= \mathbb{P}(Z_1>b_1\cap Z_2>b_2\cap Z_3>b_3). \end{aligned}$$

Then we need to show that

$$g(\rho_{12},\rho_{13},\rho_{23})-g(\rho_{12}^*,\rho_{13}^*,\rho_{23}^*)\leq f(\rho_{12},\rho_{13},\rho_{23})-f(\rho_{12}^*,\rho_{13}^*,\rho_{23}^*). \tag{4.49}$$

The right had side of Equation (4.49) is greater than or equal to zero. This can be seen by a similar argument as for the case $K=2$, it is clear that

$$\begin{aligned} \mathbb{P}(Z_1^*>b_1\cap Z_2^*>b_2)&\leq\mathbb{P}(Z_1>b_1\cap Z_2>b_2) \\ \mathbb{P}Z_1^*>b_1\cap Z_3^*>b_3)&\leq\mathbb{P}(Z_1>b_1\cap Z_3>b_3) \\ \mathbb{P}(Z_2^*>b_2\cap Z_3^*>b_3)&\leq\mathbb{P}(Z_2>b_2\cap Z_3>b_3) \end{aligned}$$

and this implies that $f(\rho_{12}^*,\rho_{13}^*,\rho_{23}^*)\leq f(\rho_{12},\rho_{13},\rho_{23})$. Suppose that $g(\rho_{12}^*,\rho_{13}^*,\rho_{23}^*)>g(\rho_{12},\rho_{13},\rho_{23})$, then Equation (4.49) holds so it is sufficient to consider the case $g(\rho_{12}^*,\rho_{13}^*,\rho_{23}^*)\leq g(\rho_{12},\rho_{13},\rho_{23})$.

Both sides of Equation (4.49) are greater than or equal to zero, and we know by Conditions 4.3 that $\rho_{12}^*\leq\rho_{12},\rho_{13}^*\leq\rho_{13}$ and $\rho_{23}^*\leq\rho_{23}$. For Equation (4.49) to hold, we need that the function $f(\cdot)$ is increasing at a greater rate than $g(\cdot)$ as each parameter ρ_{12},ρ_{13} and ρ_{23} increases.

We shall check that Equation (4.49) holds graphically for each parameter ρ_{12},ρ_{13} and ρ_{23} individually. First note that $f(\rho_{12},\rho_{13},\rho_{23})$ only depends on ρ_{12} through the function $f_{12}(\rho_{12})=\mathbb{P}(Z_1>b_1\cap Z_2>b_2)$ and so plotting this probability as a function of ρ_{12} is sufficient for checking the rate of change of $f(\rho_{12},\rho_{13},\rho_{23})$ with respect to ρ_{12}. We cannot check this for every combination of the parameters and so we follow the steps below to check the condition in a systematic way.

- Choose values of $\rho_{12},\rho_{13},\rho_{23},\rho_{12}^*,\rho_{13}^*$ and ρ_{23}^*
- Check that $f(\rho_{12},\rho_{13},\rho_{23})$ increases in ρ_{12} at a greater rate than $g(\rho_{12},\rho_{13},\rho_{23})$. Do this by plotting $f_{12}(\rho_{12})$ and $g(\rho_{12},\rho_{13},\rho_{23})$ against ρ_{12}
- Check that $f(\rho_{12}^*,\rho_{13},\rho_{23})$ increases in ρ_{13} at a greater rate than $g(\rho_{12}^*,\rho_{13},\rho_{23})$. Do this by plotting $f_{13}(\rho_{13})$ and $g(\rho_{12}^*,\rho_{13},\rho_{23})$ against ρ_{13}
- Check that $f(\rho_{12}^*,\rho_{13}^*,\rho_{23})$ increases in ρ_{23} at a greater rate than $g(\rho_{12}^*,\rho_{13}^*,\rho_{23})$.

Do this by plotting $f_{23}(\rho_{23})$ and $g(\rho^*_{12}, \rho^*_{13}, \rho_{23})$ against ρ_{23}.

We shall consider three combinations of the parameters $\rho_{12}, \rho_{13}, \rho_{23}, \rho^*_{12}, \rho^*_{13}$ and ρ^*_{23}. The first scenario is based on equally spaced information levels and the ρ parameters are divided by a factor of 1.2 to get the ρ^* parameters. The second scenario is where the two interim analyses have higher information levels (than if information was equally spaced) and the ρ parameters are divided by 1.1. The final scenario is where the two interim analyses have lower information levels (than if information was equally spaced) and the ρ parameters are divided by 1.3. The parameter values for the three cases are presented in Table 4.15.

	$\rho_{12}, \rho_{13}, \rho_{23}$	$\rho^*_{12}, \rho^*_{13}, \rho^*_{23}$
Case 1	$0.849, 0.693, 0.980$	$0.707, 0.577, 0.816$
Case 2	$0.850, 0.765, 0.990$	$0.772, 0.695, 0.900$
Case 3	$0.751, 0.531, 0.919$	$0.577, 0.408, 0.707$

Table 4.15: Parameter choices for three cases to compare rate of change of objects $\mathbb{P}(Z_{k_1} > b_{k_1} \cap Z_{k_2} > b_{k_2})$ and $\mathbb{P}(Z_{k_1} > b_{k_1} \cap Z_{k_2} > b_{k_2} \cap Z_{k_3} > b_{k_3})$ for a group sequential trial with $K = 3$ analyses .

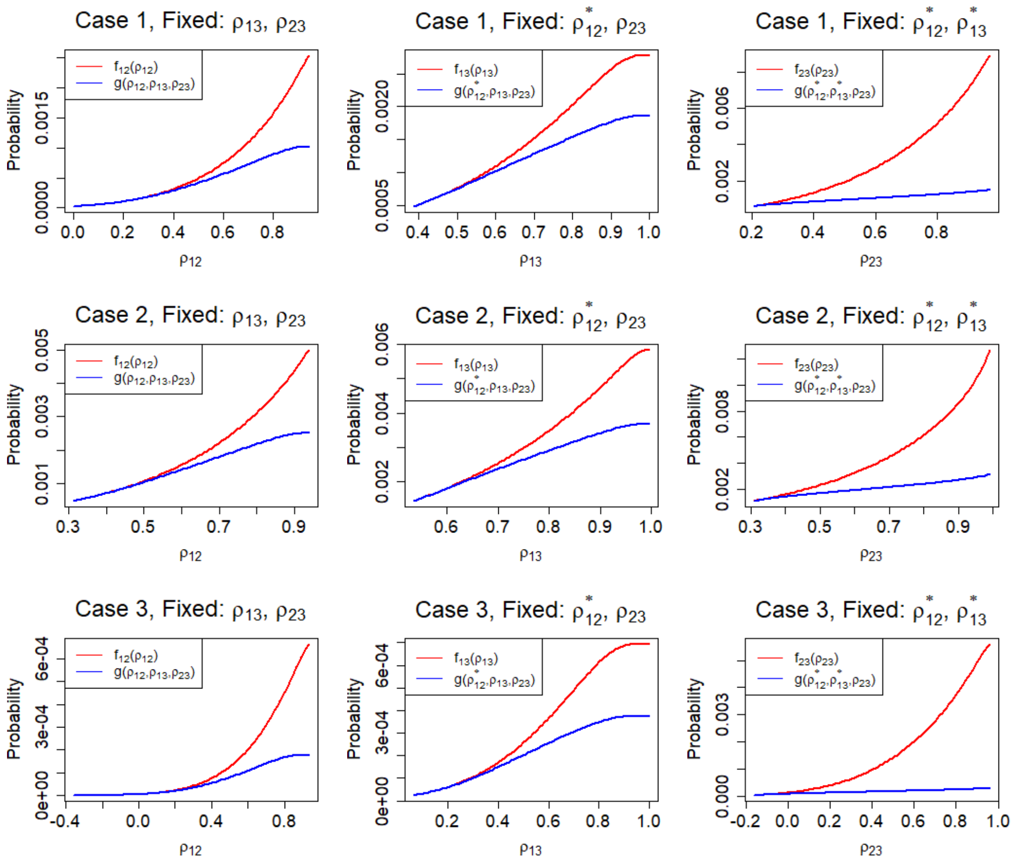


Figure 4.8: Comparison of rate of change of $f(\rho_{12}, \rho_{13}, \rho_{23})$ and $g(\rho_{12}, \rho_{13}, \rho_{23})$ for a group sequential trial with $K = 3$ analyses. Fixed values of $\rho_{12}, \rho_{13}, \rho_{23}, \rho^*_{12}, \rho^*_{13}$ and ρ^*_{23} given by Table 4.15.

Figure 4.8 compares the pairwise probabilities with the probability of crossing all three boundaries for each of the three cases in Table 4.15. In cases 1 and 3, as the parameter $\rho_{13} \to 1$, it is not clear whether $f_{13}(\rho_{13})$ increases at a greater rate than $g(\rho^*_{12}, \rho_{13}, \rho_{23})$. We have checked this numerically and found it to be true. Further the case $\rho_{13} = 1$ is the very rare case where all three analyses have the same information levels. This is not a scenario of concern. Therefore, it is clear that the function $f(\rho_{12}, \rho_{13}, \rho_{23})$ increases at a greater rate than the function $g(\rho_{12}, \rho_{13}, \rho_{23})$ in each parameter ρ_{12}, ρ_{13} and ρ_{23} and hence we have shown convincing evidence to prove that Equation (4.49) holds.

At the interim analyses of a group sequential test, estimates of the correlations $\rho_{12}, \rho_{13}, \rho_{23}, \rho^*_{12}, \rho^*_{13}$ and ρ^*_{23} become available. Hence, this check can be repeated as we learn about actual parameter values. In the rare event that the results are not in the right direction, the second method for performing a group sequential test when the canonical joint distribution does not hold should be employed. The three cases given in Table 4.15 represent the cases where information levels are equally spaced, information accruing earlier than equally spaced and also information accruing later than equally spaced and therefore these presented cases are representative of all possible scenarios.

4.4.4 | Method 2: Use an estimate of the true covariance structure

The second method for dealing with estimates from the joint model does not rely on the canonical joint distribution assumption. Instead, we calculate the group sequential boundaries using the complete structure of the variance-covariance matrix for the sequence of treatment effect estimates across analyses. This differs from when the canonical joint distribution is assumed because in such a case, only the variances are required and the covariances are ignored. We shall see that this method poses some practical difficulties; for example, a trial may yield a non-positive-definite estimate for the variance-covariance matrix and hence, calculation of the boundaries cannot be performed.

Suppose that a group sequential trial with K analyses yields the sequence of treatment effect estimates $\hat{\theta}_1, \dots, \hat{\theta}_K$. The distribution of the sequence of treatment effect estimates is

$$(\hat{\theta}_1, \dots, \hat{\theta}_K)^T \sim N((\theta, \dots, \theta)^T, \Sigma).$$

To estimate the variance-covariance matrix Σ, the matrices $\hat{A}^{(1)}, \dots, \hat{A}^{(K)}$ and $\hat{B}^{(1)}, \dots, \hat{B}^{(K)}$ are calculate using Equations (4.38) and (4.39), then elements of the estimate $\hat{\Sigma}$ are given by

$$\begin{aligned} \hat{\Sigma}_{kk} &= \left[(\hat{A}^{(k)})^{-1}\hat{B}^{(k)}((\hat{A}^{(k)})^{-1})^T\right]_{22} \qquad && \text{for } k = 1, \dots, K \\ \hat{\Sigma}_{k_1 k_2} &= \left[(\hat{A}^{(k_1)})^{-1}\hat{B}^{(k_1)}((\hat{A}^{(k_2)})^{-1})^T\right]_{22} \qquad && k_1 < k_2. \end{aligned}$$

The information levels for $\hat{\theta}_1, \dots, \hat{\theta}_K$ are given by $\mathcal{I}_k = 1/\hat{\Sigma}_{kk}$ which is the same as in method 1. Under H_0, the amount of type 1 error spent at analysis k is $\alpha^{(k)} = \mathbb{P}_{\theta=0}(\text{Continue to analysis } k \text{ and cross the upper boundary at analysis } k)$. Under H_A, when $\theta = \delta$, the amount of type 2 error spent at analysis k is $\beta^{(k)} = \mathbb{P}_{\theta=\delta}(\text{Continue to analysis } k \text{ and cross the upper lower boundary at analysis } k)$. Using error spending functions $f(t) = \min\{\alpha t^2, \alpha\}$ and $g(t) = \min\{\beta t^2, \beta\}$, at analysis k, we design the trial with

$$\begin{aligned}
\alpha^{(1)} &= f(\mathcal{I}_1/\mathcal{I}_{max}) \\
\beta^{(1)} &= g(\mathcal{I}_1/\mathcal{I}_{max}) \\
\alpha^{(k)} &= f(\mathcal{I}_k/\mathcal{I}_{max}) - f(\mathcal{I}_{k-1}/\mathcal{I}_{max}) \qquad \text{for } k = 2, \dots, K \\
\beta^{(k)} &= g(\mathcal{I}_k/\mathcal{I}_{max}) - g(\mathcal{I}_{k-1}/\mathcal{I}_{max}) \qquad \text{for } k = 2, \dots, K.
\end{aligned}$$

For this method, we calculate the boundary values for the group sequential trial using the full structure of $\hat{\Sigma}$. Let $\hat{\Sigma}_k$ denote the $k \times k$ matrix that is the first k rows and columns of $\hat{\Sigma}$. The boundary points, $\tilde{a}_k$ and $\tilde{b}_k$ are therefore calculated as the solutions to the following equations:

$$\alpha^{(k)} = \int_{b_k}^{\infty} \int_{a_{k-1}}^{b_{k-1}} \cdots \int_{a_1}^{b_1} \frac{\exp\{-\frac{1}{2}\mathbf{x}^T \hat{\Sigma}_k^{-1} \mathbf{x}\}}{\sqrt{(2\pi)^k |\hat{\Sigma}_k|}} dx_1 \dots dx_k$$

$$\beta^{(k)} = \int_{-\infty}^{a_k} \int_{a_{k-1}}^{b_{k-1}} \cdots \int_{a_1}^{b_1} \frac{\exp\{-\frac{1}{2}(\mathbf{x} - (\delta, \dots, \delta)^T)^T \hat{\Sigma}_k^{-1} (\mathbf{x} - (\delta, \dots, \delta)^T)\}}{\sqrt{(2\pi)^k |\hat{\Sigma}_k|}} dx_1 \dots dx_k.$$

This integration calculation can be performed numerically using the R package mvtnorm by Genz et al. (2020). In an earlier paper, Genz (1992) describes the numerical algorithm and shows that for 10 variables or fewer, this calculation is computationally efficient.

It is also possible to calculate boundary constants based on the standardised statistics $Z_k = \hat{\theta}_k \sqrt{\mathcal{I}_k}$. The boundary points on the Z-scale are given by $a_k = \tilde{a}_k \sqrt{\mathcal{I}_k}$ and $b_k = \tilde{b}_k \sqrt{\mathcal{I}_k}$ for $k = 1, \dots, K$ where $\tilde{a}_k, \tilde{b}_k$ are boundary points for $\hat{\theta}_k$.

During the conditional score method, the matrix Σ is estimated with error which can sometimes result in a non positive-definite estimate $\hat{\Sigma}$. In particular, suppose that we have reached analysis 2 and find that $\hat{\Sigma}_2$ is not invertible, then the boundary calculations cannot be performed. We have found, through simulation, that problems do not occur after analysis 2. This is because if the matrix $\hat{\Sigma}_2$ is invertible, then $\hat{\Sigma}_3, \dots, \hat{\Sigma}_K$ are also likely to be invertible. Therefore, we shall only consider this problem at the second analysis. We have performed a check for when this computation is not possible due to the covariance matrix being non-invertible. We are unable to invert $\hat{\Sigma}_k$ when the determinant is less than or equal to 0. In Table 4.16 we have counted the number of times, out of 10^4 simulations, which have terminated because we have determinant $d = det(\hat{\Sigma}_2) \leq 0$. This is shown in the column headed "$d \leq 0$". For extremely noisy longitudinal data with $\sigma^2 = 100$ this problem occurs roughly 50% of the time. However, for small measurement error of the longitudinal data, when σ^2 is small, this problem occurs infrequently.

γ	σ^2	η= 0			η= −0.5		
		$d \leq 0$	$d > 0$	$\mathbb{P}$(reject H_0)	$d \leq 0$	$d > 0$	$\mathbb{P}$(reject H_0)
0	0	0	10000	0.0224	0	10000	0.8917
0	1	0	10000	0.024	4	9996	0.8924
0	10	0	10000	0.023	1	9999	0.8893
0	100	4914	5086	0.0225	4408	5592	0.8886
0.03	0	0	10000	0.0221	0	10000	0.891
0.03	1	0	10000	0.0246	0	10000	0.8912
0.03	10	0	10000	0.0235	0	10000	0.8946
0.03	100	3736	6264	0.0276	4306	5694	0.8915
0.06	0	0	10000	0.0221	0	10000	0.8936
0.06	1	0	10000	0.026	0	10000	0.8912
0.06	10	1	9999	0.0228	0	10000	0.8941
0.06	100	3595	6405	0.0264	4240	5760	0.8912
0.09	0	0	10000	0.023	0	10000	0.8931
0.09	1	0	10000	0.0229	0	10000	0.8837
0.09	10	0	10000	0.025	0	10000	0.8898
0.09	100	4042	5958	0.029	4416	5584	0.8723

Table 4.16: Method 2: Simulation results for paramater values $\gamma = 0, 0.03, 0.06, 0.09$ and $\sigma^2 = 0, 1, 10, 100$ using 10^4 replicates. Counts out of 10^4 cases of problematic simulations. $\mathbb{P}$(reject H_0) calculated as the proportion out of the counts $d > 0$.

We now investigate the error rates using a simulation study for a group sequential trial using method 2. We have performed 10^4 replicates of a trial each with sample size $n = 343$ and in each case we have found boundary points and analysed the trial using method 2 described above. Out of these 10^4 replicates, we take the proportion of replicates that reject H_0 out of the cases where $d = det(\hat{\Sigma}_2) > 0$. This is given by the column headed "$\mathbb{P}$(reject H_0)" in Table 4.16. For the rows of the table where the counts of $d > 0$ are equal to 10^4, the value of $\mathbb{P}$(reject H_0) is equal to the type 1 error when $\eta = 0$ or power when $\eta = -0.5$. We see that none of the simulation type 1 errors are significantly different from $\alpha = 0.025$. Further, all simulation type 2 errors are close to $1 - \beta = 0.9$. The sample size calculation in Section 4.3.3 resulting in $n = 343$ is only approximate here which explains the deviation from power $1 - \beta$.

In summary, this method makes no assumption about the covariance structure of the sequence of treatment effect estimates and therefore the type 1 error is preserved. However, this method is problematic for large longitudinal measurement error σ^2, where roughly 50% of trials have a non-invertible covariance matrix $\hat{\Sigma}_2$. The trial may have stopped at the first analysis before finding a problem at the second analysis and hence, we cannot correct for this method. Comparing this with method 1, the error in estimating the covariance matrix Σ creates problems for calculating the boundary points. The benefit that occurs in method 1 from not having to estimate covariances outweighs the small issue that the trial does not reach the planned significance level α.

4.4.5 | Method 3: Create a new efficient estimator

For the final method, we aim to create a new estimator that is asymptotically efficient. The efficient estimate at analysis k is a linear combination of the original estimates at analyses up to and including k. We choose the weights of the linear combination using a Lagrange multiplier method in such a way that the variance is minimised. We can easily prove theoretically that this new estimator has the correct canonical distribution, and hence the methods in Section 2.1 can be used without hesitation. However, we show that in practice there are limitations to this method as it relies too heavily on accurately estimating the covariance matrix of the parameter estimates.

Jennison and Turnbull (1997) prove a simple result, that all asymptotically efficient estimators have the canonical joint distribution and this is the motivation for this method. It is not intuitively obvious why taking linear combinations of past estimates results in asymptotic efficiency, however because the canonical joint distribution does not hold for our sequence of treatment effect estimates, there must exist a more efficient estimate. Using the conditional score method, we obtain estimates that are unbiased and asymptotically normally distributed and we appeal to the closure properties of the multivariate normal distribution. It remains that the most efficient estimate has the smallest variance and hence we seek to minimize the variance of this linear combination.

We present the Lagrange multiplier method that takes a sequence of parameter estimates $\hat{\theta}_1, \ldots, \hat{\theta}_K$ as input and returns a new sequence of estimates $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_K$. This new sequence of estimates will have the canonical distribution. At analysis k,

$\hat{\theta}_k^*$ will be a linear combination of $\hat{\theta}_1, \ldots, \hat{\theta}_k$, given by

$$\hat{\theta}_k^* = \sum_{i=1}^{k} c_i^{(k)} \hat{\theta}_i \tag{4.50}$$

where $c_1^{(k)}, \ldots, c_k^{(k)}$ are scalars which are yet to be determined. The values of these scalars are chosen to minimise the estimated variance of the new estimate, which by definition is given by

$$\begin{aligned}\widehat{Var}(\hat{\theta}_k^*) &= \sum_{i=1}^{k} \sum_{j=1}^{k} c_i^{(k)} c_j^{(k)} \widehat{Cov}(\hat{\theta}_i, \hat{\theta}_j) \\ &= \sum_{i=1}^{k} \sum_{j=1}^{k} c_i^{(k)} c_j^{(k)} \hat{\Sigma}_{ij}. \end{aligned} \tag{4.51}$$

Here, the matrix Σ is notation for the variance-covariance matrix of the vector $(\hat{\theta}_1, \ldots, \hat{\theta}_K)^T$ and $\hat{\Sigma}$ is an estimate of Σ which can be found using Equations (4.38) and (4.39).

The new sequence of estimates must satisfy all three conditions of the canonical distribution of definition 2.2. One property is that we must have $\mathbb{E}(\hat{\theta}_k^*) = \theta$ for each $k = 1, \ldots, K$. We can ensure this property holds by imposing the constraint that $c_1^{(k)} + \cdots + c_k^{(k)} = 1$ for each $k = 1, \ldots, K$. This is because expectation is linear and we have by Theorem 4.4 that $\mathbb{E}(\hat{\theta}_i) = \theta$ for all $i = 1, \ldots, K$. Therefore, the problem is to minimise $\widehat{Var}(\hat{\theta}_k^*)$ subject to the constraint $c_1^{(k)} + \cdots + c_k^{(k)} = 1$. The Lagrangian function is then

$$\begin{aligned}\mathcal{L}(\hat{\theta}_1, \ldots, \hat{\theta}_k, c_1^{(k)}, \ldots, c_k^{(k)}, \lambda_k) &= \sum_{i=1}^{k} \sum_{j=1}^{k} c_i^{(k)} c_j^{(k)} \hat{\Sigma}_{ij} + \lambda_k (c_1^{(k)} + \cdots + c_k^{(k)} - 1) \\ &= 2 \sum_{i=1}^{k} \sum_{j \neq i} c_i^{(k)} c_j^{(k)} \hat{\Sigma}_{ij} + \sum_{i=1}^{k} \left(c_i^{(k)} \right)^2 \hat{\Sigma}_{ii} + \lambda_k (c_1^{(k)} + \cdots + c_k^{(k)} - 1)\end{aligned}$$

where λ_k is the scalar Lagrange multiplier.

It remains to find the stationary points of $\mathcal{L}(\hat{\theta}_1, \ldots, \hat{\theta}_k, c_1^{(k)}, \ldots, c_k^{(k)}, \lambda_k)$ as a function of $c_1^{(k)}, \ldots, c_k^{(k)}$ and λ_k. The partial derivatives are given by

$$\begin{aligned}\frac{\partial}{\partial c_m^{(k)}} \mathcal{L}(\hat{\theta}_1, \ldots, \hat{\theta}_k, c_1^{(k)}, \ldots, c_k^{(k)}, \lambda_k) &= \sum_{i=1}^{k} 2 c_i^{(k)} \hat{\Sigma}_{mi} + \lambda_k \qquad \text{for } m = 1, \ldots, k \\ \frac{\partial}{\partial \lambda_k} \mathcal{L}(\hat{\theta}_1, \ldots, \hat{\theta}_k, c_1^{(k)}, \ldots, c_k^{(k)}, \lambda_k) &= c_1^{(k)} + \cdots + c_k^{(k)} - 1.\end{aligned}$$

Note that the function $\mathcal{L}(\cdot)$ is quadratic in $c_i^{(k)}$ and $\mathcal{L}(\cdot) \to \infty$ as any $c_i^{(k)} \to \infty$. Therefore, the resulting stationary points will give a minimum of $\mathcal{L}(\cdot)$.

Setting each of the partial derivatives equal to zero, we have a set of $k+1$ equations which are linear functions of $c_1^{(k)}, \dots, c_k^{(k)}$ and λ_k. In matrix form this is

$$\begin{bmatrix} 2\hat{\Sigma}_{11} & 2\hat{\Sigma}_{12} & \dots & 2\hat{\Sigma}_{1k} & 1 \\ 2\hat{\Sigma}_{21} & 2\hat{\Sigma}_{22} & & 2\hat{\Sigma}_{2k} & 1 \\ \vdots & & \ddots & & \vdots \\ 2\hat{\Sigma}_{k1} & 2\hat{\Sigma}_{k2} & & 2\hat{\Sigma}_{kk} & 1 \\ 1 & 1 & \dots & 1 & 0 \end{bmatrix} \begin{bmatrix} c_1^{(k)} \\ c_2^{(k)} \\ \vdots \\ c_k^{(k)} \\ \lambda_k \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}. \tag{4.52}$$

To find the values of the scalars $c_1^{(k)}, \dots, c_k^{(k)}$, one can then simply rearrange and solve Equation (4.52). The constants $c_1^{(k)}, \dots, c_k^{(k)}$ are then used in Equation (4.50) to find the efficient estimate at analysis k. This process is repeated for each $k = 1, \dots, K$ to build the sequence of estimates $\hat{\theta}_1^*, \dots, \hat{\theta}_K^*$.

The new sequence of efficient estimates $\hat{\theta}_1^*, \dots, \hat{\theta}_K^*$ is easily seen to have the canonical joint distribution of Definition 2.2. The first property follows by closure under linear combinations of multivariate normal random variables. That is, if $(\hat{\theta}_1, \dots, \hat{\theta}_K)$ is multivariate normal, then $\hat{\theta}_k^* = \sum_{i=1}^k c_i^{(k)} \hat{\theta}_i$ is normally distributed for each $k = 1, \dots, K$. Then, the joint distribution $(\hat{\theta}_1^*, \dots, \hat{\theta}_K^*)$ is multivariate normal. For the second property, it is clear that the estimate $\hat{\theta}^*$ is unbiased for θ for each $k = 1, \dots, K$. In the calculation below, the first line follows by linearity of expectation and the second line is because the estimates $\hat{\theta}_1, \dots, \hat{\theta}_K$ are each unbiased for θ. Then, the result holds because of the constraint that $c_1^{(k)} + \dots + c_k^{(k)} = 1$.

$$\begin{aligned} \mathbb{E}(\hat{\theta}_k^*) &= \sum_{i=1}^k c_i^{(k)} \mathbb{E}(\hat{\theta}_i) \\ &= \theta \sum_{i=1}^k c_i^{(k)} = \theta. \end{aligned}$$

The information matrix for the new estimate $\hat{\theta}_k^*$ is given by $\mathcal{I}_k^* = 1/\widehat{Var}(\hat{\theta}^*)$, and we define the second property as $\hat{\theta}_k^* \sim N(\theta, \mathcal{I}_k^*)$. The final property follows by construction of $\hat{\theta}_k^*$. We have designed the estimate to have minimum possible variance, which ensures that the estimates are asymptotically efficient for each $k = 1, \dots, K$. Jennison and Turnbull (1997) show that all asymptotically efficient estimates have the covariance structure in the canonical joint distribution. By these arguments, we have that $\widehat{Cov}(\hat{\theta}_{k_1}^*, \hat{\theta}_{k_2}^*) = \widehat{Var}(\hat{\theta}_{k_2}^*) = (\mathcal{I}_{k_2}^*)^{-1}$ as required.

There are some numerical problems that can occur with this approach which must be considered; in some extreme cases, the new estimate obtained by the process described above has $\sum_{i=1}^{n}\sum_{j=1}^{n} c_i^{(k)} c_j^{(k)} \hat{\Sigma}_{ij} < 0$ and $\widehat{Var}(\hat{\theta}_k^*)$ appears to be negative.

We shall consider how this problem occurs through use of an example with $K = 2$ analyses. Estimation of the covariance matrix is data dependent, so there are many sequences $\hat{\Sigma}_{11}, \hat{\Sigma}_{12}, \hat{\Sigma}_{22}$ that are possible. We only consider cases such that $\hat{\Sigma}_{11} \geq \hat{\Sigma}_{12} \geq \hat{\Sigma}_{22}$ as other sequences are extremely rare and have not occurred in our simulations. The case $\hat{\Sigma}_{12} = \hat{\Sigma}_{22}$ means that the estimates are asymptotically efficient, and so we consider the affect as $\hat{\Sigma}_{12}$ becomes large and close to $\hat{\Sigma}_{11}$. To do so, we consider the parameterisation

$$\hat{\Sigma}_{12} = \hat{\Sigma}_{11}^{p} \hat{\Sigma}_{22}^{1-p} \tag{4.53}$$

and assess the variance as a function of p. We can calculate the parameter p as

$$p = \frac{\log(\hat{\Sigma}_{12}) - \log(\hat{\Sigma}_{22})}{\log(\hat{\Sigma}_{11}) - \log(\hat{\Sigma}_{22})}.$$

It is clear that for the first analysis, we always have that $c_1^{(1)} = 1$ and hence, $\hat{\theta}_1^* = \hat{\theta}_1$. For the second analysis, we shall temporarily drop the superscript notation for simplicity so that c_i replaces $c_i^{(2)}$. The constraint here is then $c_1 + c_2 = 1$. Therefore, using some simple algebraic manipulation and equation (4.51), the variance for the new estimate, in terms of c_2, is given by

$$\widehat{Var}(\hat{\theta}_2^*) = (\hat{\Sigma}_{11} - 2\hat{\Sigma}_{12} + \hat{\Sigma}_{22})c_2^2 + 2(\hat{\Sigma}_{12} - \hat{\Sigma}_{11})c_2 + \hat{\Sigma}_{11}.$$

The roots of this quadratic equation in c_2 are

$$\begin{aligned} x_1 &= \frac{\hat{\Sigma}_{11} - \hat{\Sigma}_{12} + \sqrt{\hat{\Sigma}_{12}^2 - \hat{\Sigma}_{11}\hat{\Sigma}_{22}}}{\hat{\Sigma}_{11} - 2\hat{\Sigma}_{12} + \hat{\Sigma}_{22}} \\ x_2 &= \frac{\hat{\Sigma}_{11} - \hat{\Sigma}_{12} - \sqrt{\hat{\Sigma}_{12}^2 - \hat{\Sigma}_{11}\hat{\Sigma}_{22}}}{\hat{\Sigma}_{11} - 2\hat{\Sigma}_{12} + \hat{\Sigma}_{22}}. \end{aligned}$$

Using the constraints $\hat{\Sigma}_{11} \geq \hat{\Sigma}_{12} \geq \hat{\Sigma}_{22}$, the function $\widehat{Var}(\hat{\theta}_2^*)$ is always a positive quadratic function of c_2. Further, by assessing where the roots x_1 and x_2 take positive and real values, and using Equation (4.53), we find that

$$\begin{aligned}\widehat{Var}(\hat{\theta}_2^*) \geq 0 &\iff \hat{\Sigma}_{11}^{2p}\hat{\Sigma}_{22}^{2-2p} \leq \hat{\Sigma}_{11}\hat{\Sigma}_{22} \\ &\iff p \leq \frac{1}{2} \\ \widehat{Var}(\hat{\theta}_2^*) < 0 &\iff p > \frac{1}{2}.\end{aligned}$$

We find that, in practice, every problem occurs at the second analysis, which is the first opportunity to find an estimate $\hat{\Sigma}_{12}$ for the covariance Σ_{12}. This is likely to be because of the estimation error in $\hat{\Sigma}_{11}$ and $\hat{\Sigma}_{12}$ due to the small number of observed events at the first interim analysis.

The columns headed "$p > 0.5$" of Table 4.17 show the counts out of 10^4 simulations where method 3 is problematic because we obtain $\widehat{Var}(\hat{\theta}_2^*) < 0$. These counts are each for a group sequential trial with $K = 5$ analyses, where the problems all occur at the second analysis. The table shows that problems occur for $\sigma^2 = 100$. This finding coincides with Tables 4.7 and 4.9 which show that for large σ^2, the matrix $(\hat{A}^{(1)})^{-1}\hat{B}^{(1)}$ is far from the identity matrix.

We would also like to avoid implausible situations where the variance is positive but close to zero. We see that as $p \uparrow \frac{1}{2}$ then $\widehat{Var}(\hat{\theta}_2^*) \downarrow 0$. Figure 4.9 shows an example of the boundaries calculated in a group sequential with $K = 5$ analyses when the value of p is very close to 0.5. The information levels, before the Lagrange multiplier method is applied, and the information levels after correction, are

- $(\mathcal{I}_1, \ldots, \mathcal{I}_5) = (1.77, 11.04, 16.83, 23.98, 35.12)$
- $(\mathcal{I}_1^*, \ldots, \mathcal{I}_5^*) = (1.77, 63.09, 152.04, 399.07, 502.63)$.

The estimated covariance of the treatment effect estimates between the first two analyses was $\hat{\Sigma}_{12} = 0.224$ which results in a value $p = 0.495$. We have used an error spending design and because $\mathcal{I}_2^* > \mathcal{I}_{max}$, the trial stops and concludes at the second analysis. Figure 4.9 shows the boundaries for this group sequential trial. Black lines show the boundaries based on information levels $\mathcal{I}_1, \ldots, \mathcal{I}_K$ and red lines give the boundaries using information levels $\mathcal{I}_1^*, \ldots, \mathcal{I}_K^*$.

We have found by inspection, that for values $0.49 < p < 0.5$, the Lagrange multiplier method creates estimates with implausibly small variance estimates. For values $p \leq 0.49$, this method is a reasonable solution to the problem that the canonical joint distribution does not hold. The columns headed "$0.49 < p \leq 0.5$" of Table 4.17 show the counts out of of 10^4 simulations where method 3 is problematic because it creates an estimate with implausibly small positive variance.

$\boldsymbol{\gamma}$	$\boldsymbol{\sigma^2}$	$\boldsymbol{\eta}$= **0**				$\boldsymbol{\eta}$= **−0.5**			
		$\mathbf{p \leq 0.49}$	$\mathbf{0.49 <}$ $\mathbf{p \leq 0.5}$	$\mathbf{p > 0.5}$	$\mathbb{P}$**(reject** $\mathbf{H_0}$)	$\mathbf{p \leq 0.49}$	$\mathbf{0.49 <}$ $\mathbf{p \leq 0.5}$	$\mathbf{p > 0.5}$	$\mathbb{P}$**(reject** $\mathbf{H_0}$)
0	0	10000	0	0	0.0278	10000	0	0	0.8895
0	1	10000	0	0	0.0280	9996	0	4	0.8910
0	10	10000	0	0	0.0289	9999	0	1	0.8893
0	100	9807	3	190	0.0240	9776	4	220	0.8876
0.03	0	10000	0	0	0.0239	10000	0	0	0.8913
0.03	1	10000	0	0	0.0261	10000	0	0	0.8882
0.03	10	10000	0	0	0.0254	10000	0	0	0.8955
0.03	100	9805	1	194	0.0278	9805	2	193	0.8911
0.06	0	10000	0	0	0.0227	10000	0	0	0.8930
0.06	1	10000	0	0	0.0270	10000	0	0	0.8895
0.06	10	10000	0	0	0.0231	10000	0	0	0.8925
0.06	100	9816	1	183	0.0270	9851	3	146	0.8910
0.09	0	10000	0	0	0.0231	10000	0	0	0.8929
0.09	1	10000	0	0	0.0233	10000	0	0	0.8836
0.09	10	10000	0	0	0.0258	10000	0	0	0.8916
0.09	100	9815	4	181	0.0295	9843	3	154	0.8716

Table 4.17: Method 3: Simulation results for parameter values $\gamma = 0, 0.03, 0.06, 0.09$ and $\sigma^2 = 0, 1, 10, 100$ using 10^4 replicates. Counts out of 10^4 cases of problematic simulations. $\mathbb{P}$(reject H_0) calculated as the proportion out of the counts $p \leq 0.49$.

Table 4.17 shows the simulation results from using method 3 for a clinical trial where the canonical joint distribution does not hold. All parameter values are chosen as in Section 4.3.1. We have performed 10^4 replicates of a trial each with sample size $n = 343$ which is calculated using methods described in Section 4.3.3 and in each case we have found boundary points and analysed the trial using method 3 described above. Out of these 10^4 replicates, we take the proportion of replicates that reject H_0 out of the cases where $p \leq 0.49$. This is given by the column headed "$\mathbb{P}$(reject H_0)" in Table 4.16. For the rows of the table where the counts of $p \leq 0.49$ are equal to 10^4, the value of $\mathbb{P}$(reject H_0) is equal to the type 1 error when $\eta = 0$ or power when $\eta = -0.5$. There are some signs in this method that type 1 error is greater than 0.025. Major problems can occur from the inaccuracy in $\widehat{Var}(\hat{\theta}_1)$ and $\widehat{Cov}(\hat{\theta}_1, \hat{\theta}_2)$ and this casts doubts on the advisability of using these estimates for the method 3 construction, even when you don't see a negative variance estimate.

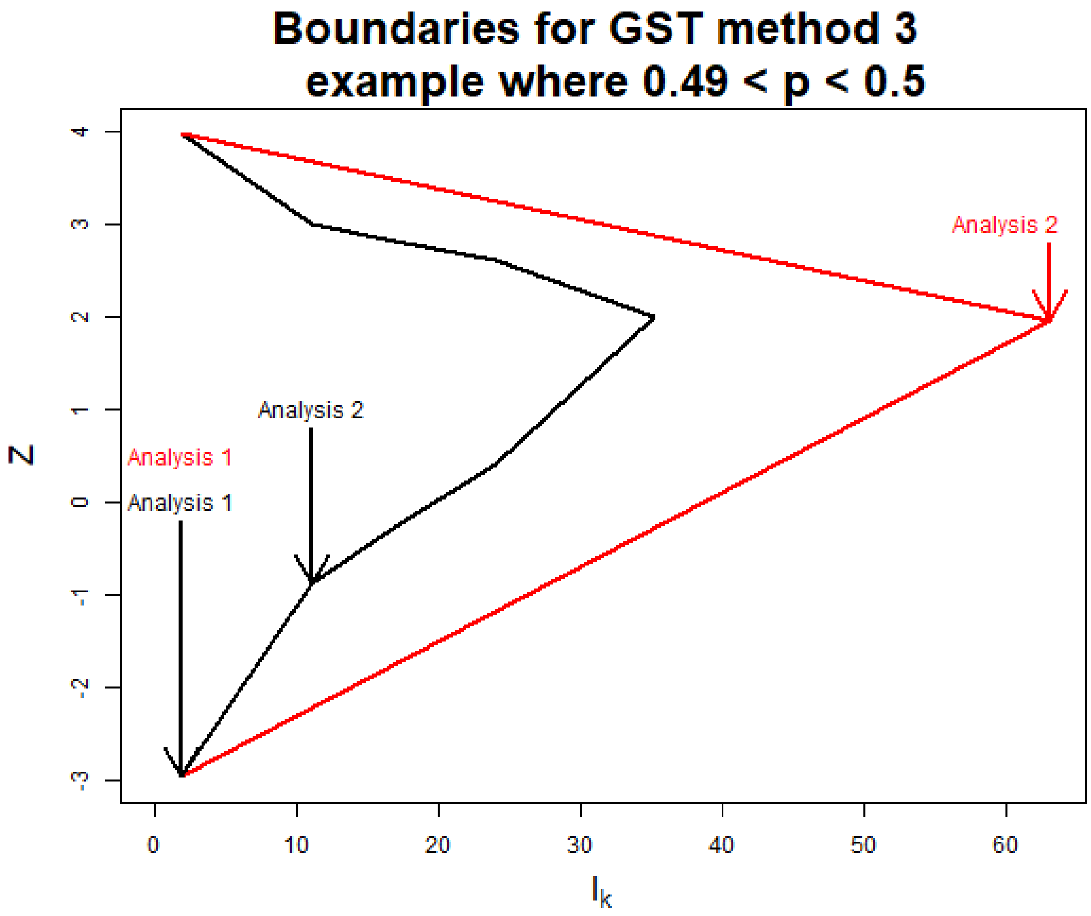


Figure 4.9: Comparison of boundaries for methods 1 and 3 in a group sequential trial with $K = 5$ analyses when $0.49 < p < 0.5$. Black lines are the boundaries from method 1 and red lines are boundaries from method 3.

4.4.6 | Comparison

We have presented three methods for creating a group sequential trial when the canonical joint distribution does not hold. We have seen that each method results in type 1 error close to the desired significance level α and (without accurately identifying a sample size) power $1 - \beta$. We now aim to compare these methods by calculating the amount of error spent at each analysis and see whether this matches the error spending designs.

We consider how the boundaries differ between designs. For each method we determine estimates $\hat{\theta}_1, \ldots, \hat{\theta}_K$ and information levels $\mathcal{I}_1, \ldots, \mathcal{I}_K$. Under H_0, the amount of type 1 error spent at analysis k is

$$\alpha^{(k)} = \mathbb{P}_{\theta=0}(\text{Continue to analysis } k \text{ and cross the upper boundary at analysis } k).$$

Under H_A, when $\theta = \delta$, the amount of type 2 error spent at analysis k is

$\beta^{(k)} = \mathbb{P}_{\theta=\delta}$(Continue to analysis k and cross the upper lower boundary at analysis k).

Using error spending functions $f(t) = \min\{\alpha t^2, \alpha\}$ and $g(t) = \min\{\beta t^2, \beta\}$, at analysis k, we design the trial with

$$\begin{aligned}
\alpha^{(1)} &= f(\mathcal{I}_1/\mathcal{I}_{max}) \\
\beta^{(1)} &= g(\mathcal{I}_1/\mathcal{I}_{max}) \\
\alpha^{(k)} &= f(\mathcal{I}_k/\mathcal{I}_{max}) - f(\mathcal{I}_{k-1}/\mathcal{I}_{max}) \qquad \text{for } k = 2, \dots, K \\
\beta^{(k)} &= g(\mathcal{I}_k/\mathcal{I}_{max}) - g(\mathcal{I}_{k-1}/\mathcal{I}_{max}) \qquad \text{for } k = 2, \dots, K.
\end{aligned}$$

The information levels are calculated in the same way for method 1 and method 2. Therefore, the same values are passed into the error spending functions and both methods aim to spend the same amount of error at each analysis. The location of the boundary points will be different. The information levels are calculated differently for methods 1 and 3, which affects the boundary positions.

Tables 4.18 and 4.19 show the amount of error that is spent at each analysis. For 10^4 replicates, we simulate a data set with $n = 343$ patients, and can calculate $\alpha^{(1)}, \dots, \alpha^{(K)}$ for each replicate. The average values of these over all the simulations is shown in the column headed "$\mathbb{E}(\alpha^{(k)})$" in Table 4.18. The columns headed "Simulation" gives the proportion of trials which stop to reject H_0 at analysis k. Similarly in Table 4.19, the columns headed "$\mathbb{E}(\beta^{(k)})$" give the average amount of type 2 error that is planned to be spent at analysis k. The parameter values here are $\gamma = 0.03, \sigma^2 = 1$ and the two tables show 10^4 Monte Carlo simulation study results calculated using $\eta = 0$ and $\eta = -0.5$ respectively.

Analysis	Method 1		Method 2		Method 3	
	Simulation	$\mathbb{E}(\boldsymbol{\alpha}^{(\mathbf{k})})$	**Simulation**	$\mathbb{E}(\boldsymbol{\alpha}^{(\mathbf{k})})$	**Simulation**	$\mathbb{E}(\boldsymbol{\alpha}^{(\mathbf{k})})$
1	0.0003	0.0012	0.0003	0.0012	0.0003	0.0012
2	0.0034	0.0038	0.0034	0.0038	0.0048	0.0040
3	0.0050	0.0064	0.0050	0.0064	0.0057	0.0065
4	0.0097	0.0087	0.0096	0.0087	0.0093	0.0088
5	0.0063	0.0050	0.0063	0.0050	0.0060	0.0046
Total	0.0247	0.0250	0.0246	0.0250	0.0261	0.0250

Table 4.18: Comparison of efficiency correction methods under the null hypothesis $\eta = 0$ for paramater values $\gamma = 0.03$ and $\sigma^2 = 1$. Probability of crossing each boundary in a group sequential trial with $K = 5$ analyses according to simulation and expected probability of error spending test.

Analysis	Method 1		Method 2		Method 3	
	Simulation	$\mathbb{E}(\boldsymbol{\beta}^{(\mathbf{k})})$	**Simulation**	$\mathbb{E}(\boldsymbol{\beta}^{(\mathbf{k})})$	**Simulation**	$\mathbb{E}(\boldsymbol{\beta}^{(\mathbf{k})})$
1	0.0025	0.0028	0.0025	0.0028	0.0025	0.0028
2	0.0112	0.0095	0.0114	0.0095	0.0146	0.0099
3	0.0215	0.0169	0.0214	0.0169	0.0216	0.0172
4	0.0243	0.0247	0.0244	0.0247	0.0257	0.0249

Table 4.19: Comparison of efficiency correction methods under the alternative hypothesis $\eta = -0.5$ for parameter values $\gamma = 0.03$ and $\sigma^2 = 1$. Probability of crossing each boundary in a group sequential trial with $K = 5$ analyses according to simulation and expected probability of error spending test.

The tables show that each of the three methods perform adequately with respect to the probability of crossing each boundary. The number of times each boundary is crossed closely matches the desired amount of error to be spent for each boundary. We have not included analysis 5 in Table 4.19 because it is common that the trials either over-run or under-run and therefore the boundary is not placed in the correct position for power to equal $1 - \beta$.

In method 1, we have shown that type 1 error is almost exactly equal to the planned significance level α as the large sample results in Section 4.4.1 show that $Cov(\hat{\theta}^{(k_1)}, \hat{\theta}^{(k_2)})$ is very close to $Var(\hat{\theta}^{(k_2)})$ for $k_1 < k_2$. The simulation results for method 1 show that the impact of the difference on type 1 error and power is extremely small. Further, we showed theoretically that for trials with a non-binding futility bounday, when there are differences in type 1 error and planned significance α, the test is conservative. This is supported by simulation studies of the whole process for moderately sized trials.

Given that $Cov(\hat{\theta}^{(k_1)}, \hat{\theta}^{(k_2)})$ is very close to $Var(\hat{\theta}^{(k_2)})$ for $k_1 < k_2$, there is not a lot to be gained by the more complex methods 2 and 3. The calculations are sensitive to errors in estimates of $Cov(\hat{\theta}^{(k_1)}, \hat{\theta}^{(k_2)})$ and $Var(\hat{\theta}^{(k_2)})$. In some cases, the errors in these calculations lead to these methods simply not working. This raises doubts about how well they work in less extreme cases. Despite these problems, methods 2 and 3 perform adequately in simulation studies. However, this does not change our view that method 1 is preferable.

4.5 | Efficiency gain from including longitudinal data in the joint model

We aim to assess the efficiency gain when the longitudinal data is included in the analysis compared to when this longitudinal data is available, yet ignored. In this case, we believe that our joint model is true and therefore, we shall simulate clinical trial data from the true joint model and analyse it in two separate ways. The first way is to fit the data to the joint model using the conditional score method to find a treatment effect estimate and the second way is to ignore the longitudinal data and fit the survival observations to a Cox model without the longitudinal data and find the maximum partial likelihood estimate of the treatment effect. We are interested in comparing the sample sizes required in each method to achieve the same power. A comparison of these sample sizes reflects the efficiency of the inclusion of the longitudinal data.

For clarity, the joint model is given by

$$W_i(t) = b_{0i} + b_{1i}t + \epsilon_i(t) \tag{4.54}$$

$$h_i(t) = h_0(t)\exp\{\gamma(b_{0i} + b_{1i}t) + \eta_J Z_i\}. \tag{4.55}$$

where

- $\begin{bmatrix} b_{i0} \\ b_{i1} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_0 \\ \mu_1 \end{bmatrix}, \begin{bmatrix} \phi_0^2 & 0 \\ 0 & \phi_1^2 \end{bmatrix}\right)$
- $\epsilon_i(t)|\mathbf{b}_i \sim N(0, \sigma^2)$.

To perform a fixed sample analysis using this model, we shall test the one-sided hypothesis

$$H_0^{(J)} : \eta_J \geq 0, \qquad H_A^{(J)} : \eta_J < 0.$$

We fit the joint model using the conditional score method to find a treatment effect estimate $\hat{\eta}_J$ in order to perform this hypothesis test.

We first consider a fixed sample size trial design. Our aim is to find the sample size, n_J, required using the conditional score method to achieve Type 1 error rate $\alpha = 0.025$ when the true treatment effect is $\eta_J = 0$ and power $1 - \beta = 0.9$ when $\eta_J = -0.5$. An estimate of the sample size is denoted by $\hat{n}_J$, and will be calculated by simulating clinical trials. As laid out in Section 4.3, the trial is designed with 2 years recruitment and 3 years follow-up. When increasing the sample size, we do so by increasing the rate of recruitment so accrual and follow-up periods in the trial design stay fixed. This is to ensure that differences in power are purely due to the sample size and not changes in the trial design as sample size increases.

In Appendix 4.A we present a method for estimating the root of a function which is measured with error, and we use this method to find an estimate $\hat{n}_J$ for the sample size as defined above. The parameter values for this algorithm are also discussed and presented in Appendix 4.A. To perform this estimation, data is simulated from this model with parameter values chosen and described in Section 4.3. The common values used in simulations are $(\mu_0, \mu_1) = (6, 3), \phi_0 = 3.5, \phi_1 = 2.5$ and $h_0(t) = 5.5$. We are interested in differences when we simulate using the values $\gamma = 0, 0.03, 0.06, 0.09$ and $\sigma^2 = 0, 1, 10, 100$. Further, these simulations are performed under H_A with the case $\eta_J = -0.5$. Sample size estimates $\hat{n}_J$ are given in Table 4.20. The column "Naive model" will shortly be described.

	$\sigma^2 = 0$	$\sigma^2 = 1$	$\sigma^2 = 10$	$\sigma^2 = 100$	Naive model
$\gamma = 0$	322	322	320	329	316
$\gamma = 0.03$	328	329	331	330	358
$\gamma = 0.06$	326	328	326	370	464
$\gamma = 0.09$	327	332	334	460	517

Table 4.20: Sample sizes required for power 0.9 in true and naive model for a fixed sample clinical trial.

The sample size increases as σ^2 increases. This reflects that noisy longitudinal data is associated with high variance or small information levels. Sample sizes are particularly high in each case where $\sigma^2 = 100$, which has been chosen as an extreme value. Further, sample sizes appear to increase slightly with γ.

We now consider the analysis when the longitudinal data is ignored. We believe the joint model to be true and correct, however we shall fit the data to a Cox model. To do so, we shall simulate data from the joint model and then fit this data to a misspecified Cox proportional hazards model. The Cox model is given by:

$$\lambda_i(t) = \tilde{\lambda}_0(t) \exp\{\eta_C Z_i\}. \tag{4.56}$$

For this clinical trial, we test the hypothesis

$$H_0^{(C)} : \eta_C \geq 0, \qquad H_A^{(C)} : \eta_C < 0 \tag{4.57}$$

and we find a treatment effect estimate $\hat{\eta}_C$ using the maximum partial likelihood method as described in Section 3.2.1.

Although this model is misspecified, type 1 error is not affected. This is because, under $H_0^{(J)}$ we have $\eta_J = 0$ and there is no difference between treatment groups in overall survival. When fitting this data to the Cox model, the longitudinal data trajectory is reflected in the function $\tilde{\lambda}_0(t)$ and we also have that $\eta_C = 0$. Hence, $H_0^{(C)}$ is also true.

Let n_C be the sample size such that we achieve type 1 error $\alpha = 0.025$ when $\eta_J = 0$ and power $1 - \beta = 0.9$ when $\eta_J = -0.5$ when we perform the hypothesis test in (4.57). A sample size estimate for this model is denoted $\hat{n}_C$ and is found using a similar method to the calculation for $\hat{n}_J$. Note that data is simulated under $H_A^{(J)}$ with values $(\mu_0, \mu_1) = (6, 3), \phi_0 = 3.5, \phi_1 = 2.5, h_0(t) = 5.5$ and $\eta_J = -0.5$. Further, the value of γ for simulation is varied. We do not need to worry about σ^2 since this plays no role in simulating survival times, and the longitudinal data, which is affected by σ^2, is ignored. The column "Naive model" of Table 4.20 displayed estimates $\hat{n}_C$. As the value of γ increases, the sample size estimate $\hat{n}_C$ increases. This represents that as the longitudinal data has more weight in the survival hazard rate, ignoring the longitudinal data results in an increasingly inefficient clinical trial. When $\gamma = 0$, this represents the case where longitudinal data is available yet has no influence on the survival function. In this case, $\hat{n}_C < \hat{n}_J$ and it is more efficient to fit the data to the simple Cox model.

To compare the sample sizes obtained using the joint model and the naive misspecified Cox model, we define "relative efficiency" to be

$$RE = \frac{n_C}{n_J}.$$

Using this definition, when $RE > 1$ we interpret this as the joint model analysis being the more efficient model to use and similarly when $RE < 1$, the Cox model analysis is the more efficient analysis method.

Table 4.21 shows the relative efficiency results for the fixed sample clinical trial. We see that in general, $RE > 1$ and the joint model is performing more efficiently than the analysis with the simple Cox model. In the most extreme case, 1.58 times as many patients are required using the Cox model to achieve the same power as when the joint modelling framework is used. Relative efficiency increases as γ increases, which is interpreted as the impact of the longitudinal data on the survival hazard rate. We also see that when $\gamma = 0$, e.g when longitudinal data is available but has no impact on the survival endpoint, we have $RE < 1$. Hence, when this is the case, it is slightly more efficient to use the Cox model analysis. For small and sensible values of σ^2 relative efficiency is not affected by changes in σ^2. However, relative efficiency is lower when $\sigma^2 = 100$. This reflects that using the joint model to analyse the data is particularly efficient when the longitudinal measurement error is not extreme.

	$\sigma^2 = 0$	$\sigma^2 = 1$	$\sigma^2 = 10$	$\sigma^2 = 100$
$\gamma = 0$	0.98	0.98	0.99	0.96
$\gamma = 0.03$	1.09	1.09	1.08	1.09
$\gamma = 0.06$	1.42	1.41	1.42	1.25
$\gamma = 0.09$	1.58	1.56	1.55	1.12

Table 4.21: Ratio of sample sizes required for power 0.9 in true and naive mode for a fixed sample clinical trial.

We shall now extend these sample size and relative efficiency results to group sequential trials. The parameter values remain the same as in Section 4.3.3. We have chosen to use an error spending design with parameter $\rho = 2$ with 2 years recruitment and 3 years follow up. Further details of the trial design can be found in Section 4.3.3. Similarly to the fixed sample case, when the sample size is increased, we increase the rate of recruitment so that the sample size is the only variable affecting power. The root finding algorithm described in Appendix 4.A is used to find the values in Table 4.22

Table 4.22 shows the maximum sample sizes required to achieve power $1-\beta = 0.9$ when $\eta_J = -0.5$ for the group sequential trial. The first analysis, at 19 months, occurs just before the end of the recruitment period which is 2 years. Trials that terminate at the first interim analysis may recruit less than n_J or n_C patients, however this occurs with very small probability. Hence, the expected sample size will be very close to the maximum sample size for each model and therefore the maximum sample size is a useful measure to compare methods. Clearly, a similar pattern is seen to the fixed sample case; using the true model, sample sizes increase with σ^2 and γ. When the misspecified Cox model is used to analyse the data, the sample size increases with γ.

	$\sigma^2 = 0$	$\sigma^2 = 1$	$\sigma^2 = 10$	$\sigma^2 = 100$	Naive model
$\gamma = 0$	363	364	364	373	363
$\gamma = 0.03$	365	365	364	374	421
$\gamma = 0.06$	364	365	365	420	528
$\gamma = 0.09$	365	369	375	522	607

Table 4.22: Maximum sample sizes required for power 0.9 in true and naive model for a group sequential clinical trial.

Comparing Tables 4.20 and Table 4.22 we see that the maximum sample size for the group sequential trial is roughly 1.1 times the sample size for the fixed sample trial. This corresponds to the inflation factor R described by Jennison and Turnbull (2000) for a group sequential test with $K = 5$ equally spaced analyses. This inflation factor describes the increase in information, and therefore sample size, when no early stopping occurs in the group sequential trial. The discrepancies of these ratios from 1.1 is because the group sequential tests are designed to have equally spaced information levels but in practice this will not be the case.

The maximum sample sizes needed to achieve power of 0.9 of these two analysis methods for a group sequential trial is compared. Table 4.23 shows the relative efficiency, $RE = n_C/n_J$, results for this group sequential trial.

	$\sigma^2 = 0$	$\sigma^2 = 1$	$\sigma^2 = 10$	$\sigma^2 = 100$
$\gamma = 0$	1.00	1.00	1.00	0.97
$\gamma = 0.03$	1.16	1.16	1.16	1.13
$\gamma = 0.06$	1.45	1.45	1.45	1.26
$\gamma = 0.09$	1.67	1.65	1.62	1.16

Table 4.23: Ratio of maximum sample sizes required for power 0.9 in true and naive mode for a group sequential clinical trial.

Similarly to the fixed sample analysis, we see that relative efficiency increases with γ and remains constant with σ^2 apart from the case where $\sigma^2 = 100$ which reflects extremely noisy data. Also, we see that $RE = 0.97$ when $\gamma = 0$ and $\sigma^2 = 100$ which indicates that when the longitudinal data is not correlated with the survival endpoint and the longitudinal data is noisy, the simple Cox model is a slightly more efficient method for estimating the treatment effect. Apart from the case where $\gamma = 0$, it is always more efficient to analyse the data using the joint modelling approach. Even when $\gamma = 0$, fitting the data to the simple Cox model for survival data is only marginally more efficient than fitting the data to the joint model. In the extreme case, 1.67 times as many patients are required to analyse the data using the Cox model as when the joint modelling framework is used. A reduction in sample size of 67% is incredibly beneficial and the results are overwhelmingly conclusive that when longitudinal observations are available, it is more efficient to fit the data to the joint model than the simple Cox model for survival data. This is true even for the case where the biomarker is only slightly correlated with survival.

APPENDIX

4.A | Root finding algorithm for power calculations

4.A.1 | Stochastic observations

In this Section, we introduce an algorithm for sample size calculations when power is estimated using simulation. The power function can be measured with error at different places along a curve, and we would like to find a best estimate for the root of the true function. We shall introduce a two-step algorithm which aims to accurately calculate the root of the true function, and we shall discuss how to optimise the algorithm given a finite number of simulations. The accuracy in the calculation of the stochastic function estimate increases with Monte Carlo sample size, however, estimating the function is computationally expensive. Therefore, we discuss where to sample along the curve when the number of Monte Carlo replicates to be performed is limited.

The Robbins-Monro algorithm created by Robbins and Monro (1951) aims to find the root of a function which can be measured with error at points along a curve. The Robbins-Monro algorithm is an iterative method and the authors prove the convergence of the sequence of estimates to the true value. We create a new algorithm which places importance on the number of iterations that are performed and the need for computational efficiency. Further, since we are interested in power estimates only, we shall fit the power observations to a specified model. Our algorithm takes influence from the Robbins-Monro algorithm, which converges to the truth, however our algorithm takes advantage of at least partial knowledge about the form of the power function.

Let $p(n) : \mathbb{R} \rightarrow [0,1]$ be a strictly increasing continuous function. For this analysis, we restrict attention to the case where the codomain of the function $p(\cdot)$ is the closed set $[0,1]$, therefore $p(\cdot)$ represents a probability. Suppose that the form of $p(\cdot)$ is unknown, but instead, $\hat{p}(n)$ can be calculated as an estimate of $p(n)$ so that

$$\hat{p}(n) = p(n) + \epsilon(n)$$

where $\epsilon(n)$ is a random error term. Figure 4.A.1 gives a graphical representation of the problem, where the red line is the unobservable function. The black dots show the simulated estimates of the function at different values of n. These observations can be used to estimate the function.

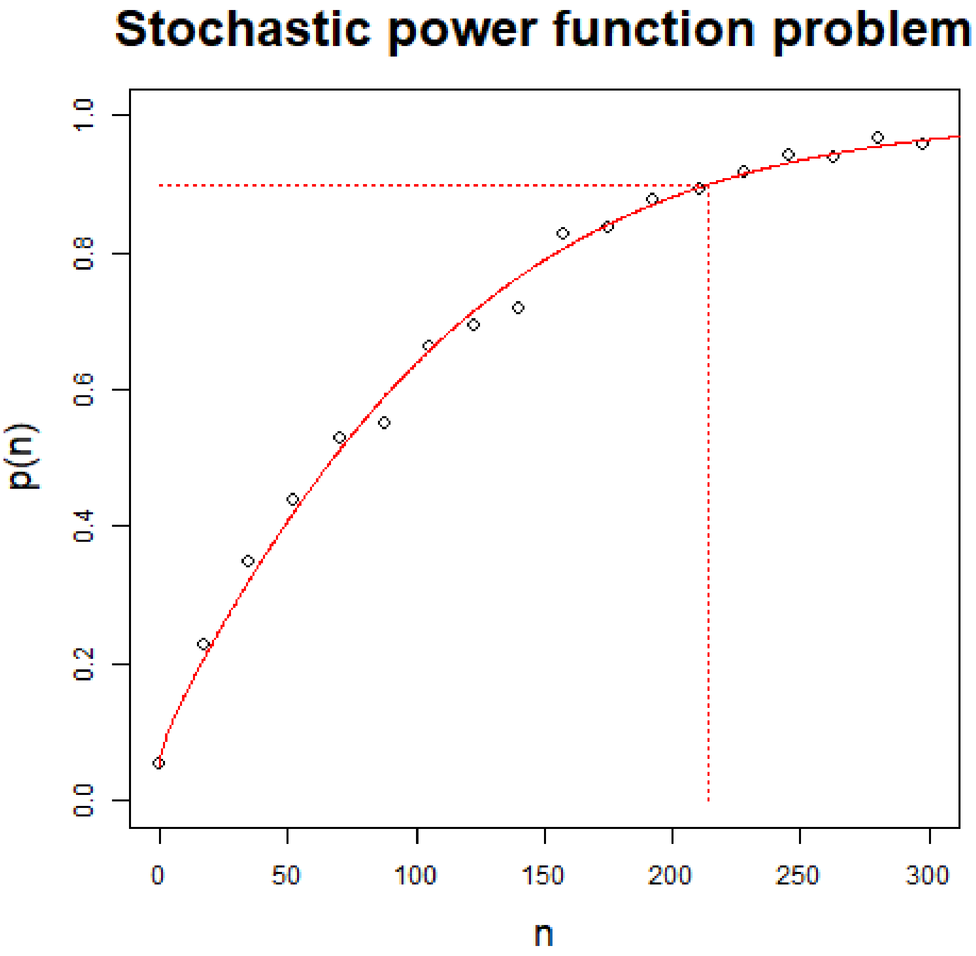


Figure 4.A.1: Stochastic approximation problem. Red line is true, unknown function and red dotted line shows the root for $p(n) = 0.9$. Black dots are simulated estimates of the function.

In the context of this Thesis, the function $p(n)$ will be the power of a test for sample size n, and $\hat{p}(n)$ will be the Monte Carlo calculation that simulates clinical trials and takes the power estimate as the proportion of times the null hypothesis is rejected. In this sample size and power example, we are interested in finding the value of n such that $p(n) = 0.9$, e.g the sample size that attains 90% power.

The information that n and $p(n)$ represent sample size and power can be used to specify a form for $p(\cdot)$. In clinical trials, it is often the case that the treatment effect estimate is asymptotically normally distributed. This is shown for the joint model in Section 4.2.1 for a fixed sample trial and also Section 4.2.2 for the group sequential trial. Let θ be the treatment effect and suppose that, given $\theta = \delta$, the treatment effect estimate is distributed as $\hat{\theta} \sim N(\delta, \sigma^2/n)$ where σ^2 is a constant. Then, for significance level α, the power of a one-sided hypothesis test for $H_0 : \theta = 0, H_A : \theta > 0$ is given by

$$p_1(n; \sigma) = \Phi\left(\frac{\delta}{\sigma}\sqrt{n} - \Phi^{-1}(1-\alpha)\right). \tag{4.58}$$

This result holds when the treatment effect is normally distributed, which we have proven asymptotically. Hence, for small sample results we cannot be sure that this is the correct form for the power curve. Power is known to be an increasing function of n and another option for fitting the power curve is

$$p_2(n; \beta_0, \beta_1) = \beta_0 + \beta_1 n. \tag{4.59}$$

Although this function may seem a poor approximation in general, we may apply this approximation locally. For any function $p(n)$, a Taylor expansion around the point n shows that locally, this linear approximation is appropriate. We will show, through simulation, that this function provides an accurate estimate for power when we restrict attention to a small interval.

4.A.2 | Two-step algorithm

We shall now describe the root-finding algorithm in two stages. We would like to find a value n^* such that $p(n^*) = y$. The form of the stochastic function is unknown and the outcome of this algorithm will be the estimate $\hat{n}^*$ for n^*. The first step of this algorithm is presented below:

1. Calculate $\hat{p}(n_0)$ using N_0 Monte Carlo simulations, where n_0 is a starting guess for n, and the choice for N_0 is discussed later in this Section.

2. Using Equation (4.58), fit the point $(n_0, \hat{p}(n_0))$ to the curve to get an estimate

$\hat{\sigma}$ and a fitted curve $p_1(n;\hat{\sigma})$.

3. Find n_1, n_2 and n_3 such that

$$\begin{aligned} p_1(n_1;\hat{\sigma}) &= y - r \\ p_1(n_2;\hat{\sigma}) &= y \\ p_1(n_3;\hat{\sigma}) &= y + r \end{aligned}$$

where the choice for r is discussed later in this section.

Figure 4.A.2 represents the first stage of the algorithm. The red line shows the true underlying function $p_1(n;\sigma)$. We choose $n_0 = 100$ which results in $\hat{p}(100) = 0.638$. Then, the function $p_1(n;\hat{\sigma})$ is calculated and shown in black with the dotted lines indicating where to place points n_1, n_2 and n_3.

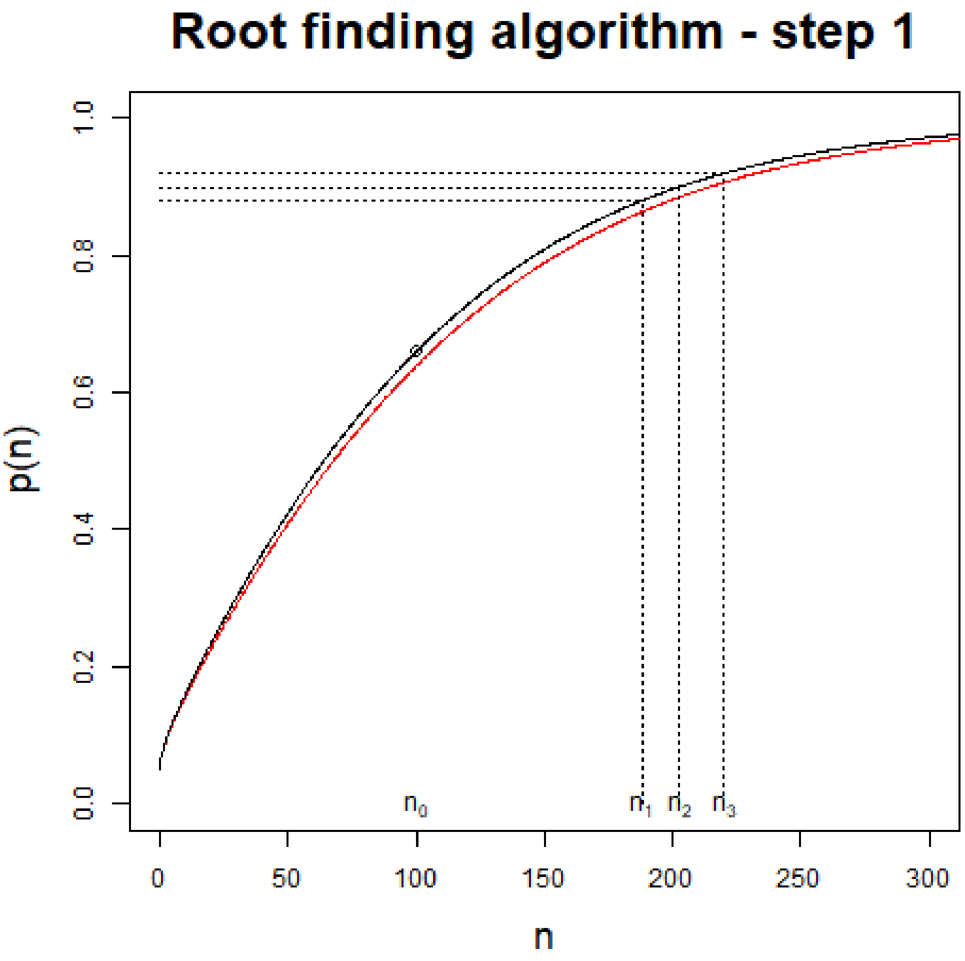


Figure 4.A.2: First stage of the stochastic root finding algorithm. $p_1(n,\sigma)$ given by red line and $p_1(n;\hat{\sigma})$ given by black line. Dotted lines show where to place n_1, n_2 and n_3.

The second stage of the algorithm is as follows:

1. Calculate $\hat{p}(n_1), \hat{p}(n_2)$ and $\hat{p}(n_3)$ using N Monte Carlo replicates at each point.
2. Check that $\hat{p}(n_1) < y$ and $\hat{p}(n_2) > y$. If not
   (a) Using Equation (4.58) fit the points $(n_1, \hat{p}(n_1)), (n_2, \hat{p}(n_2))$ and $(n_3, \hat{p}(n_3))$ to obtain the fitted curve $p_1(n; \hat{\sigma})$.
   (b) Find n_1, n_2 and n_3 such that
   $$\begin{aligned} p_1(n_1; \hat{\sigma}) &= y - r \\ p_1(n_2; \hat{\sigma}) &= y \\ p_1(n_3; \hat{\sigma}) &= y + r. \end{aligned}$$
   (c) Repeat steps 1 and 2.
3. Using Equation (4.59) fit the linear model to the points $(n_1, \hat{p}(n_1)), (n_2, \hat{p}(n_2))$ and $(n_3, \hat{p}(n_3))$ to obtain the fitted line $p_2(n; \hat{\beta}_0, \hat{\beta}_1)$.
4. Calculate the final estimate for n^* as the solution to $p_2(n^*; \hat{\beta}_0, \hat{\beta}_1) = y$.

The step of the algorithm that we have just described fits a curve to three points. Suppose that the three points do not surround y so that either $\hat{p}(n_1) > y$ and/or $\hat{p}(n_3) < y$. Then, the final prediction for $\hat{n}^*$ will be taken by extrapolating outside of the data. This is particularly problematic when the points $(n_1, \hat{p}(n_1)), (n_2, \hat{p}(n_2))$ and $(n_3, \hat{p}(n_3))$ are fitted to a straight line as extrapolation could result in a highly inaccurate estimate, $\hat{n}^*$. This is the reason why we have introduced a check between steps 1 and 2 of the second stage of the algorithm which is repeated until the condition is satisfied that $\hat{p}(n_1) < y$ and $\hat{p}(n_3) > y$.

In Figure 4.A.3, there is a visual representation of the second stage of the root finding algorithm. Starting with values $n_1 = 202.8, n_2 = 214.5$ and $n_3 = 228.0$ given by stage 1, the Monte Carlo approximation gives $\hat{p}(n_1) = 0.892, \hat{p}(n_2) = 0.903$ and $\hat{p}(n_3) = 0.908$. The red line indicates the true underlying function $p(\cdot)$ and the black line is the data fitted to Equation (4.59). The final estimate for n^* is $\hat{n}^* = 213.5$.

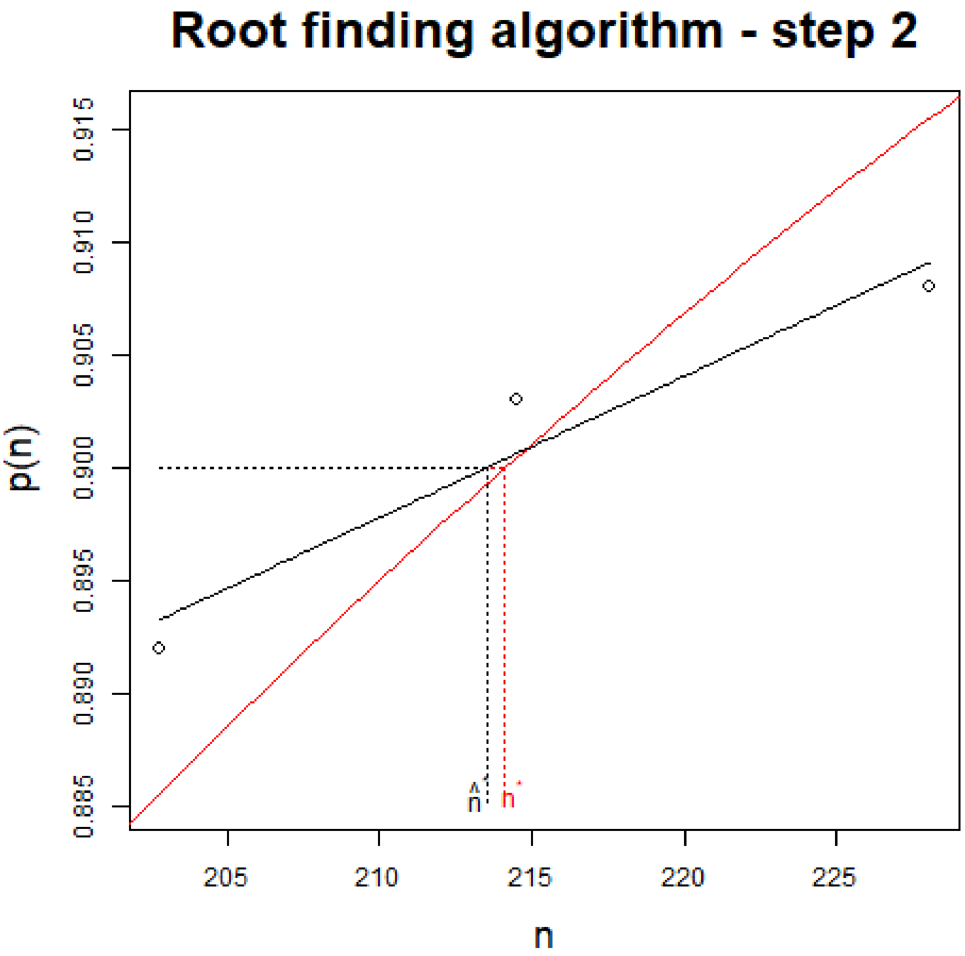


Figure 4.A.3: Second stage of the stochastic root finding algorithm. $p_1(n, \sigma)$ shown in red, $p_2(n; \hat{\beta}_0, \hat{\beta}_1)$ shown in black. Dotted lines show where to place $\hat{n}^*$ with a comparison to the true value n^* given by the red dotted line.

The red line in Figure 4.A.3 represents the true underlying power curve. The black points are simulated from this model and then the black line is the fitted linear model for these points. It is apparent that a linear approximation for the power curve Equation (4.58) is appropriate here because the red line appears straight in this small interval. Further, it is clear that the difference between the true power curve in red and the linear fitted line in black is due to large vertical distance between the true power curve and the power observations. Hence, the difference between these two lines is dominated by the simulation error and not the error in the linear approximation. The points n_1, n_2 and n_3 are far enough apart from each other so that we have $\hat{p}(n1) < \hat{p}(n_2) < \hat{p}(n_3)$ and the slope β_1 is accurately estimated (note the small range on the $y-$ axis). The distance between points is controlled using the parameter r and we will shortly discuss this choice in detail. It is also clear that the value r has been chosen appropriately so that the true power curve appears linear in this local interval. We see that the difference between n^* and $\hat{n}^*$ is very small and is primarily because of the simulation error in calculating $\hat{p}(n_1), \hat{p}(n_2)$ and $\hat{p}(n_3)$.

4.A.3 | Optimising the root finding algorithm

Some elements of the algorithm design contribute to the efficiency and accuracy of the algorithm. We simulate data from a distribution where the true curve $p(\cdot)$ is known, so that n^* is known and compare to the results of the algorithm that generate an estimate $\hat{n}^*$.

The parameters in the algorithm to be chosen are: N_0 and N, the number of Monte Carlo simulations at the starting guess n_0, and points n_1, n_2, n_3, and r, the distance from y which determines where to place the points n_1 and n_3.

The smaller the value of r, the more likely that the conditions $\hat{p}(n_1) > y$ and $\hat{p}(n_3) < y$ are true, so that the values of n_1, n_2 and n_3 must be re-calculated. This re-calculation will result in further simulations being performed, which decrease the efficiency. However, when a straight line is fitted to the data with an underlying concave function, the resulting estimate (along the x-axis) will be an overestimate of the truth. The level of over-estimation varies with the range of the data, as the approximation of a curve to a linear becomes more accurate as localisation increases. Therefore, a small r value reduces the bias in $\hat{n}^*$ and we see that there is a bias-variance trade-off associated with the choice of r.

Since the number of possible simulations is restricted, our choice for N is constant at 30,000. High accuracy in the estimate $\hat{p}(n_0)$ allows n_1, n_2 and n_3 to be chosen with accuracy, but we expect this to be insignificant in comparison to the accuracy in the estimates $\hat{p}(n_1), \hat{p}(n_2)$ and $\hat{p}(n_3)$. Hence, N_0 is chosen to be small compared to N.

We generate data where the treatment effect estimate is distributed by

$$\hat{\theta} \sim N\left(0.6, \frac{2.5^2}{n}\right). \tag{4.60}$$

By Equation (4.58), for power equal to 0.9, $n^* = 214.1$ is required. Table 4.A.1 shows the results of the stochastic approximation method when the algorithm is performed 10^5 times for each combination of n_0, N_0 and r. “Total simulations” gives the mean number of simulations that occur due to extrapolation plus $N_0 + 3N$ for the initially allocated number of simulations.

$\mathbf{n_0}$	$\mathbf{N_0}$	**r**	**Total**	**Linear method**		**Normal method**	
			simulations	$\hat{\mathbf{n}}^*$	**MSE**	$\hat{\mathbf{n}}^*$	**MSE**
175	10000	0.0075	102515.5	214.2 (0.93)	0.87	214.1 (0.83)	0.69
175	10000	0.01	100299.7	214.3 (0.90)	0.87	214.1 (0.83)	0.69
175	10000	0.0125	100015.3	214.5 (0.89)	0.92	214.1 (0.84)	0.70
175	15000	0.0075	106006.2	214.2 (0.92)	0.85	214.1 (0.83)	0.69
175	15000	0.01	105066.6	214.3 (0.90)	0.86	214.1 (0.84)	0.70
175	15000	0.0125	105001.8	214.5 (0.88)	0.91	214.1 (0.84)	0.70
175	20000	0.0075	110515.7	214.2 (0.91)	0.85	214.1 (0.83)	0.69
175	20000	0.01	110026.1	214.3 (0.89)	0.84	214.1 (0.83)	0.70
175	20000	0.0125	110000.9	214.5 (0.87)	0.90	214.1 (0.83)	0.70
200	10000	0.0075	102734.2	214.2 (0.93)	0.87	214.1 (0.83)	0.69
200	10000	0.01	100349.2	214.3 (0.91)	0.87	214.1 (0.83)	0.69
200	10000	0.0125	100019.8	214.5 (0.88)	0.91	214.1 (0.83)	0.70
200	15000	0.0075	106094.4	214.2 (0.92)	0.86	214.1 (0.83)	0.69
200	15000	0.01	105071.1	214.3 (0.90)	0.86	214.1 (0.84)	0.70
200	15000	0.0125	105001.8	214.5 (0.88)	0.91	214.1 (0.84)	0.70
200	20000	0.0075	110554.4	214.2 (0.92)	0.85	214.1 (0.83)	0.70
200	20000	0.01	110025.2	214.3 (0.89)	0.84	214.1 (0.84)	0.70
200	20000	0.0125	110000.9	214.5 (0.87)	0.91	214.1 (0.84)	0.70
225	10000	0.0075	102943.0	214.2 (0.92)	0.87	214.1 (0.83)	0.68
225	10000	0.01	100371.7	214.3 (0.91)	0.88	214.1 (0.84)	0.70
225	10000	0.0125	100027.0	214.5 (0.89)	0.92	214.1 (0.84)	0.71
225	15000	0.0075	106242.0	214.2 (0.92)	0.85	214.1 (0.83)	0.69
225	15000	0.01	105099.9	214.3 (0.90)	0.86	214.1 (0.84)	0.70
225	15000	0.0125	105004.5	214.5 (0.88)	0.91	214.1 (0.84)	0.70
225	20000	0.0075	110638.1	214.2 (0.91)	0.85	214.1 (0.83)	0.69
225	20000	0.01	110029.7	214.3 (0.89)	0.85	214.1 (0.84)	0.70
225	20000	0.0125	110000.0	214.5 (0.87)	0.90	214.1 (0.84)	0.70

Table 4.A.1: Root finding algorithm results for normally distributed treatment effect estimate.

The column "Normal method" in Table 4.A.1 is included for comparison. This shows the results when the function Equation (4.58) is fitted in step 2 of the algorithm. This is the model in which the data has been simulated from and so the estimate will be unbiased.

The results in Table 4.A.1 show that Equation (4.58) should be used in the root finding algorithm to find the sample size for a given power requirement when the treatment effect estimate is known to be normally distributed. This is obvious since

Equation (4.58) defines the power of a test for a normally distributed estimate. This is because the mean squared error (MSE) is smaller for the normal method than the linear method in all cases. The starting point n_0 has very little impact on the accuracy and efficiency of the algorithm. This is because the true, known model is used to fit the curve to the starting point. Also, we can see that increasing the number of starting Monte Carlo simulations, N_0, has little impact on the MSE, but increases the total number of simulations. We believe that the gain in accuracy at the starting point is not worth the increase in computation time. Finally, the value of r has an impact on the total number of simulations, with a smaller value of r corresponding to larger simulation times. Larger values of r also imply higher MSE when Equation (4.58) is used in step 2 of the algorithm, however when Equation (4.59) is used in setup 2 of the algorithm, we see that overall, the MSE is smallest for $r = 0.01$.

We now consider an example where the treatment effect estimate is not normally distributed. In most clinical trial scenarios, it is possible to show that the treatment effect estimate is asymptotically normally distributed, however the true distribution in small samples is unknown. Therefore, it is desirable to find the optimum parameters for this algorithm for a case where the treatment effect estimate follows a distribution other than normal. The distribution of the estimate is given by

$$\hat{\theta} \sim t\left(1.25, \theta\sqrt{n}\right). \tag{4.61}$$

Table 4.A.2 shows the results of using this algorithm when the treatment effect estimate $\hat{\theta}$ has the distribution in Equation (4.61). In this example, although we known the true function $p(\cdot)$, we shall still use the Equations (4.58) and (4.59) and assess the results, as this is what would happen if the form of $p(\cdot)$ was unknown. Using the true distribution of $\hat{\theta}$ given by Equation (4.61), we can calculate that to obtain significance $\alpha = 0.025$ and power $1 - \beta = 0.9$ when $\theta = 0.5$, we require $n^* = 676.2$.

$\mathbf{n_0}$	$\mathbf{N_0}$	**r**	**Total**	**Linear method**		**Normal method**	
			simulations	$\mathbf{\hat{n}^*}$	**MSE**	$\mathbf{\hat{n}^*}$	**MSE**
200	10000	0.0125	280525.6	677.1 (4.08)	17.47	671.1 (2.74)	33.22
200	10000	0.015	279681.4	677.5 (4.05)	18.06	670.8 (2.85)	37.37
200	10000	0.0175	275868.1	677.5 (4.55)	22.41	669.9 (3.83)	54.02
200	15000	0.0125	285493.2	677.1 (4.08)	17.47	671.1 (2.73)	33.06
200	15000	0.015	284733.6	677.5 (4.05)	18.02	670.8 (2.83)	37.32
200	15000	0.0175	281216.4	677.5 (4.54)	22.34	669.9 (3.79)	53.34
200	20000	0.0125	290484.2	677.1 (4.11)	17.69	671.1 (2.73)	33.16
200	20000	0.015	289748.0	677.5 (4.02)	17.82	670.8 (2.83)	37.35
200	20000	0.0175	286292.9	677.5 (4.55)	22.50	669.9 (3.77)	53.17
400	10000	0.0125	203112.1	675.9 (4.00)	16.13	669.1 (3.36)	61.42
400	10000	0.015	191944.0	676.9 (4.20)	18.08	668.3 (2.85)	70.44
400	10000	0.0175	190105.3	677.5 (4.13)	18.89	667.9 (2.77)	76.36
400	15000	0.0125	207766.5	675.8 (3.98)	15.97	669.1 (3.34)	61.81
400	15000	0.015	196808.1	676.9 (4.20)	18.17	668.3 (2.85)	70.22
400	15000	0.0175	195087.3	677.5 (4.14)	19.00	667.9 (2.75)	76.27
400	20000	0.0125	212477.6	675.8 (3.99)	16.01	669.1 (3.32)	61.78
400	20000	0.015	201627.2	676.9 (4.18)	17.90	668.3 (2.80)	70.36
400	20000	0.0175	200081.0	677.5 (4.13)	18.95	667.9 (2.74)	76.14
600	10000	0.0125	110031.4	676.4 (4.12)	17.01	670.3 (3.45)	46.48
600	10000	0.015	102380.5	677.1 (4.08)	17.52	669.7 (3.36)	53.91
600	10000	0.0175	100367.2	677.7 (3.96)	18.11	669.3 (3.32)	58.74
600	15000	0.0125	112716.6	676.4 (4.10)	16.89	670.1 (3.27)	47.20
600	15000	0.015	106323.9	677.2 (4.08)	17.65	669.6 (3.14)	53.26
600	15000	0.0175	105127.8	677.8 (3.98)	18.41	669.3 (3.13)	57.48
600	20000	0.0125	116443.1	676.5 (4.10)	16.90	670.0 (3.17)	47.89
600	20000	0.015	110864.9	677.2 (4.10)	17.89	669.6 (3.03)	52.87
600	20000	0.0175	110050.4	677.8 (3.96)	18.28	669.3 (3.01)	57.12

Table 4.A.2: Root finding algorithm results for non-normally distributed treatment effect estimate.

Table 4.A.2 shows that for this example, the linear method has a smaller MSE than the normal method in all cases and is less biased since the estimates $\hat{n}^*$ are closer to the true n^*. The starting point n_0 is crucial because a large difference in total simulations is seen when the starting point approaches the true value n^*. When this is the case, further effort should be giving to choosing a suitable n_0. This could be by performing simulations at multiple values of n_0 and choosing the one with $\hat{p}(n_0)$ closest to 0.9. Similarly to the previous example, the value of N_0 has little impact on the accuracy of the algorithm and simply contributes to the total number of simulations. Finally, the value of r is influential since increasing r decreases the average number of simulations however the MSE for the linear method decreases.

In Section 4.5, this root finding algorithm is implemented to find the required sample size for a test based on the joint model. We present an example implementation of this root finding algorithm for the case $\gamma = 0.03, \sigma^2 = 100$. Using the results of Tables 4.A.1 and 4.A.2 we have chosen to use $N_0 = 10^4$ since an increase in N_0 results in very little gain on the accuracy of the algorithm. Some thought should be given to the starting point n_0, especially for trials where the sample size is small and the normality of the treatment effect estimate is questionable. In Section 4.3.2 we describe a method for calculating a suitable starting value, which uses a very large dataset to accurately estimate the variance of the treatment effect. We show that a suitable initial sample size estimate is given by by $n_0 = 311$. When $\hat{\theta}$ is not normally distributed, the linear method out-performs the normal method with respect to the bias of the estimate, and when $\hat{\theta}$ is normally distributed, there is nominal increase in the MSE when using the linear method compared to the normal method. We have chosen to use $r = 0.015$ since this provides a compromise between accuracy of the estimate $\hat{n}^*$ and computational efficiency of the algorithm. We obtain an initial power estimate of $\hat{p}(311) = 0.8853$. The first stage of the algorithm suggests placing the points at $n_1 = 311, n_2 = 327, n_3 = 345$. The results of simulations under these values with $N = 10^4$ Monte Carlo simulations at each point result in a sample size calculation of $\hat{n}^* = 322.8$. In practice, this sample size is then rounded up to $\hat{n}^* = 323$.

CHAPTER 5

JOINT MODEL WITH BOTH LONGITUDINAL AND SURVIVAL TREATMENT EFFECTS

5.1 | Analysis of clinical trials using the restricted mean survival time

5.1.1 | Motivation for designing tests for a joint model with both longitudinal and survival treatment effects

The joint model in Section 4.1 adjusts for the longitudinal data in the survival model by considering the biomarker trajectory as a covariate. This does not cover the case where the biomarker may also be influenced by treatment. The US Food and Drug Administration (2019) cautions against adjusting for "covariates measured after randomisation because they could be affected by the treatments." In this chapter, we shall present a joint model for longitudinal and survival data where there is a treatment effect acting through the biomarker and another directly affecting survival. The causal effect diagram in Figure 5.1.1 represents this directional effect of treatment.

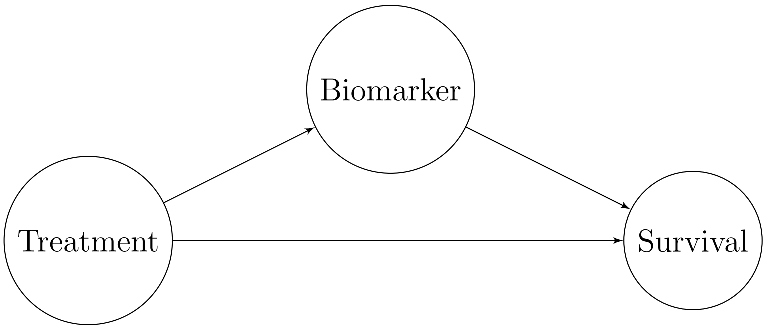


Figure 5.1.1: Causal effect diagram for the treatment effects in the joint model.

In this chapter, we introduce a joint model with both a longitudinal treatment effect and a survival treatment effect, therefore taking into account both treatment effect pathways. Model fitting is straightforward. However, the challenge arises in creating a group sequential trial which combines two treatment effect estimates. We shall us the Restricted Mean Survival Time (RMST) framework in order to summarise the treatment effect as a single variable. RMST, described by Royston and Parmar (2013), is a popular method for analysing survival data when it is expected that the proportional hazards assumption does not hold. We shall investigate general use of RMST methods. However, it is important to remember that the motivation for using RMST is because of the complex nature of the model which we believe to be true.

RMST methods have recently become popular for designing and analysing fixed sample clinical trials. The novel aspect of the research in this Chapter is the design of group sequential clinical trials using RMST. Further, we show, through simulation, that this design is more efficient than the fixed sample trial which uses RMST. The efficiency is with respect to the stopping time of the group sequential trial. In some cases, the trial stops 1.5 years early on average when using the group sequential design compared to the fixed sample trial which lasts for a total of 5 years. We also show that the group sequential version of this clinical trial results in fewer hospital visits by patients and shorter patient follow-up times.

5.1.2 | Restricted mean survival time

To perform a clinical trial based on the joint model, we require a single one dimensional summary statistic that summarises the treatment effect. We propose using the Restricted Mean Survival Time (RMST) method to combine the longitudinal and survival treatment effects into a single useful parameter. RMST has recently become a widely accepted method for dealing with survival data when the proportional hazards assumption does not hold. In our case, the motivation for using RMST is because of the complex nature of the model with multiple pathways for the treatment effect. In this section we introduce and define RMST and provide some results and examples for designing and analysing fixed and group sequential trials under this framework.

Royston and Parmar (2013) define RMST as the area under the survival curve up to time t^*. This value of t^* is fixed at the design stage. The choice of t^* may have an impact on some properties of the analysis and we shall discuss this choice in Section 5.1.6. Let F be the time-to-failure random variable and let $S(t; \boldsymbol{\theta})$ be the survival function, integrated over any patient-specific random effects. Here $\boldsymbol{\theta}$ is a vector of parameters in the model. Then RMST is defined as

$$\begin{aligned} RMST &= \int_0^{t^*} S(t; \boldsymbol{\theta}) dt \\ &= \mathbb{E}[\min(F, t^*)]. \end{aligned}$$

Since we are interested in a statistic that summarises the effect of treatment, we are interested in the difference in RMST between treatment arms. Suppose that F_0 and F_1 are time-to-failure random variables for patients on the control and treatment arms respectively, and that $S_0(t;\boldsymbol{\theta})$ and $S_1(t;\boldsymbol{\theta})$ are the corresponding survival functions integrated over any patient specific random effects. Then the parameter of interest, Δ, is

$$\begin{aligned}\Delta(t^*;\boldsymbol{\theta}) &= \int_0^{t^*} \left[S_1(t;\boldsymbol{\theta}) - S_0(t;\boldsymbol{\theta})\right] dt && (5.1)\\ &= \mathbb{E}[\min(F_1,t^*)] - \mathbb{E}[\min(F0,t^*)]. && (5.2)\end{aligned}$$

Restricted mean survival time is popular within fixed sample clinical trials. We shall show that a group sequential trial can also be created using RMST as the analysis approach, however many design aspects must be carefully and thoughtfully chosen.

5.1.3 | Parametric and non-parametric models

To find an estimate for $\Delta(t^*;\boldsymbol{\theta})$, there are many available methods. It is most common to analyse RMST using non-parametric methods. Zhao et al. (2016) describe a method for finding the confidence band for RMST and apply this result to a data set from a cardiovascular clinical trial. Some options for estimating $\Delta(t^*;\boldsymbol{\theta})$ include integration under the Kaplan-Meier survival curve and bootstrap methods for estimating the variance. One design aspect of each of these methods is that the value of t^* must be smaller than the final observation time to ensure that the estimate is identifiable. Chen and Tsiatis (2001) compare parametric and non-parametric RMST methods when the model includes confounding covariates with the treatment effect. The authors conclude that the non-parametric Kaplan-Meier estimator is severely biased in the presence of confounding variables.

Murray and Tsiatis (1999) show that it is possible to create a group sequential trial using non-parametric RMST estimates by proving the independent increments property, which is required in order for the canonical joint distribution to hold. Further, Lu and Tian (2020) consider some of the practical design challenges for group sequential trials using non-parametric RMST estimates. This non-parametric analysis uses a Kaplan-Meier estimator. If the final observation is censored, then the Kaplan-Meier estimator is non-identifiable after the final follow-up time. As a result, the authors change the value of the truncation time t^* between analyses of a group sequential trial. An estimand describes what is to be estimated based on the trial objectives and Akacha et al. (2017) discuss the potential impact that an estimand can have on the design of a clinical trial. Research into estimands was motivated by the National Research Council (2010) highlighting a need to more clearly define the measurement of interest in a clinical trial. We seek a group sequential design in which the definition of the estimand remains constant between analyses and the truncation time t^* is clinically motivated rather than data-dependent.

We have chosen to use a fully parametric approach to estimating $\Delta(t^*; \boldsymbol{\theta})$. The parameters $\boldsymbol{\theta}$ of the model are estimated using maximum likelihood and these estimates are substituted into Equation (5.1) to find an estimate $\Delta(t^*; \hat{\boldsymbol{\theta}})$. The advantage of this approach is that as long as the parameters are identifiable, $\Delta(t^*; \hat{\boldsymbol{\theta}})$ can be calculated even for t^* greater than the final follow-up time. This requires extrapolation which is not desirable but has the advantage of giving accurate results if the model is correct. Further, using this parametric approach means that a group sequential trial can be created where the truncation time t^* remains the same across analyses. It is also important to remember that the motivation for using RMST in this instance is to summarise the multi-directional treatment effect as a single parameter. If the model is regarded as correct, the parametric approach is appropriate.

A consideration with the parametric modelling approach is that the model must be fully specified, including the the form of the baseline hazard function. This is to ensure that the difference in RMST values, in Equation (5.1), can be calculated. Common choices for baseline hazard functions include Weibull, piecewise constant and spline functions.

5.1.4 | The delta method

An estimate for the difference in RMST between treatment arms, $\Delta(t^*; \hat{\boldsymbol{\theta}})$, is found by calculating parameter estimates $\hat{\boldsymbol{\theta}}$ using maximum likelihood and substituting these into Equation (5.1). To perform a group sequential trial, we need a method for estimating the variance of the estimate (and hence information) and also its distribution. We shall use the Delta method to find this distribution and also prove that this distribution has the canonical distribution in Definition 2.2, which is important for creating a group sequential trial.

The Delta method can be used whenever a transformation is to be made to a set of multivariate normally distributed parameters.

Theorem 5.1. *Let $\hat{\boldsymbol{\theta}}_n$ be a $p \times 1$ vector which is a consistent estimate for the parameter vector $\boldsymbol{\theta}$ and suppose that the vector $\hat{\boldsymbol{\theta}}_n$ has the following asymptotic distribution*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} N(0, \Sigma) \quad \text{as } n \to \infty.$$

Then for a scalar function $f : \mathbb{R}^p \to \mathbb{R}$,

$$\sqrt{n}(f(\hat{\boldsymbol{\theta}}_n) - f(\boldsymbol{\theta})) \xrightarrow{d} N\left(0, \left[\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]^T \Sigma \left[\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]\right) \qquad \textit{as } n \to \infty.$$

The proof of Theorem 5.1 is given by Doob (1935). This reasoning is based on a Taylor approximation to the function $f(\cdot)$ around $\boldsymbol{\theta}$.

Suppose that the estimates $\hat{\boldsymbol{\theta}}_n$ are found using maximum likelihood, then we know that these estimates are asymptotically normally distributed. Therefore, we can apply the Delta method to the estimate for the difference in restricted mean survival times. Let the covariance matrix of the parameter estimates be given by $Var(\hat{\boldsymbol{\theta}}_n) = \frac{1}{n}\Sigma$, then using Theorem 5.1, the estimate $\Delta(t^*; \hat{\boldsymbol{\theta}}_n)$ has the following distribution

$$\sqrt{n}\left(\Delta(t^*; \hat{\boldsymbol{\theta}}_n) - \Delta(t^*; \boldsymbol{\theta})\right) \xrightarrow{d} N\left(0, \left[\frac{\partial \Delta(t^*; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]^T \Sigma \left[\frac{\partial \Delta(t^*; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]\right).$$

It is often necessary in clinical trials to find a standardised statistic for the test statistic, that is a statistic with unit variance. This is found by dividing the test statistic by its standard deviation. The variance must be estimated and this is done by substituting the estimate $\hat{\boldsymbol{\theta}}_n$ for $\boldsymbol{\theta}$. Following this method of standardisation the statistic of interest is given by

$$Z(t^*; \hat{\boldsymbol{\theta}}_n) = \frac{\sqrt{n}\Delta(t^*; \hat{\boldsymbol{\theta}}_n)}{\sqrt{\left[\frac{\partial \Delta(t^*; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]^T \Sigma \left[\frac{\partial \Delta(t^*; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]}}. \tag{5.3}$$

This method for standardisation results in a statistic Z which is distributed approximately as $N(\mu, 1)$ for some constant μ. Under the null hypothesis, when $\Delta(t^*; \boldsymbol{\theta}) = 0$, the distribution gives a more accurate fit and we show this through examples in Section 5.1.7 and Section 5.1.8.

5.1.5 | Sample size calculation

We shall present a method for calculating sample sizes for a hypothesis test based on the difference in parametric RMST estimates. To do so, we make use of the structure of the standardised statistic $Z(t^*; \hat{\boldsymbol{\theta}}_n)$. Let $\hat{\boldsymbol{\theta}}_n$ be the maximum likelihood estimate for the vector of parameters $\boldsymbol{\theta}$ in the statistical model. Both $\hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}$ are $p \times 1$ vecots. Suppose that the $p \times p$ matrix Σ is the variance-covariance matrix for $\hat{\boldsymbol{\theta}}_n$. The difference in parametric RMST estimates is given by $\Delta(t^*; \hat{\boldsymbol{\theta}}_n)$. A one-sided clinical trial based on the difference in RMST estimates can be constructed in terms of the statistic $Z(t^*; \hat{\boldsymbol{\theta}}_n)$ in Equation (5.3). We shall test

$$H_0 : \Delta(t^*; \boldsymbol{\theta}) \leq 0 \quad \text{versus} \quad H_A : \Delta(t^*; \boldsymbol{\theta}) > 0$$

and we shall reject H_0 if $Z(t^*; \hat{\boldsymbol{\theta}}_n) > c$ for some constant c. Using Theorem 5.1, the estimate $\Delta(t^*; \hat{\boldsymbol{\theta}}_n)$ is normally distributed and its expectation and variance are used in the sample size calculation. For significance level α and power $1 - \beta$ at $\Delta(t^*; \boldsymbol{\theta}) = \delta$, the fixed trial sample size calculation is given by

$$n = \frac{\left[\frac{\partial \Delta(t^*;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]^T \Sigma \left[\frac{\partial \Delta(t^*;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] \left(\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\right)^2}{\delta^2}. \tag{5.4}$$

This calculation requires differentiation of the difference in restricted mean survival times, $\Delta(t^*; \boldsymbol{\theta})$. We can often perform this differentiation analytically or if this is not possible then numerical differentiation methods can be applied. Also, the covariance matrix Σ must be known for this sample size calculation. It is rare that the theoretical true value of Σ is known. If this matrix is not known or difficult to calculate, then we suggest simulating a large sample of patients under a set of parameter values chosen to represent a likely scenario, and using maximum likelihood methods to estimate Σ.

5.1.6 | Choice of t^*

At the design stage of a clinical trial which uses RMST, a suitable value of t^* must be chosen. In their paper, Royston and Parmar (2013) suggest choosing t^* as the value that minimises the sample size in the calculation shown in Equation (5.4). The data motivated approach suggests choosing t^* as a trade-off between maximising Δ and minimising $Var(\hat{\Delta})$. Royston and Parmar (2013) further explain that the value of t^* should ideally not exceed the maximum censored or uncensored event time to avoid extrapolation of RMST estimates. Combining these two ideas of minimising the sample size but avoiding extrapolation of estimates, we suggest choosing t^* based on its “clinical meaning” and then adjusting the time of analysis to avoid excessive extrapolation. The clinical meaning will typically be at least as great as the expected median survival time which allows for possible changes in the shape of the treatment and control survival curves early in the study. This idea will be expanded upon in Section 5.3 where we show an example of designing a trial using RMST.

In the following sections, we show examples of fixed sample clinical trials which use the difference in RMST as an endpoint. We shall give details about the sample size calculation and also show how the sample size changes with t^*. For these examples, the differences in sample size are only very small and we suggest that the choice of t^* should be based on clinical meaning with some consideration given to how this choice affects the design of the trial.

5.1.7 | Example: Exponential survival distributions

We shall now consider a simple parametric model in which we can see the effect of the choice of t^* on the sample size. This simple survival model is characterised by constant hazard rates for each treatment arm. For patient i, let $Z_i = \mathbb{I}\{\text{patient } i \text{ receives the treatment}\}$ be the treatment indicator and let λ_0 and λ_1 be hazard rates for patients on the control and treatment arms respectively. For patient i, the time-to-event random variable, F_i, therefore has the following distribution

$$\begin{aligned} F_i | Z_i = 0 &\sim \text{Exp}\{\lambda_0\} \\ F_i | Z_i = 1 &\sim \text{Exp}\{\lambda_1\}. \end{aligned}$$

We shall compare three methods; the hypothesis test based on the difference in estimated hazard rates, the parametric RMST analysis and the non-parametric RMST analysis. The first method is included because it is known to be the most efficient hypothesis test by the Neyman-Pearson Lemma and we would like to compare the efficiency that is lost when using the RMST method.

Define the parameters $\theta_0 = \log(\lambda_0)$ and $\theta_1 = \log(\lambda_1)$ and suppose that the parameter estimates $\hat{\lambda}_0$ and $\hat{\lambda}_1$ are fitted by maximum likelihood and therefore $\hat{\theta}_0 = \log(\hat{\lambda}_0)$ and $\hat{\theta}_0 = \log(\hat{\lambda}_0)$ are also MLEs. In this example, with n patients on each treatment arm and with no censoring, it is possible to show that

$$\sqrt{n} \begin{bmatrix} \hat{\theta}_0 - \theta_0 \\ \hat{\theta}_1 - \theta_1 \end{bmatrix} \xrightarrow{d} N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) \quad \text{as } n \to \infty. \tag{5.5}$$

We can easily create a test of the null hypothesis that the hazard ratio is equal to one, or equivalently the difference in the log-hazard rates is equal to zero. We shall test

$$H_0^{(1)} : \theta_0 - \theta_1 = 0 \quad \text{versus} \quad H_A^{(1)} : \theta_0 - \theta_1 > 0$$

and we shall reject $H_0^{(1)}$ whenever $\hat{\theta}_0 - \hat{\theta}_1 > c_1$ for some constant c_1. For this hypothesis test with significance level α and power $1 - \beta$ when $\theta_0 - \theta_1 = D_1$, we require the total sample size $n^{(1)}$ given by

$$n^{(1)} = \frac{2\,V_1\left(\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\right)^2}{D_1^2} \tag{5.6}$$

where $V_1 = 2$. This form for $n^{(1)}$ will be useful for a comparison between the three analysis methods. For each method $i = 1, 2, 3$ we shall present the form of the variance V_i and the mean of the treatment effect under H_A, given by D_i.

We now consider a hypothesis test based on RMST for this exponential example. The corresponding survival functions, $S_0(t)$ and $S_1(t)$, for patients on the control and treatment arm respectively are given by

$$\begin{aligned} S_0(t) &= \exp\{-\lambda_0 t\} \\ S_1(t) &= \exp\{-\lambda_1 t\} \end{aligned}$$

and by integrating the difference between survival functions, up to t^*, we find the difference in restricted mean survival times for this model to be

$$\Delta(t^*; \lambda_0, \lambda_1) = \frac{1 - \exp\{-\lambda_0 t^*\}}{\lambda_0} - \frac{1 - \exp\{-\lambda_1 t^*\}}{\lambda_1}.$$

We estimate this function by substituting $\hat{\lambda}_0$ and $\hat{\lambda}_1$ for λ_0 and λ_1 respectively to obtain $\Delta(t^*; \hat{\lambda}_0, \hat{\lambda}_1)$.

For the asymptotic distribution of the estimate $\Delta(t^*; \hat{\lambda}_0, \hat{\lambda}_1)$, we shall apply the Delta method of Theorem 5.1. The estimates $\hat{\theta}_0 = \log(\hat{\lambda}_0)$ and $\hat{\theta}_1 = \log(\hat{\lambda}_1)$ are asymptotically normally distributed and hence, we apply the Delta method under this parameterisation. Therefore, we shall take derivatives with respect to the parameters θ_0 and θ_1. These are

$$\begin{aligned} \frac{\partial}{\partial \theta_0}\Delta(t^*; \lambda_0, \lambda_1) &= \frac{\partial}{\partial \lambda_0}\Delta(t^*; \lambda_0, \lambda_1) \cdot \frac{d\lambda_0}{d\theta_0} \\ &= \frac{\exp\{-\lambda_0 t^*\}(\lambda_0 t^* + 1) - 1}{\lambda_0} \qquad (5.7) \\ \frac{\partial}{\partial \theta_1}\Delta(t^*; \lambda_0, \lambda_1) &= \frac{\partial}{\partial \lambda_1}\Delta(t^*; \lambda_0, \lambda_1) \cdot \frac{d\lambda_1}{d\theta_1} \\ &= \frac{1 - \exp\{-\lambda_1 t^*\}(\lambda_1 t^* + 1)}{\lambda_1}. \qquad (5.8) \end{aligned}$$

Note that the parameter estimates $\hat{\theta}_0$ and $\hat{\theta}_1$ are independent. In applying

Theorem 5.1 and using the form of the variance matrix Σ given by Equation (5.5), the variance of the RMST estimate $Var(\Delta(t^*;\hat{\lambda}_0,\hat{\lambda}_1))$, with n patients on each treatment arm, is given by

$$\begin{aligned} & Var(\Delta(t^*;\hat{\lambda}_0,\hat{\lambda}_1)) \\ =& Var(\Delta(t^*;\hat{\theta}_0,\hat{\theta}_1)) \\ =& \frac{1}{n}\begin{bmatrix}\frac{\partial}{\partial\theta_0}\Delta(t^*;\lambda_0,\lambda_1)\\ \frac{\partial}{\partial\theta_1}\Delta(t^*;\lambda_0,\lambda_1)\end{bmatrix}^T\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}\frac{\partial}{\partial\theta_0}\Delta(t^*;\lambda_0,\lambda_1)\\ \frac{\partial}{\partial\theta_1}\Delta(t^*;\lambda_0,\lambda_1)\end{bmatrix} \\ =& \frac{1}{n}\left[\left(\frac{\exp\{-\lambda_0 t^*\}(\lambda_0 t^*+1)-1}{\lambda_0}\right)^2+\left(\frac{1-\exp\{-\lambda_1 t^*\}(\lambda_1 t^*+1)}{\lambda_1}\right)^2\right]. \end{aligned} \tag{5.9}$$

Now consider a test based on the difference in RMST estimates between control and treatment arms. We shall test

$$H_0^{(2)}:\Delta(t^*;\lambda_0,\lambda_1)=0 \quad \text{versus} \quad H_A^{(2)}:\Delta(t^*;\lambda_0,\lambda_1)>0$$

and we shall reject $H_0^{(2)}$ when $\Delta(t^*;\hat{\lambda}_0,\hat{\lambda}_1)>c_2$ for some constant c_2.

For this test with significance level α and power $1-\beta$ when $\Delta(t^*;\lambda_0,\lambda_1)=D_2$, we require the total sample size $n^{(2)}$ given by

$$n^{(2)} \leq \frac{2\,V_2\left(\Phi^{-1}(1-\alpha)+\Phi^{-1}(1-\beta)\right)^2}{D_2^2}$$

where $V_2 = nVar(\Delta(t^*;\hat{\lambda}_0,\hat{\lambda}_1))$ given in Equation (5.9).

Finally, we describe a method for performing an analysis using a non-parametric RMST estimate. The estimate $\hat{\Delta}(t^*;\lambda_0,\lambda_1)$ is the Kaplan-Meier estimate for $\Delta(t^*;\lambda_0,\lambda_1)$. One should note the difference in notation between the parametric estimate $\Delta(t^*;\hat{\lambda}_0,\hat{\lambda}_1)$ and the non-parametric estimate $\hat{\Delta}(t^*;\lambda_0,\lambda_1)$. The non-parametric estimate for the difference in RMST and the variance of the estimate are given by

$$\begin{aligned} \hat{\Delta}(t^*;\lambda_0,\lambda_1) &= \mathbb{E}[\min(F_i|Z_i=1,t^*)]-\mathbb{E}[\min(F_i|Z_i=0,t^*)] \\ Var\left(\hat{\Delta}(t^*;\lambda_0,\lambda_1)\right) &= \mathbb{E}[\min(F_i|Z_i=1,t^*)^2]-\mathbb{E}[\min(F_i|Z_i=0,t^*)^2]-\Delta(t^*;\lambda_0,\lambda_1)^2. \end{aligned}$$

The estimates $\hat{\Delta}(t^*;\lambda_0,\lambda_1)$ and $Var(\hat{\Delta}(t^*;\lambda_0,\lambda_1)$ are found using simulation with a large sample of patients. It is possible to calculate these expressions using simulation because, in this example, there is no censoring.

For this non-parametric Kaplan-Meier analysis, we shall test

$$H_0^{(3)} : \Delta(t^*; \lambda_0, \lambda_1) \leq 0 \quad \text{versus} \quad H_A^{(3)} : \Delta(t^*; \lambda_0, \lambda_1) > 0$$

and we shall reject $H_0^{(3)}$ when $\hat{\Delta}(t^*; \lambda_0, \lambda_1) > c_3$ for some constant c_3.

The total sample size $n^{(3)}$ required for this test with significance level α and power $1 - \beta$ when $\Delta(t^*; \lambda_0, \lambda_1) = D_3$ is given by

$$n^{(3)} = \frac{2\, V_3 \left(\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\right)^2}{D_3^2}$$

where $V_3 = nVar(\hat{\Delta}(t^*; \lambda_0, \lambda_1)$.

In summary, we have proposed three hypothesis tests and found an expression for the sample size for each of these tests. Each sample size calculation has the following structure

$$n^{(i)} = \frac{2\, V_i \left(\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\right)^2}{D_i^2}.$$

In Table 5.1.1, a comparison of these sample sizes is made. For all calculations we have chosen significance level $\alpha = 0.025$ and power $1 - \beta = 0.9$. The range of t^* values from 0.8 to 3.2 represents the 50^{th} and 95^{th} percentiles of the survival distributions under H_A when $D_1 = \theta_0 - \theta_1 = 0.3$. For the case $t^* = \infty$, the restricted mean survival time is equivalent to the mean survival. It is important to note that the calculation of the non-parametric RMST estimate with $t^* = \infty$ is only possible because there is no censoring in this example. Further, in some clinical trials it might also not be possible to calculate the Kaplan-Meier estimate $\hat{\Delta}(t^*; \lambda_0, \lambda_1)$ for $t^* = 3.2$ in the presence of censoring. This is because the maximum uncensored event time might be less than the value $t^* = 3.2$.

t^*	θ_0	θ_1	D_1	V_1	$n^{(1)}$	D_2	V_2	$n^{(2)}$	D_3	V_3	$n^{(3)}$
0.8	0	-0.35	0.35	2	171.5	0.06	0.06	173.1	0.06	0.15	420.9
1.6	0	-0.35	0.35	2	171.5	0.16	0.42	169.6	0.16	0.65	262.5
2.4	0	-0.35	0.35	2	171.5	0.25	0.99	168.7	0.25	1.24	212.4
3.2	0	-0.35	0.35	2	171.5	0.31	1.56	169.5	0.31	1.78	191.4
Inf	0	-0.35	0.35	2	171.5	0.42	3.01	180.3	0.42	3.00	181.0
0.8	0	-0.30	0.30	2	233.5	0.05	0.06	234.9	0.05	0.15	565.9
1.6	0	-0.30	0.30	2	233.5	0.14	0.43	231.4	0.14	0.65	354.0
2.4	0	-0.30	0.30	2	233.5	0.21	0.99	230.7	0.21	1.23	286.6
3.2	0	-0.30	0.30	2	233.5	0.26	1.54	231.6	0.27	1.74	260.0
Inf	0	-0.30	0.30	2	233.5	0.35	2.82	242.3	0.35	2.82	235.3
0.8	0	-0.25	0.25	2	336.2	0.04	0.06	337.5	0.04	0.15	806.2
1.6	0	-0.25	0.25	2	336.2	0.12	0.43	334.1	0.12	0.64	490.5
2.4	0	-0.25	0.25	2	336.2	0.18	0.99	333.5	0.17	1.22	419.7
3.2	0	-0.25	0.25	2	336.2	0.22	1.52	334.5	0.22	1.70	365.5
Inf	0	-0.25	0.25	2	336.2	0.28	2.65	345.0	0.28	2.65	344.0
0.8	0	-0.20	0.20	2	525.4	0.04	0.07	526.5	0.04	0.15	1239.5
1.6	0	-0.20	0.20	2	525.4	0.09	0.44	523.1	0.09	0.64	781.2
2.4	0	-0.20	0.20	2	525.4	0.14	0.99	522.6	0.14	1.20	614.1
3.2	0	-0.20	0.20	2	525.4	0.17	1.50	523.8	0.18	1.65	566.7
Inf	0	-0.20	0.20	2	525.4	0.22	2.49	534.1	0.22	2.50	539.7

Table 5.1.1: Comparing the efficiency of three methods for the proportional hazards model.

For this example we know, by the Neyman-Pearson Lemma, that the test based on maximum likelihood estimates $\hat{\theta}_0$ and $\hat{\theta}_1$ (or the equivalent test based on $\hat{\lambda}_0$ and $\hat{\lambda}_1$) is optimal. For the test based on differences in parametric RMST estimates, the sample size $n^{(2)}$ does not change much with t^* and is not often much higher than $n^{(1)}$. This shows that this test is efficient. Surprisingly, for some values e.g $t^* = 2.4$, we have that $n^{(2)} < n^{(1)}$. This is because the Delta method is approximate and a test with $n^{(2)} < n^{(1)}$ will be under-powered. The sample size $n^{(3)}$ is heavily influenced by the choice of t^*, and much higher than the value of $n^{(1)}$ especially when t^* is small. This indicates that non-parametric methods are much less efficient for moderate values of t^*. However, there are fewer model assumptions made when using non-parametric analyses.

5.1.8 | Example: Non-proportional hazards

We now perform a similar analysis for a parametric model that does not follow the proportional hazards assumption. This is a common reason for using the restricted mean survival time as an analysis method. For this model we will consider a test based on the maximum likelihood estimate compared with the tests based on parametric and non-parametric RMST. The analysis shows that the sample size required for an RMST analysis is largely dependent on the value of t^* chosen.

Consider a survival model where Z_i is the treatment indicator that patient i receives the new treatment. The hazard rate is given by

$$h(t|Z_i) = \exp\{\beta_0 + (\beta_1 - \beta_2 Z_i)t\}. \tag{5.10}$$

We shall follows the steps in Section 2.2 to derive the survival function for this model. The cumulative hazard function is

$$\begin{aligned} H(t|Z_i) &= \int_0^t \exp\{\beta_0 + (\beta_1 - \beta_2 Z_i)t\} \\ &= \frac{\exp\{\beta_0\}}{\beta_1 - \beta_2 Z_i} \left[\exp\{(\beta_1 - \beta_2 Z_i)t\} - 1\right]. \end{aligned}$$

This leads to the survival function

$$\begin{aligned} S(t|Z_i) &= \exp\left(-H(t|Z_i)\right) \\ &= \exp\left(\frac{\exp\{\beta_0\}}{\beta_1 - \beta_2 Z_i}\left[1 - \exp\{(\beta_1 - \beta_2 Z_i)t\}\right]\right). \end{aligned} \tag{5.11}$$

In this parametric model, β_2 is the parameter that governs the treatment effect as this parameter describes how survival differs between the two treatment arms. Clearly, the hazards are not proportional as the hazard ratio for the treatment group versus control group is equal to $\exp\{\beta_2 t\}$ which is a function of time t. However, for $\beta_2 = 0$ we have equivalence in survival curves between treatment and control groups. Let $\hat{\beta}_2$ be the maximum likelihood estimate for β_2. Then we shall test the hypothesis

$$H_0^{(1)} : \beta_2 \leq 0 \quad \text{versus} \quad H_A^{(1)} : \beta_2 > 0$$

and we shall reject $H_0^{(1)}$ when $\hat{\beta}_2 = c_1$ for some constant c_1.

The sample size calculation for this one-sided hypothesis test requires knowledge of the value of $Var(\hat{\beta}_2)$. We now describe a method for estimating this value, which uses maximum likelihood estimation and a very large sample size. Suppose that $F_1, \ldots, F_n$ are time-to-failure random variables for patients $1, \ldots, n$ which are distributed according to the hazard rate given in equation (5.10) and let the observed event times be $t_1, \ldots, t_n$. For simplicity, suppose there is no censoring and therefore the log-likelihood function is given by

$$\begin{aligned}\ell(\boldsymbol{\beta}; t_1, \ldots, t_n) &= \sum_{i=1}^{n} \log\{f(t_i|\boldsymbol{\beta})\} \\ &= \sum_{i=1}^{n} \log\{\lambda(t_i|Z_i)\} + \log\{S(t|Z_i)\} \\ &= \sum_{i=1}^{n} \beta_0 + (\beta_1 - \beta_2 Z_i)t_i + \frac{\exp\{\beta_0\}}{\beta_1 - \beta_2 Z_i}\left[1 - \exp\{(\beta_1 - \beta_2 Z_i)t_i\}\right].\end{aligned}$$

In Section 3.1, we presented the asymptotic distribution for maximum likelihood estimates. In applying these results, we shall substitute the vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$ for $\boldsymbol{\theta}$ and the data $t_1, \ldots, t_n$ shall be substituted for $\mathbf{x}_n^{(k)}$. The variance-covariance matrix for the the MLEs is the inverse of the Fisher information matrix in Equation (3.6). Therefore, the variance covariance matrix for $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)^T$ is given by

$$\Sigma = \frac{1}{n}\left(-\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\ell(\boldsymbol{\beta}; t_1, \ldots, t_n)\right)^{-1} \tag{5.12}$$

and this matrix Σ can be accurately estimated by simulation using a large sample of $n = 10^4$ patients.

For this analysis, we are only interested in the parameter β_2 and hence we take $Var(\hat{\beta}_2) = \Sigma_{33}$. For the hypothesis test with significance level α and power $1 - \beta$ when $\beta_2 = D_1$, we require a sample size

$$n^{(1)} = \frac{V_1\left(\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\right)^2}{D_1^2} \tag{5.13}$$

where $V_1 = n\Sigma_{33}$ and Σ is given in Equation (5.12).

Next, consider the RMST analysis based on maximum likelihood estimates for the parametric model. The difference in RMST is

$$\begin{aligned}\Delta(t^*; \boldsymbol{\beta}) = \int_0^{t^*} &\exp\left(\frac{\exp\{\beta_0\}}{\beta_1}\left[1 - \exp\{\beta_1 t\}\right]\right) \\ &- \exp\left(\frac{\exp\{\beta_0\}}{\beta_1 - \beta_2}\left[1 - \exp\{(\beta_1 - \beta_2)t\}\right]\right) dt. \end{aligned} \tag{5.14}$$

The parametric estimate $\Delta(t^*; \hat{\boldsymbol{\beta}})$ is found by substituting the MLE $\hat{\boldsymbol{\beta}}$ into equation (5.14) and can be evaluated using numerical integration. We shall test the hypothesis

$$H_0^{(2)} : \Delta(t^*; \boldsymbol{\beta}) \leq 0 \quad \text{versus} \quad H_A^{(2)} : \Delta(t^*; \boldsymbol{\beta}) > 0$$

and we shall reject $H_0^{(2)}$ when $\Delta(t^*; \hat{\boldsymbol{\beta}}) > c_2$ for some constant c_2.

We shall apply the Delta method of Theorem 5.1 where the MLE $\hat{\boldsymbol{\beta}}$ is substituted for $\hat{\boldsymbol{\theta}}$. Let Σ be the covariance matrix of $\hat{\boldsymbol{\beta}}$ in Equation (5.12). Then, the variance of the parametric estimate for the difference in RMST is

$$Var(\Delta(t^*; \hat{\boldsymbol{\beta}})) = \frac{1}{n} \left[\frac{\partial}{\partial \boldsymbol{\beta}} \Delta(t^*; \boldsymbol{\beta}) \right]^T \Sigma \left[\frac{\partial}{\partial \boldsymbol{\beta}} \Delta(t^*; \boldsymbol{\beta}) \right]. \tag{5.15}$$

To test this hypothesis with significance level α and power $1-\beta$ when $\Delta(t^*; \boldsymbol{\beta}) = D_2$, we require the sample size

$$n^{(2)} = \frac{V_2 \left(\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\right)^2}{D_2^2}$$

where $V_2 = nVar(\Delta(t^*; \hat{\boldsymbol{\beta}}))$ given in Equation (5.15).

Finally, we shall find the sample size required when we perform an analysis using a non-parametric Kaplan-Meier RMST estimate. The non-parametric estimate $\hat{\Delta}(t^*; \boldsymbol{\beta})$ is found using similar methods to the exponential distribution example. Since there is no censoring, $\hat{\Delta}(t^*; \boldsymbol{\beta})$ and $Var(\hat{\Delta}(t^*; \boldsymbol{\beta}))$ are calculated using simulation. These values can be accurately estimated by using a large sample of $n = 10^4$ patients. We shall test the hypothesis

$$H_0^{(3)} : \Delta(t^*; \boldsymbol{\beta}) \leq 0 \quad \text{versus} \quad H_A^{(3)} : \Delta(t^*; \boldsymbol{\beta}) > 0$$

and we shall reject $H_0^{(3)}$ when $\hat{\Delta}(t^*; \boldsymbol{\beta}) > c_3$ for some constant c_3.

For this test with significance level α and power $1 - \beta$ when $\Delta(t^*; \boldsymbol{\beta}) = D_3$, we require the sample size

$$n^{(3)} = \frac{V_3 \left(\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\right)^2}{D_3^2}$$

where $V_3 = nVar(\hat{\Delta}(t^*; \boldsymbol{\beta}))$.

We shall now compare the sample sizes $n^{(1)}, n^{(2)}$ and $n^{(3)}$. For this comparison, we have chosen significance level $\alpha = 0.025$ and power $1-\beta = 0.9$. Table 5.1.2 shows the results. For this example, the range of t^* is from 3 years to 8 years. These values are roughly equal to the 60^{th} and 99^{th} percentiles of the time-to-event observations under H_A when $D_1 = \beta_2 = 0.1$. Similarly to Section 5.1.7, there is no censoring in this example so non-parametric estimates for $\hat{\Delta}(t^*;\boldsymbol{\beta})$ can be calculated for all values of t^*. However for models with censoring, calculation of $\hat{\Delta}(t^*;\boldsymbol{\beta})$ is likely to be unobtainable for $t^* > 6$ since 6 years is roughly equal to the 95^{th} percentile in each model.

$\boldsymbol{t^*}$	β_2	D_1	V_1	$n^{(1)}$	D_2	V_2	$n^{(2)}$	D_3	V_3	$n^{(3)}$
3	0.08	0.08	0.42	697.6	0.06	0.24	697.7	0.06	2.04	6542.5
4	0.08	0.08	0.42	697.6	0.12	0.90	698.3	0.11	3.50	2810.3
5	0.08	0.08	0.42	697.6	0.18	2.06	695.9	0.17	4.76	1681.1
6	0.08	0.08	0.42	697.6	0.22	3.32	692.7	0.22	5.57	1177.4
7	0.08	0.08	0.42	697.6	0.25	4.22	692.9	0.26	5.99	953.1
8	0.08	0.08	0.42	697.6	0.27	4.68	697.8	0.26	6.13	940.1
Inf	0.08	0.08	0.42	697.6	0.27	4.92	707.2	0.27	6.19	897.9
3	0.10	0.10	0.42	436.2	0.07	0.23	436.3	0.08	2.05	3774.5
4	0.10	0.10	0.42	436.2	0.15	0.88	437.1	0.14	3.54	1882.0
5	0.10	0.10	0.42	436.2	0.22	2.03	435.0	0.22	4.85	1062.8
6	0.10	0.10	0.42	436.2	0.28	3.31	431.7	0.28	5.73	764.6
7	0.10	0.10	0.42	436.2	0.32	4.30	431.3	0.32	6.21	625.9
8	0.10	0.10	0.42	436.2	0.34	4.87	435.4	0.35	6.44	567.9
Inf	0.10	0.10	0.42	436.2	0.35	5.27	446.2	0.35	6.54	573.5
3	0.12	0.12	0.41	297.2	0.09	0.22	297.2	0.09	2.05	2570.1
4	0.12	0.12	0.41	297.2	0.17	0.86	298.1	0.17	3.57	1243.0
5	0.12	0.12	0.41	297.2	0.27	2.00	296.3	0.27	4.92	694.4
6	0.12	0.12	0.41	297.2	0.34	3.31	292.9	0.34	5.88	526.3
7	0.12	0.12	0.41	297.2	0.40	4.38	291.9	0.40	6.46	432.5
8	0.12	0.12	0.41	297.2	0.42	5.07	295.0	0.42	6.72	392.0
Inf	0.12	0.12	0.41	297.2	0.44	5.70	307.3	0.43	6.96	386.7

Table 5.1.2: Comparing the efficiency of three method for the non-proportional hazards model.

The results of the non-proportional hazards model are similar to the results from the proportional hazards model; the test based on differences in parametric RMST estimates is efficient as $n^{(2)}$ is only slightly higher than $n^{(1)}$ in most cases. Again, when $n^{(2)} < n^{(1)}$, this is because the Delta method is only approximate and the tests are under-powered for these cases. The test based on the difference in non-parametric RMST estimates is highly inefficient as $n^{(3)}$ is much greater than $n^{(1)}$ in all cases. The efficiency loss is greater for this non-proportional hazards model than for the exponential distributions example.

We conclude that a test based on differences in parametric RMST estimates is appropriate when the fitted model is correct and this test is close to optimal efficiency. This test is invariant to the value of t^*. The test based on differences in non-parametric RMST estimates is highly dependent on t^* and is inefficient. However, this test requires fewer model assumptions.

5.2 | Joint modelling

5.2.1 | Joint model

We consider the same form of joint model as presented in Section 4.1, but now we add a treatment effect to the model for the biomarker value. In this section, we shall present the joint model, discuss model fitting by maximum likelihood using the Expectation Maximisation (EM) algorithm, and present some asymptotic results for the RMST estimate. Then, we shall be equipped to perform both fixed sample and group sequential trials based on the RMST estimate for this joint model.

Suppose that $X_i(t)$ is the true value of the biomarker at time t for subject i and that $W_i(t)$ is the observed value of the biomarker at time t for patient i. Let $Z_i = \mathbb{I}\{\text{patient } i \text{ receives the new treatment}\}$ be the indicator function for treatment. Then the longitudinal model takes the form

$$\begin{aligned} W_i(t) &= b_{0i} + (b_{1i} + b_2 Z_i)t + \epsilon_i(t) \\ &= X_i(t) + \epsilon_i(t) \end{aligned}$$

where $\mathbf{b}_i = (b_{i0}, b_{i1})$ is a vector of patient specific random effects, b_2 is a fixed treatment effect and $\epsilon_i(t)$ is the measurement error. The purpose of b_2 is to ensure that the means of the slope of the longitudinal trajectory differ between treatment groups.

For model fitting, we must impose some distributional assumptions upon this model. We require that the measurement errors in the longitudinal data independent and identically distributed for each patient. Suppose that the biomarker for patient i is measured at times $t_{i1}, \dots t_{im_i}$, then $\epsilon_i(t_{ij})|\mathbf{b}_i \sim N(0, \sigma^2)$ for $j = 1, \dots, m_i$ and $\epsilon_i(t)$ and $\epsilon_i(t')$ are independent for $t \neq t'$. We shall assume that the random effects, $\mathbf{b}_1, \mathbf{b}_2$, are from a Gaussian distribution, specifically

$$\begin{bmatrix} b_{i0} \\ b_{i1} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_0 \\ \mu_1 \end{bmatrix}, \begin{bmatrix} \phi_0^2 & 0 \\ 0 & \phi_1^2 \end{bmatrix} \right).$$

This allows use of the Gauss-Hermite quadrature rule for efficient computation of maximum likelihood estimates. There is no immediately obvious reason to suggest that the random effects b_{i0} and b_{i1} should be correlated and for this reason, we have assumed independence. This assumption means that there is one less parameter to estimate and simplifies the calculation of the maximum likelihood estimate. However, if there is reason to believe that the random effects are correlated then the methodology can be readily extended. In Section 2.3.2, we described how the Gauss-Hermite quadrature rule can be used to evaluate integrals over random effects in both cases where the random effects are independent and correlated.

For the survival endpoint, the longitudinal data is modelled as a time-varying covariate. In the general model, the hazard function is

$$h_i(t) = h_0(t)\exp\{\gamma X_i(t) + \eta^T Z_i\}, \tag{5.16}$$

where $h_0(\cdot)$ is the baseline hazard function. The longitudinal data has corresponding scalar coefficient γ. The remaining covariates for patient i are given by the $p \times 1$ column vector Z_i which has corresponding coefficient vector η of length p. For the example in this chapter, we consider a single patient coefficient which is the treatment indicator $Z_i = \mathbb{I}\{\text{patient } i \text{ receives the new treatment}\}$. Therefore, the parameter η is a scalar coefficient and summarises the treatment effect that directly affects survival. In summary, the joint model has the form

$$W_i(t) = b_{0i} + (b_{1i} + b_2 Z_i)t + \epsilon_i(t) \tag{5.17}$$

$$h_i(t) = h_0(t)\exp\{\gamma(b_{0i} + (b_{1i} + b_2 Z_i)t) + \eta Z_i\}. \tag{5.18}$$

To perform and analyse a trial using the joint model, the data that must be collected for each individual $i = 1, \dots, n$ is the vector $\{W_i(t_{ij}), j = 1, \dots, m_i\}, Z_i, t_i, \delta_i)$ where

- $(W_i(t_{i1}), \dots, W_i(t_{im_i}))$ are biomarker measurements,
- $Z_i = \mathbb{I}\{\text{patient } i \text{ receives the new treatment}\}$,
- t_i is the event time,
- $\delta_i = \mathbb{I}\{F_i \leq C_i\}$ is the indicator function for censoring, so that $\delta_i = 1$ implies an exact observation.

One should note the difference between t_i and t_{ij}, where t_i is the survival event time and t_{ij} represents a longitudinal observation time.

5.2.2 | Parameter estimation

To find an estimate for the difference in restricted mean survival times between treatment arms for this model, we must have a method for estimating all the parameters of the model. We shall assume that this model is fully parametric so that the form of $h_0(t)$ is specified and the parameters in $h_0(t)$ can be estimated.

Let the vector $\boldsymbol{\theta}$ represent all the parameters in the joint model Equations (5.17)-(5.18). The vector $\boldsymbol{\theta}$ is given by $(\mu_0, \mu_1, \phi_0, \phi_1, b_2, \sigma^2, \gamma, \eta)$ and any parameters in the baseline hazard function. Denote the maximum likelihood estimate for $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}$. First, we can find survival functions $S_0(t; \hat{\boldsymbol{\theta}})$ and $S_1(t; \hat{\boldsymbol{\theta}})$ using methods described in Section 2.2 and substituting in $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$. Then, these survival functions are substituted into Equation (5.1) to calculate the difference in RMST, $\Delta(t^*; \hat{\boldsymbol{\theta}})$. We shall discuss how the Expectation Maximisation (EM) algorithm can be used to find maximum likelihood estimates for models with random effects.

In Section 3 of their paper, Tsiatis and Davidian (2004) present an expression for the full likelihood function of a joint model. For the joint model of Equations (5.17)–(5.18), the likelihood function is

$$\prod_{i=1}^{n} \int\int h_i(t_i)^{\delta_i} \exp\left[-\int_0^{t_i} h_i(u)du\right] \frac{1}{(2\pi\sigma^2)^{m_i/2}} \\ \times \exp\left[-\sum_{j=1}^{m_i} \frac{W_i(t_{ij}) - (b_{0i} + (b_{1i} + b_2 Z_i)t)^2}{2\sigma^2}\right] f(b_{0i}, b_{1i}) db_{0i} db_{1i} \qquad (5.19)$$

where

$$h_i(u) = h_0(u) \exp\{\gamma(b_{0i} + (b_{1i} + b_2 Z_i)u) + \eta Z_i\}.$$

It is clear that the computation of this likelihood will be time consuming because of the integration over the random effects b_{0i} and b_{1i} for each $i = 1, \ldots, n$.

Dempster et al. (1977) present an algorithm called the Expectation Maximisation (EM) algorithm which is a method for finding maximum likelihood estimates (MLEs) when there are latent variables (in this case random effects) in the model. The EM algorithm consists of two steps; the E-step, where a function for the expectation of the complete data log-likelihood is found using the current estimate for the parameters, and the M-step, where the parameters that maximise this function are calculated. Note that during the E-step, the complete data log-likelihood function is constructed as if the values of the random effects are known. This is different to the observed data log-likelihood seen in Equation (5.19). These steps are iterated until a local maximum is found. Further, Dempster et al. (1977) discuss the convergence of the EM algorithm in the finite sample case and prove that this algorithm truly returns a maximum likelihood estimate.

Rizopoulos (2012) presents an R package for joint modelling which implements the EM algorithm to find the parameter estimates of the joint model. Further, the R package by Rizopoulos (2012) makes use of the Gauss-Hermite quadrature rule for use in the M-step, which was introduced in Section 2.3.2. Note that using Gauss-

Hermite integration is only appropriate when the random effects are assumed to be normally distributed, however it can reduce computation times dramatically, which is an attractive feature.

5.2.3 | Asymptotic results for $\hat{\boldsymbol{\theta}}$ and RMST

We wish to design a group sequential trial with K analyses. Let $\boldsymbol{\theta}$ be a $p \times 1$ vector of parameters in the joint model given by Equation(5.18). Suppose that we have a trial with K analyses and that $\hat{\boldsymbol{\theta}}^{(k)}$ is the vector of estimates found at analysis k for each $k = 1, \dots, K$. These are maximum likelihood estimates found using the EM algorithm. In Section 4.1 we proved the asymptotic distribution for a sequence of estimates that are the solutions to estimating equations. We also showed that for MLEs, the sequence of estimates has the canonical joint distribution. Let Σ_k be the covariance matrix for $\hat{\boldsymbol{\theta}}^{(k)}$ at analysis k. Then, for this joint model, we have

1. $(\hat{\boldsymbol{\theta}}^{(1)}, \dots, \hat{\boldsymbol{\theta}}^{(K)})$ is multivariate normally distributed
2. $\hat{\boldsymbol{\theta}}^{(k)} \sim N\left(\boldsymbol{\theta}, \Sigma_k\right)$ for $1 \leq k \leq K$
3. $Cov(\hat{\boldsymbol{\theta}}^{(k_1)}, \hat{\boldsymbol{\theta}}^{(k_2)}) = \Sigma_{k_2}$ for $1 \leq k_1 \leq k_2 \leq K$.

At analysis k, the difference in restricted mean survival times is estimated by $\Delta(t^*; \hat{\boldsymbol{\theta}}_k)$. We aim to show that the sequence $\Delta(t^*; \hat{\boldsymbol{\theta}}_1), \dots, \Delta(t^*; \hat{\boldsymbol{\theta}}_K)$ has the canonical joint distribution Definition 2.2. We have previously seen in Section 5.1.4 that the difference in RMST estimates is asymptotically normally distributed. Therefore, it is appropriate to perform a fixed sample test using RMST methods. We now extend this theory to show that a GST design is also possible using the RMST framework.

We follow the method in Section 2.2 to calculate the survival functions. The survival functions can then be substituted into Equation (5.1) to calculate an estimate for the difference in RMST. We begin with the hazard rate for the joint model, and in order to find a difference in RMST between treatment groups we shall use different notation for the hazard function. The hazard function in Equation (5.18) is

$$h_i(t; \boldsymbol{\theta}, Z_i, b_{0i}, b_{1i}) = h_0(t) \exp\{\gamma(b_{0i} + (b_{1i} + b_2 Z_i)t) + \eta Z_i\}.$$

The cumulative hazard function is given by

$$H_i(t; \boldsymbol{\theta}, Z_i, b_{0i}, b_{1i}) = \int_0^t h_0(u) \exp\{\gamma(b_{0i} + (b_{1i} + b_2 Z_i)u) + \eta Z_i\} du.$$

Let $S_0(t;\boldsymbol{\theta})$ and $S_1(t;\boldsymbol{\theta})$ be the survival functions, integrated over the random effects b_0 and b_1, for patients on the control and treatment arms respectively. Following the methods in Section 2.2, the survival functions are given by

$$\begin{aligned}
S_0(t) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\{-H_i(t;\boldsymbol{\theta},0,b_{0i},b_{1i})\}\, db_{0i}\, db_{1i} \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left\{-\int_0^t h_0(u)\exp\{\gamma(b_{0i}+b_{1i}u)\}\, du\right\} db_{0i}\, db_{1i} \\
S_1(t) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\{-H_i(t;\boldsymbol{\theta},1,b_{0i},b_{1i})\}\, db_{0i}\, db_{1i} \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left\{-\int_0^t h_0(u)\exp\{\gamma(b_{0i}+(b_{1i}+b_2)u)+\eta\}\, du\right\} db_{0i}\, db_{1i}
\end{aligned}$$

Then the difference in restricted mean survival times is

$$\Delta(t^*;\boldsymbol{\theta}) = \int_0^{t^*} [S_1(t;\boldsymbol{\theta}) - S_0(t;\boldsymbol{\theta})]dt.$$

Gauss-Hermite integration can be used to efficiently calculate the integrals over b_0 and b_1.

The information levels, $\mathcal{I}_1,\ldots,\mathcal{I}_K$ are the reciprocals of the variance of $\Delta(t^*;\hat{\boldsymbol{\theta}}^{(1)}),\ldots,\Delta(t^*;\hat{\boldsymbol{\theta}}^{(K)})$. For each analysis $k=1,\ldots,K$ we have the $p\times p$ variance matrix of the the parameter estimates given by $\Sigma_k = Var(\hat{\boldsymbol{\theta}}^{(k)})$ and the information for $\Delta(t^*;\hat{\boldsymbol{\theta}}^{(k)})$ is

$$\mathcal{I}_k^{-1} = \frac{1}{n}\left[\left.\frac{\partial\Delta(t^*;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(k)}}\right]^T \Sigma_k \left[\left.\frac{\partial\Delta(t^*;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(k)}}\right]. \tag{5.20}$$

We now show that the covariance structure of the canonical joint distribution holds. In the following proof, we use the Taylor expansion theory which requires some regularity conditions to hold. We assume that Conditions 3.2 hold where the function $\mathbf{G}_n(\boldsymbol{\theta},\mathbf{x}_n^{(k)})$ represents the function $\Delta(t^*;\hat{\boldsymbol{\theta}}^{(k)})$, dependent on data $\mathbf{x}_n^{(k)}$.

Theorem 5.2. *Let $\boldsymbol{\theta}$ be a $p\times 1$ vector of parameters in model* (5.18) *and let $\boldsymbol{\theta_0}$ be the true value of $\boldsymbol{\theta}$. Suppose that $\hat{\boldsymbol{\theta}}^{(k)}$ is the maximum likelihood estimate for $\boldsymbol{\theta}$ found at analysis k of a group sequential trial with K analyses. Let the estimated difference in restricted mean survival times at analysis k be $\Delta(t^*;\hat{\boldsymbol{\theta}}^{(k)})$ and let $\mathcal{I}_1,\ldots,\mathcal{I}_K$ be the information levels for $\Delta(t^*;\hat{\boldsymbol{\theta}}^{(k)}),\ldots,\Delta(t^*;\hat{\boldsymbol{\theta}}^{(K)})$ given by Equation* (5.20)*. Then, the canonical joint distribution holds for the sequence of estimates $\Delta(t^*;\hat{\boldsymbol{\theta}}^{(1)}),\ldots,\Delta(t^*;\hat{\boldsymbol{\theta}}^{(K)})$. That is*

1. $\left(\Delta(t^*;\hat{\boldsymbol{\theta}}^{(1)}),\ldots,\Delta(t^*;\hat{\boldsymbol{\theta}}^{(K)})\right)$ *is multivariate normally distributed*

2. $\Delta(t^*; \hat{\boldsymbol{\theta}}^{(k)}) \sim N(\boldsymbol{\theta}, \mathcal{I}_k^{-1})$ *for* $1 \leq k \leq K$

3. $Cov\left(\Delta(t^*; \hat{\boldsymbol{\theta}}^{(k_1)}), \Delta(t^*; \hat{\boldsymbol{\theta}}^{(k_2)})\right) = \mathcal{I}_{k_2}^{-1}$ *for* $1 \leq k_1 \leq k_2 \leq K$.

Proof. The proof of this theorem uses the Taylor expansion of the function $\Delta(t^*; \boldsymbol{\theta})$. First note that the parameter $\boldsymbol{\theta}$ is a $p \times 1$ column vector and that $\boldsymbol{\theta} - \boldsymbol{\theta_0}$ has the same dimension. Further, the function $\Delta(t^*; \boldsymbol{\theta})$ returns a scalar, so that $\partial/\partial\boldsymbol{\theta}\,(\Delta(t^*; \boldsymbol{\theta}))$ will be a $p \times 1$ column vector. Regularity conditions 6 and 7 of conditions 3.2 allow us to perform the following Taylor expansion. The Taylor expansion of $\Delta(t^*; \boldsymbol{\theta})$ around $\boldsymbol{\theta_0}$ is

$$\Delta(t^*; \boldsymbol{\theta}) = \Delta(t^*, \boldsymbol{\theta_0}) + \left[\frac{\partial}{\partial\boldsymbol{\theta}} \Delta(t^*; \boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right]^T (\boldsymbol{\theta} - \boldsymbol{\theta_0})$$

where $\boldsymbol{\theta}^*$ lies on the line segment between $\boldsymbol{\theta_0}$ and $\boldsymbol{\theta}$.

For each $k = 1, \ldots, K$, we shall apply this Taylor expansion at the point $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{(k)}$. Therefore, for each $k = 1, \ldots, K$, we have

$$\Delta(t^*; \hat{\boldsymbol{\theta}}^{(k)}) = \Delta(t^*, \boldsymbol{\theta_0}) + \Delta'(t^*; \boldsymbol{\theta}^{*(k)})^T (\hat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta_0})$$

where $\boldsymbol{\theta}^{*(k)}$ lies on the line segment between $\boldsymbol{\theta_0}$ and $\hat{\boldsymbol{\theta}}^{(k)}$ and $\Delta'(t^*; \boldsymbol{\theta}^{*(k)}) = \partial/\partial\boldsymbol{\theta}\,(\Delta(t^*; \boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{*(k)}})$ is shorthand notation.

For each $k = 1, \ldots, K$, the parameter estimate $\hat{\boldsymbol{\theta}}^{(k)}$ is consistent for $\boldsymbol{\theta_0}$. We have that $\boldsymbol{\theta}^{*(k)}$ lies on the line segment between $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}^{(k)}$ and the difference between $\boldsymbol{\theta}^{*(k)}$ and $\boldsymbol{\theta_0}$ is asymptotically negligible. Therefore for each $k = 1, \ldots, K$ we have approximately

$$\Delta(t^*; \hat{\boldsymbol{\theta}}^{(k)}) = \Delta(t^*, \boldsymbol{\theta_0}) + \Delta'(t^*; \boldsymbol{\theta_0})^T (\hat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta_0}). \tag{5.21}$$

The proof that the first condition holds follows by stacking Equations (5.21) for each $k = 1, \ldots, K$. By property 1 of the canonical joint distribution for the sequence of MLEs $\hat{\boldsymbol{\theta}}^{(1)}, \ldots, \hat{\boldsymbol{\theta}}^{(K)}$, we have that $(\hat{\boldsymbol{\theta}}^{(1)}, \ldots, \hat{\boldsymbol{\theta}}^{(K)})$ is multivariate normally distributed. Then, Slutsky's Theorem can be applied to the vector of stacked equations (see Theorem 3.1 for an example of applying Slutsky's Theorem). This also follows by the multivariate version of the Delta method of Theorem 5.1 given by Doob (1935).

It is clear that the second condition holds by the Delta method of Theorem 5.1 and for each $k = 1, \ldots, K$,

$$\Delta(t^*; \hat{\boldsymbol{\theta}}^{(k)}) \sim N(\boldsymbol{\theta}, \mathcal{I}_k^{-1}).$$

It remains to prove property 3. Using the approximation in Equation (5.21), the covariance is given by

$$\begin{aligned}
&Cov\left(\Delta(t^*,\hat{\boldsymbol{\theta}}^{(k_1)}),\Delta(t^*,\hat{\boldsymbol{\theta}}^{(k_2)})\right)\\
=&Cov\left(\Delta(t^*,\boldsymbol{\theta_0})+\Delta'(t^*;\boldsymbol{\theta_0})^T(\hat{\boldsymbol{\theta}}^{(k_1)}-\boldsymbol{\theta_0}),\Delta(t^*,\boldsymbol{\theta_0})+\Delta'(t^*;\boldsymbol{\theta_0})^T(\hat{\boldsymbol{\theta}}^{(k_2)}-\boldsymbol{\theta_0})\right)\\
=&\Delta'(t^*;\hat{\boldsymbol{\theta}}^{(k_1)})^T Cov\left(\hat{\boldsymbol{\theta}}^{(k_1)},\hat{\boldsymbol{\theta}}^{(k_2)}\right)\Delta'(t^*;\hat{\boldsymbol{\theta}}^{(k_2)}).
\end{aligned}$$

By property 3 of the canonical joint distribution for the sequence of MLEs $\hat{\boldsymbol{\theta}}^{(1)},\ldots,\hat{\boldsymbol{\theta}}^{(K)}$, we have that $Cov(\hat{\boldsymbol{\theta}}^{(k_1)},\hat{\boldsymbol{\theta}}^{(k_2)})=\Sigma_{k_2}$ and we see the result

$$Cov\left(\Delta(t^*,\hat{\boldsymbol{\theta}}^{(k_1)}),\Delta(t^*,\hat{\boldsymbol{\theta}}^{(k_2)})\right)=\Delta'(t^*;\hat{\boldsymbol{\theta}}^{(k_1)})^T\Sigma_{k_2}\Delta'(t^*;\hat{\boldsymbol{\theta}}^{(k_2)})=\mathcal{I}_{k_2}^{-1}.$$

□

We have shown that the canonical joint distribution for the sequence of estimates $\Delta(t^*;\hat{\boldsymbol{\theta}}_1),\ldots,\Delta(t^*;\hat{\boldsymbol{\theta}}_K)$ holds asymptotically and hence, a group sequential trial can be performed using the RMST method. In Section 5.3 we shall perform some simulation studies to show that this is the case in a moderately sized trial.

5.3 | Simulation study of the restricted mean survival time estimates

5.3.1 | Design choices and parameter values for simulation studies

When using the restricted mean survival time as an analysis tool within a clinical trial, there are many choices to be made concerning the design of the trial. One design aspect is the functional form of the baseline hazard function. We propose using a piecewise constant baseline hazard function for computational efficiency and we discuss the number and positions of the knot points for a model that can be used in a group sequential trial. Considering all of these design aspects, we can then calculate the sample size for a fixed sample trial given a desired power. We shall also present some typical parameter values which we use in the simulation of data for analysis of the joint model given by Equation (5.18). Some reasoning for these choices is presented. In Section 5.4 we shall inspect some properties of the fixed and group sequential trials as we vary these parameters. We perform simulation studies to ensure confidence in the distributional results for the estimate of the difference in restricted mean survival times.

For clarity, the joint model is presented again here. The longitudinal data is assumed to follow a random effects model. For patients $i = 1, \dots, n$, let $Z_i = \mathbb{I}\{\text{patient } i \text{ receives the new treatment}\}$ be the indicator function for treatment, then the longitudinal observations are given by

$$W_i(t) = b_{i0} + (b_{1i} + b_2 Z_i)t + \epsilon_i(t)$$

where

$$\begin{bmatrix} b_{i0} \\ b_{i1} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_0 \\ \mu_1 \end{bmatrix}, \begin{bmatrix} \phi_0^2 & 0 \\ 0 & \phi_1^2 \end{bmatrix} \right)$$
$$\epsilon_i(t) | \mathbf{b}_i \sim N(0, \sigma^2).$$

The hazard function for the survival part of the model is given by

$$h_i(t) = h_0(t) \exp\{\gamma(b_{i0} + (b_{1i} + b_2 Z_i)t) + \eta Z_i\}.$$

Throughout this chapter, we have chosen to use a piecewise constant baseline hazard function. The main reason for this is computational efficiency as this results in an analytic solution for integrating the hazard function. This integration is needed to calculate the survival function $S_i(t)$ which is required for both the RMST estimate and its variance. When performing simulation studies with a large number of Monte Carlo replicates, this computational efficiency is an attractive feature.

Similarly to the value of t^*, we believe that the knot points should be chosen based on clinical meaning to reflect changes in the shape of the survival curve, and should be specified during the design of a clinical trial. In the examples, we shall focus on models with a single knot point. This is because of the complications that arise when designing a group sequential trial that uses a piecewise constant baseline hazard function and we shall expand on the choice of knot points in Section 5.3.3. We believe that, although simple, this model is meaningful and reflects many true scenarios regarding survival data. A piecewise constant baseline hazard function with one knot point placed at t_1 is defined by

$$h_0(t) = \begin{cases} c_1 & \text{if } t \leq t_1 \\ c_2 & \text{if } t > t_1 \end{cases}. \tag{5.22}$$

An analytic expression for the cumulative hazard function for the joint model in Equation (5.18) is therefore given in two parts.

For $t \leq t_1$:

$$\begin{aligned} H_i(t) =& \int_0^t c_1 \exp\{\gamma(b_{0i} + (b_{1i} + b_2 Z_i)u) + \eta Z_i\} du \\ =& \frac{c_1 \exp\{\gamma b_{0i} + \eta Z_i\}}{\gamma(b_{1i} + b_2 Z_i)} [\exp\{\gamma t(b_{1i} + b_2 Z_i)\} - 1]. \end{aligned}$$

For $t > t_1$:

$$\begin{aligned} H_i(t) =& \int_0^{t_1} c_1 \exp\{\gamma(b_{0i} + (b_{1i} + b_2 Z_i)u) + \eta Z_i\} du \\ &+ \int_{t_1}^t c_2 \exp\{\gamma(b_{0i} + (b_{1i} + b_2 Z_i)u) + \eta Z_i\} du \\ &= \frac{\exp\{\gamma b_{0i} + \eta Z_i\}}{\gamma(b_{1i} + b_2 Z_i)} [(c_1 - c_2)\exp\{\gamma t_1(b_{1i} + b_2 Z_i)\} + c_2 \exp\{\gamma t(b_{1i} + b_2 Z_i)\} - c_1] \end{aligned}$$

and the survival function for patient i is then

$$S_i(t) = \exp\{-H_i(t)\}, \qquad \text{for } t > 0.$$

We are using the JM package by Rizopoulos (2010) to perform the parameter estimation of this model, which uses the expectation maximisation (EM) algorithm. Alternative options for the baseline hazard function, which are permitted within this package, include Weibull and spline models. If specifying knot points is non desirable, then the user might wish to use a Weibull baseline hazard function. In this case the survival function must be calculated using numerical integration. Alternatively, one could specify a spline function as the baseline hazard function to avoid the jumps in piecewise constant function. The spline baseline hazard however requires numerical integration and specifying knot points.

Similarly to the first model, Equations (4.1)–(4.3) of Chapter 4, we have designed the study with non-informative censoring with roughly 10% of patients being censored. This is done by setting the distribution of the censoring random variable for patient i as $C_i \sim \exp\{\lambda\}$ and λ is chosen using trial and error to attain 10% of observations being censored.

We shall simulate in the case

$$\begin{gathered}(\mu_0, \mu_1) = (3, 1), \phi_0 = 1.2, \phi_1 = 0.25, b_2 = -0.4, \sigma_2 = 1, \\ \gamma = 0.035, \eta = -0.4, c_1 = 5.32, c_2 = 4.43, t_1 = 1, \lambda = 0.022\end{gathered} \tag{5.23}$$

which reflects that we are simulating under H_A. In later sections, we set $\eta = 0$ and $b_2 = 0$ when we simulate under H_0. For notation, let $\boldsymbol{\theta}$ denote the set of all parameters in the model. In Figure 5.3.1, examples of four randomly selected patients' longitudinal trajectories can be seen. This plot shows that the measurement error $\sigma^2 = 1$ is such that the true underlying biomarker trajectory can easily be estimated however there is still a realistic amount of noise surrounding the measurements. In later Sections, we shall increase this value of σ^2 to understand how the model reacts to extremely noisy measurements.

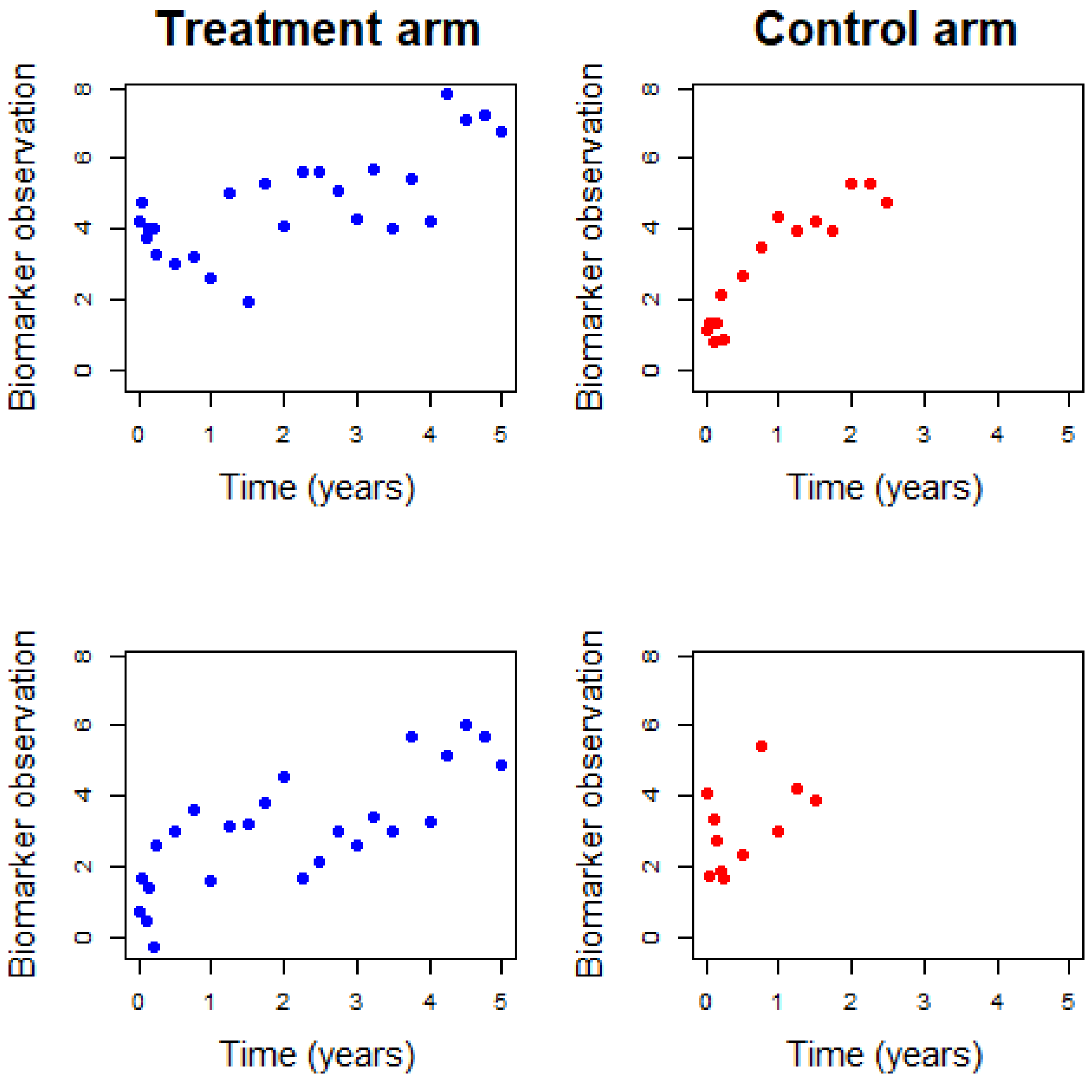


Figure 5.3.1: Longitudinal observations of four randomly selected patients with parameter values (5.23). Blue dots are patients on the treatment arm and red are for patients on control arm.

We can also see from Figure 5.3.1, that there is some variation in the slope terms of the biomarker trajectories. However, all trajectories are increasing to reflect a worsening condition over time and also patients on the treatment arm are not increasing as rapidly. The variance-covariance matrix of the random effects $\mathbf{b}_1, \ldots, \mathbf{b}_n$ determines by how much these intercept and slopes vary, and this variation carries through into the hazard function. Figure 5.3.2 shows the survival function for patients with different underlying biomarker trajectories: a survival function for a patient with mean values of b_0 and b_1 is given by the solid line, and dotted lines show the survival function for patients 1 standard deviation of both b_0 and b_1 above and below their respective means. For this model, we set

$\phi_0 = 1.2$ and $\phi_1 = 0.25$. In the first model of Chapter 4, we simulated under the case $\phi_0 = 3.5$ and $\phi_1 = 2.5$. Therefore, the variance of the random effects terms are not as great as in the first model indicating that patients trajectories are similar within a treatment arm. A similar plot using the AIDS data set in the R package JM written by Rizopoulos (2010) is included for comparison, which shows that the effect of the biomarker heavily outweighs that of the treatment effect in the hazard function. However, the treatment effect is very small in the AIDS data set. Our chosen parameter values are therefore acceptable with respect to the amount of information that comes from the longitudinal data.

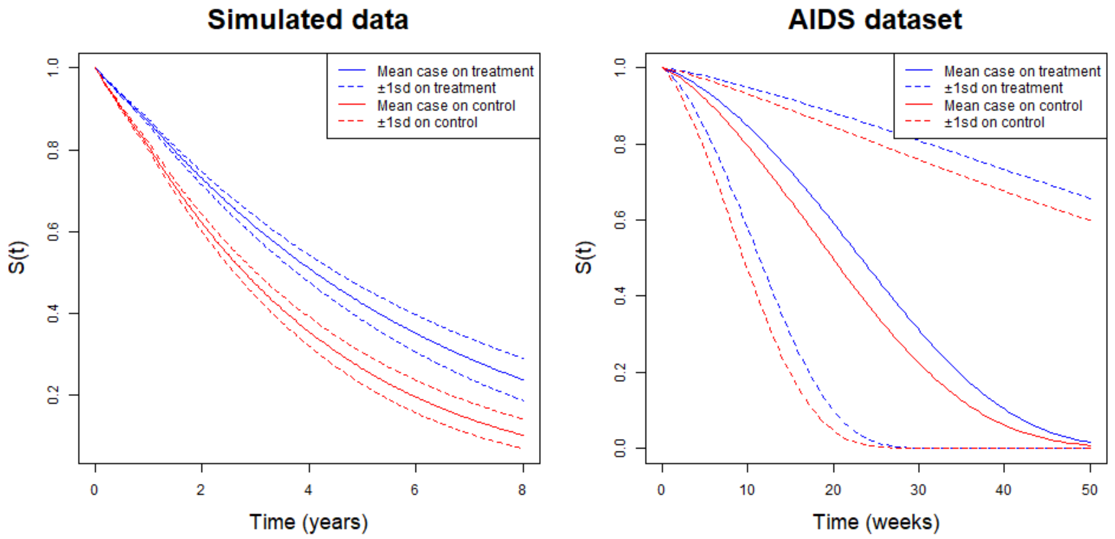


Figure 5.3.2: Survival function for the simulated data and from AIDS data set. The survival function for a patient with mean values of b_0 and b_1 is given by the solid line, and dotted lines show the survival function for patients 1 standard deviation of both b_0 and b_1 above and below their respective means.

Other variables which impact the survival function are the difference in slopes of longitudinal measurements between treatments, b_2, and the coefficient of the longitudinal data, γ. In Section 5.4 we shall vary these parameters to see how properties of the trial change. In Figure 5.3.3 we see the overall distribution of the time-to-event observations. This histogram shows the survival times of 1000 randomly generated patients under this model and parameter choices described above. At the final analysis at 5 years, roughly 60% of events have occurred, as intended. The median survival time for the data set is 3.25 years.

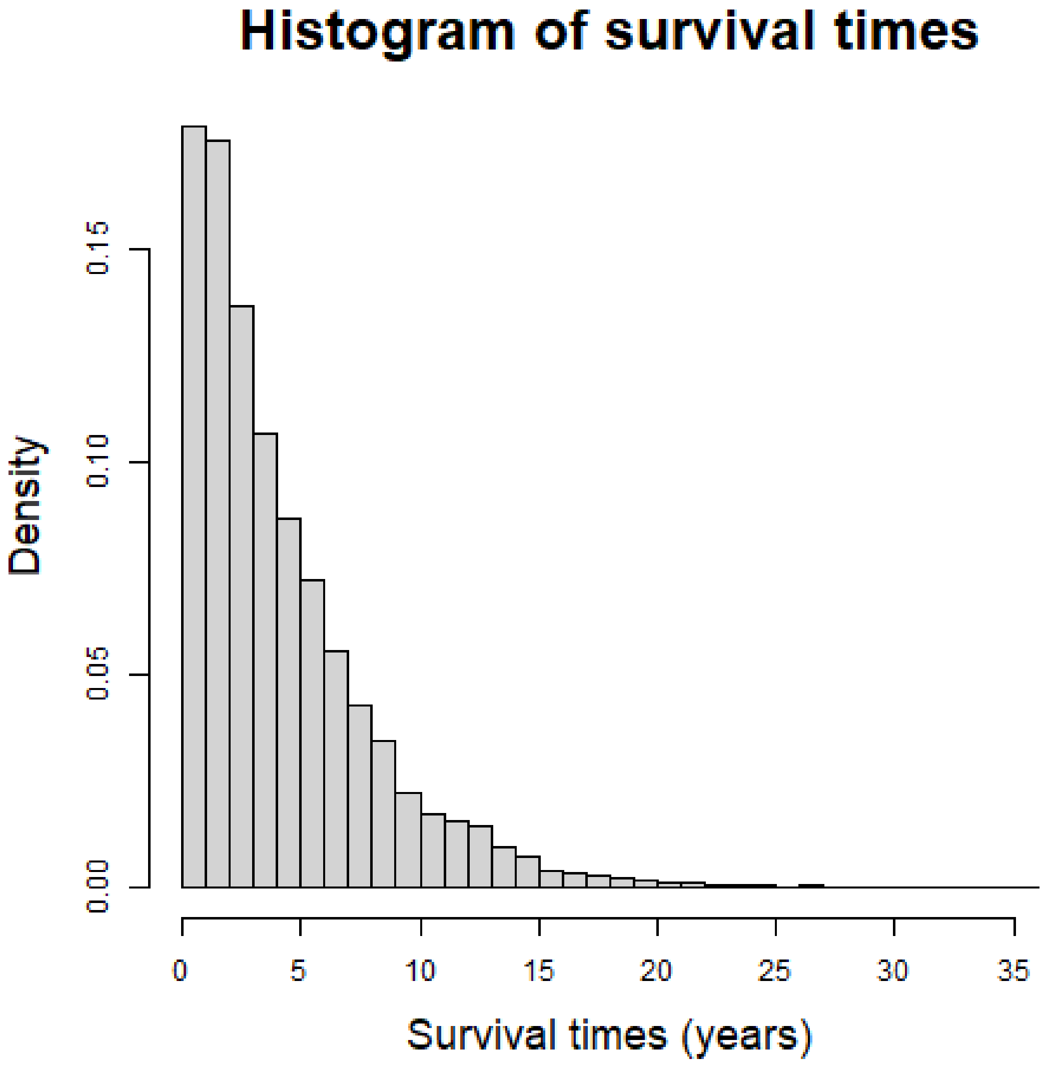


Figure 5.3.3: Histogram of survival times for generated patients on control and treatment arms combined using parameter values (5.23).

5.3.2 | Fixed sample simulations

To ensure that the large sample theoretical results of Section 5.2 hold in a moderately sized trial, we now perform some simulation studies using the parameter values described above. For a fixed sample trial, we are aiming to show that the RMST estimates are asymptotically normally distributed and that the delta method gives an accurate estimate for the variance. We shall also check that the marginal distributions of the RMST estimates in a group sequential trial are asymptotically normally distributed and that the covariances of RMST estimates between analyses have the required canonical joint distribution structure.

To perform a fixed sample clinical trial using the RMST method, we test the hypothesis

$$H_0 : \Delta(t^*; \boldsymbol{\theta}) \leq 0, \qquad H_A : \Delta(t^*; \boldsymbol{\theta}) > 0$$

where $\boldsymbol{\theta}$ is the true set of all parameters in the joint model given by Equation (5.18). We shall simulate under H_0 and H_A with values (5.23). When simulating under H_0 we use $\eta = 0$ and $b_2 = 0$ and we shall use $\eta = -0.4$ and $b_2 = -0.4$ as an example for simulating under H_A.

The main choice when performing an analysis which uses RMST methods, is the choice of t^*. We have previously discussed how this value should be chosen based on clinical meaning and throughout this report, we will consider the 3-year restricted mean survival time difference. The structure for this clinical trial is 2 years recruitment and 3 years follow-up and hence we expect to observe the trajectories of many patients past 3 years. In Section 5.1.6 we discuss that t^* should not exceed the maximum follow-up time but increasing t^* may lead to smaller sample sizes. In the results, Section 5.4, we shall we simulate a subset of clinical trials with $t^* = 5$ to observe the potential sample size reduction from increasing t^* and compare these to the results with $t^* = 3$.

In Section 5.1.4, we proved theoretically using the Delta method, that the estimate of the difference in RMST between treatment arms, $\Delta(t^*; \hat{\boldsymbol{\theta}})$ is normally distributed. Further, if the standardised statistic is considered, this is also approximately normally distributed. We shall simulate standardised estimates centered on zero. These are given by

$$Z(\eta, b_2) = \Big(\Delta(3; \hat{\boldsymbol{\theta}}) - \Delta(3; \boldsymbol{\theta})\Big) \sqrt{\mathcal{I}}.$$

Let $Z_{null} = Z(0,0)$ be the random variable simulated under H_0 with $\eta = 0$ and $b_2 = 0$ and let $Z_{alt} = Z(-0.4, -0.4)$ be the random variable for the standardised statistic simulated under H_A with $\eta = -0.4$ and $b_2 = -0.4$. We can compare the simulated centered statistics with the theoretical distributions

$$Z_{null} \sim N(0,1) \qquad Z_{alt} \sim N(0,1).$$

Similarly to Section 4.3.2, we have chosen to center these statistics on zero so that the mean of the theoretical distribution is know. If the statistics were not centered, then we would need to know the true value of $\mathcal{I}$ in order to find the mean of Z_{null} and Z_{alt}.

A method for choosing a suitable sample size n is described in Section 5.1.5. We simulate a large data set with 10^4 patients under H_A with $\eta = -0.4$ and $b_2 = -0.4$.

The data set is then truncated at time 5 years to reflect end of study censoring in the fixed sample trial. The variance-covariance matrix, Σ, for the full set of parameters, is estimated using this large data set. Using this estimate for Σ and the choice $t^* = 3$, the sample size required for significance level $\alpha = 0.025$ and power $1 - \beta = 0.9$ is given by $n = 460$.

The histograms and QQ-plots in Figure 5.3.4 show 10^4 estimates of Z_{null} and Z_{alt}. Each estimate is found by simulating a fixed sample clinical trial with $n = 460$ patients, calculating the estimates $\Delta(3; \hat{\boldsymbol{\theta}})$ and $Var(\Delta(3; \hat{\boldsymbol{\theta}}))$, and standardising the estimate. It is clear that the estimates are normally distributed. In the histograms, the estimates simulated under the null hypothesis closely match the probability density function of a $N(0, 1)$ distribution which is shown in red and the estimates simulated under the alternative hypothesis, centered on zero, closely match the probability density function of a $N(0, 1)$. Also, the quantiles of the sampled Z values clearly match the quantiles of the theoretical distributions as shown by the QQ-plots.

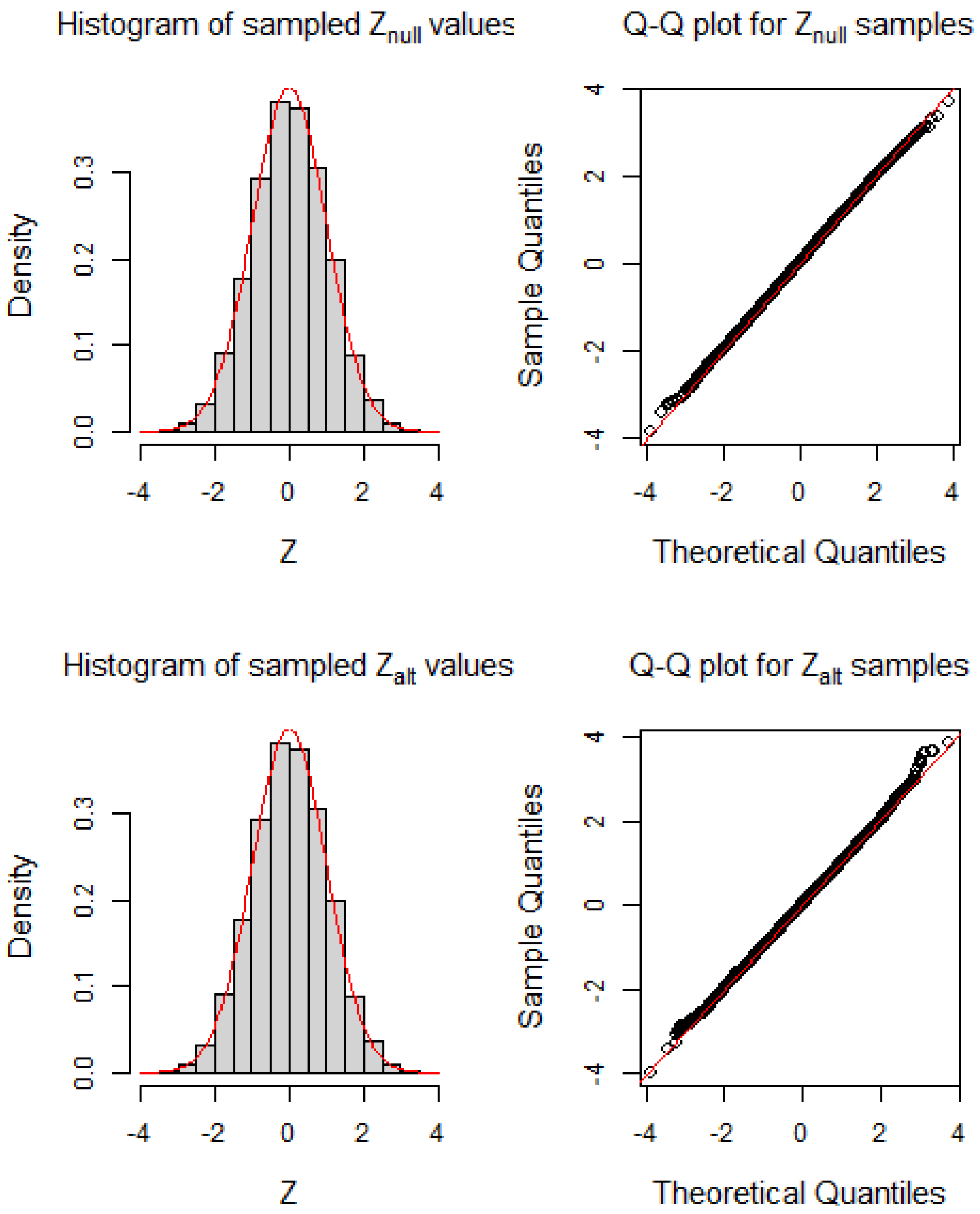


Figure 5.3.4: Fixed sample simulations for standardised statistic under null and alternative hypotheses.

5.3.3 | Group sequential simulations

For a clinical trial which uses the RMST framework, the group sequential trial analysis needs slightly more consideration than the fixed sample trial. The main challenge in designing a group sequential trial is to ensure that parameters are identifiable and also the parameter estimates follow large sample theory. At an early interim analysis, there may be no or very few individuals followed up beyond a certain knot point, and there will be no or very little information on a parameter in $h_0(t)$. We suggest there should be at least 30 events between successive knot points observed at every analysis time. Jennison and Turnbull (1989) investigate the implications of the normal approximation assumption for group sequential tests when a small number of failures occurs. They observe that problems arise when fewer than roughly 25 failures occur at the first interim analysis. Further, there may be information available to estimate parameters in $h_0(t)$, but little or no late follow up to allow checking of model assumptions and therefore, extrapolation of survival function estimates should also be taken into consideration.

In Section 5.4, we shall compare the fixed sample and group sequential designs for the joint model in Equation (5.18). We are interested in comparing the length of the trials in time in order to assess the benefits of potential early stopping in the group sequential trial. The number of patients shall remain constant between fixed and group sequential trials and therefore, to maintain equal power, the time of the analyses shall be varied in the group sequential trial. The probability of stopping before all patients are recruited is low, so the expected sample size on stopping is not likely to be reduced in the GST.

To choose the times of analyses, we first determine the relationship between calendar time and information. The true variance, Σ, of the parameter estimates $\hat{\boldsymbol{\theta}}$ in Equation (5.18) is unknown and therefore the relationship between calendar time and information, $\mathcal{I}$, must be estimated. Figure 5.3.5 provides a visual representation of the relationship and the following steps describe how this relationship is determined for a particular choice of parameter values for $\boldsymbol{\theta}$:

1. A large sample of 5000 patients' event times, biomarker observations, arrival times and censoring times are simulated with the chosen parameter values for $\boldsymbol{\theta}$. This simulation occurs under H_A with particular choices $\eta \neq 0$ and $b_2 \neq 0$.

2. For each time point along the x-axis, this data set is truncated and the information level is calculated using the delta method.

3. Note that this information is calculated using a sample size of 5000 patients and denote this information level as $\mathcal{I}_{5000}$.

4. For the fixed sample trial, we recruit n patients ($n = 460$ for parameter values (5.23), see Section 5.3.2). The asymptotic results of Section 5.1.4 determine the distribution of the estimate $\Delta(t^*; \hat{\boldsymbol{\theta}}_n)$ as $n \to \infty$ and the relationship $\mathcal{I} \propto n$ holds. Therefore, let the final information level for this time point on the x-axis be $\mathcal{I} = n\mathcal{I}_{5000}/5000$.

5. A *log* function is then fit to these corresponding time points and information levels.

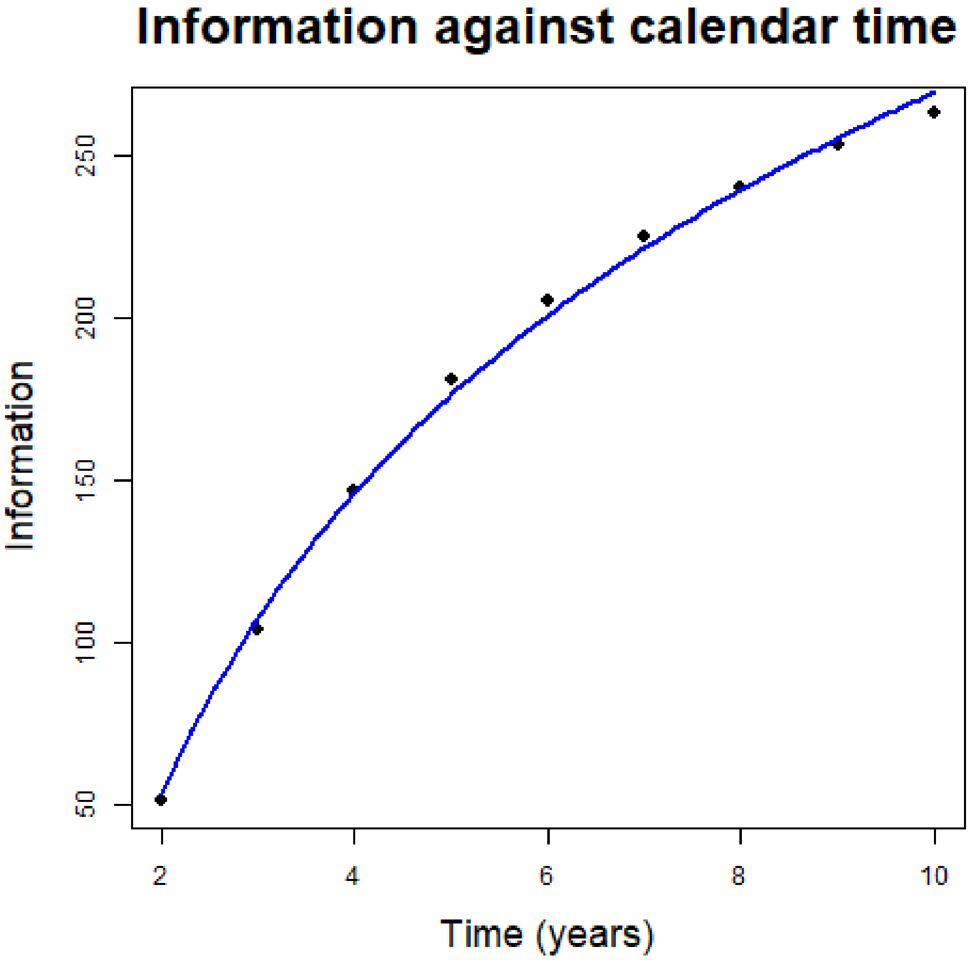


Figure 5.3.5: Information level estimates against calendar time based on a large sample of 5000 patients simulated with parameter values (5.23) for $\boldsymbol{\theta}$.

As previously alluded to, it is important that there are enough observed events occurring in between the knot points and times of interim analyses. This is required so that the large sample theory holds for parameter estimation. Let $\tau_1, \ldots, \tau_K$ be the calendar times of the interim analyses and let t_1 be the time of the knot point in the baseline hazard function. As a reminder, we have chosen to use a model with a single knot point. Suppose that $\tilde{t}_1, \ldots, \tilde{t}_n$ are the survival times of patients $1, \ldots, n$. Note that the calendar time at which the event occurs for patient i is $\tilde{t}_i + a_i$ where a_i is the arrival time of patient i. To ensure that the parameters of the model can be estimated, there must be at least one event between successive knot points and analysis times. That is we require

- $0 < \tilde{t}_1 < \tau_1 < \cdots < \tau_K$
- $\{i : 0 < \tilde{t}_i \leq t_1\} \neq \phi$
- $\{i : t_1 < \tilde{t}_i \leq \tau_1\} \neq \phi$.

For models that use $m > 1$ knot points, this theory can be extended to ensure that all the model parameters are identifiable at the first interim analysis. In particular, the final point becomes $\{i : t_m < \tilde{t}_i \leq \tau_1\} \neq \phi$ and we must also ensure that there are events occurring inbetween successive knot points so that $\{i : t_{j-1} < \tilde{t}_i \leq t_j\} \neq \phi$ for $j = 2, \ldots, m$.

Further, to ensure that the large sample theory holds, there should be a substantial number of events occurring, not just one. For this, we require the size of each of the above sets to be at least 30. We believe that 30 failures is sufficiently large so that the normal approximation of the parameter estimates is accurate based on findings by Jennison and Turnbull (1989).

We now describe a method for designing a group sequential trial and choosing where to place analysis times.

1. Using the steps above or reading off Figure 5.3.5, calculate the value of $\mathcal{I}_f$ for a fixed sample trial that terminates at calendar time 5 years.

2. Using methods described in Section 2.1.3 and error spending functions be $f(t) = \min\{\alpha t^2, \alpha\}$ and $g(t) = \min\{\beta t^2, \beta\}$, calculate the maximum information level $\mathcal{I}_{max}$. This is the value such that the trial has power $1 - \beta$ when $\Delta(t^*; \boldsymbol{\theta}) = \delta$, the final boundary points a_k and b_k are equal and is calculated under the assumption that information levels are equally spaced.

3. Let $\mathcal{I}_1^{(1)} = \mathcal{I}_{max}/K$ and set $\tau_1^{(1)}$ as the calendar time corresponding to this $\mathcal{I}_1^{(1)}$ using the *log* relationship.

4. With the simulated data from step 1 with values chosen for $\boldsymbol{\theta}$ and under H_A with particular choices $\eta \neq 0$ and $b_2 \neq 0$, obtain values $\tilde{t}_1, \ldots, \tilde{t}_{5000}$ and $a_1, \ldots, a_{5000}$.

5. Determine the time at which 30 event times occur in a sample of n patients by finding the $(30/n)$-th percentile of values $\tilde{t}_1 + a_1, \ldots, \tilde{t}_{5000} + a_{5000}$. Call this value $\tau_1^{(2)}$.

6. Let $\tau_1 = \max\{\tau_1^{(1)}, \tau_1^{(2)}\}$ and use the *log* relationship to calculate $\mathcal{I}_1$ as the information level for τ_1.

7. Set $\mathcal{I}_2, \dots, \mathcal{I}_{K-1}$ equally spaced between $\mathcal{I}_1$ and $\mathcal{I}_K$ and find corresponding analysis times $\tau_2, \dots, \tau_{K-1}$ using the *log* relationship.

As an example, we shall simulate data with values (5.23). This reflects that we are simulating under a H_A with $\eta = -0.4$ and $b_2 = -0.4$. For the hypothesis test based on the difference in 3-year restricted mean survival times with $t^* = 3$, we use an error spending test with $K = 5$ analyses. We design the trial with $\mathcal{I}_{max} = 195.2$. The resulting analysis times in months are $22.5, 30, 40, 53, 70$.

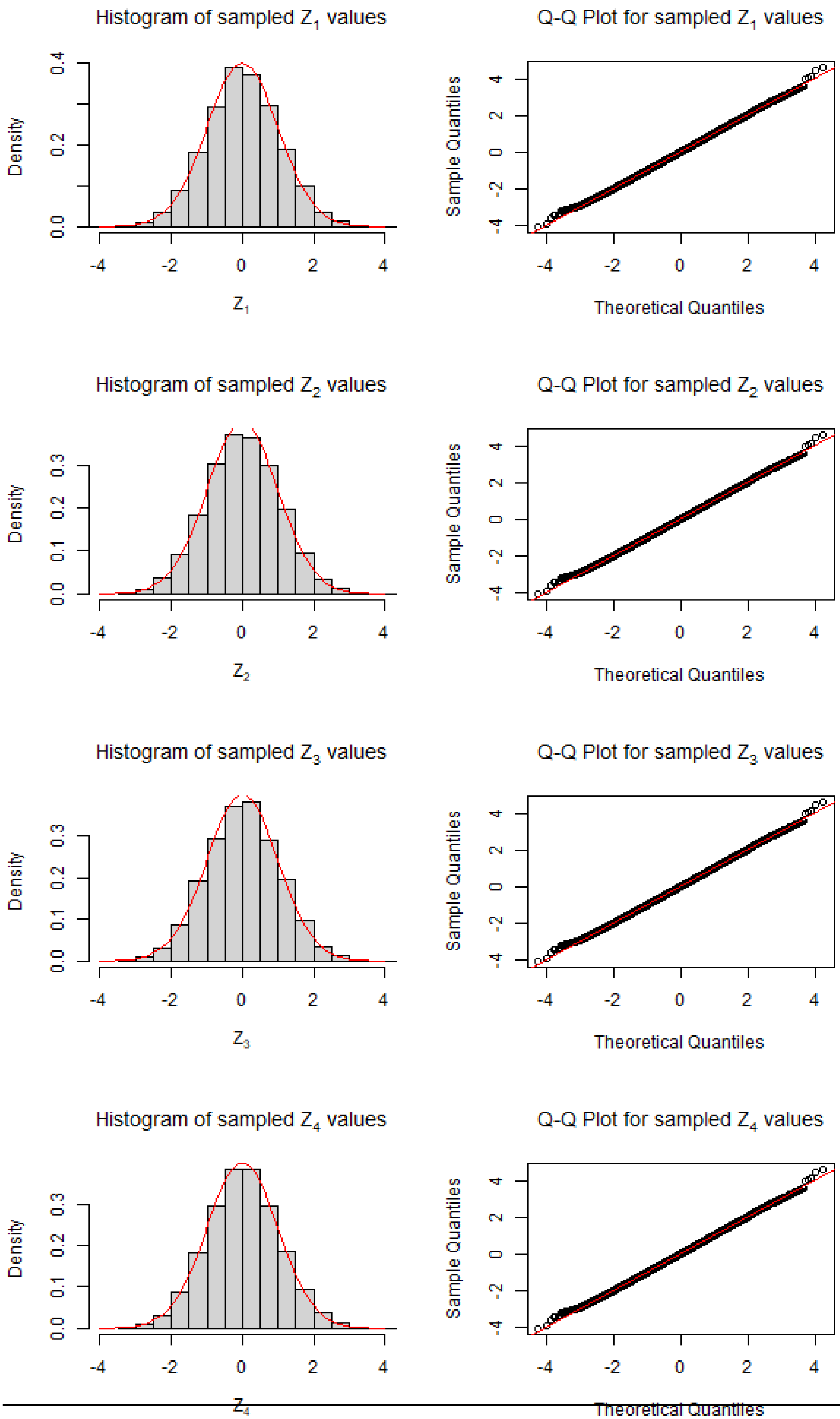
In Section 5.2.3 we proved that asymptotically the sequence of estimates $\Delta(t^*; \hat{\boldsymbol{\theta}}_1), \dots, \Delta(t^*; \hat{\boldsymbol{\theta}}_K)$, has the canonical joint distribution of Definition 2.2. This is the sequence of estimates that occurs when the RMST method is applied for a group sequential trial with K analyses. We now check that this claim is true for a moderately sized trial using simulation. To start, we investigate the marginal distributions of estimates $\Delta(3; \hat{\boldsymbol{\theta}}_k)$ for $k = 1, \dots, K$. The standardised statistic at analysis k, centered on zero, is given by

$$Z_k = \left(\Delta(3; \hat{\boldsymbol{\theta}}_k) - \Delta(3; \boldsymbol{\theta})\right) \sqrt{\mathcal{I}_k} \qquad \text{for } k = 1, \dots, K$$

and we can compare these estimates to the theoretical distribution

$$Z_k \sim N(0, 1).$$

Figures 5.3.6 and 5.3.7 show the result of this simulation study. Figure 5.3.6 was simulated under H_0 with $\eta = 0$ and $b_2 = 0$ and Figure 5.3.7 was simulated under H_A with $\eta = -0.4$ and $b_2 = -0.4$. The Q-Q plots and histograms confirm that the distribution $N(0, 1)$ gives a good fit for Z_k for each $k = 1, \dots, K$.


Histogram of sampled Z_1 values
Density
0.0
0.2
0.4
-4
-2
0
2
4
Z_1
Q-Q Plot for sampled Z_1 values
Sample Quantiles
Theoretical Quantiles
Histogram of sampled Z_2 values
Density
0.0
0.1
0.2
0.3
Z_2
Q-Q Plot for sampled Z_2 values
Histogram of sampled Z_3 values
Z_3
Q-Q Plot for sampled Z_3 values
Histogram of sampled Z_4 values
Z_4
Q-Q Plot for sampled Z_4 values

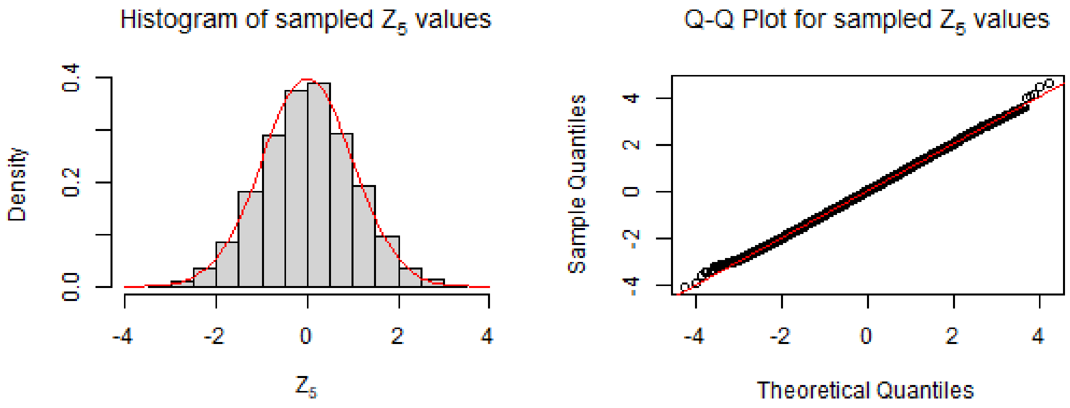
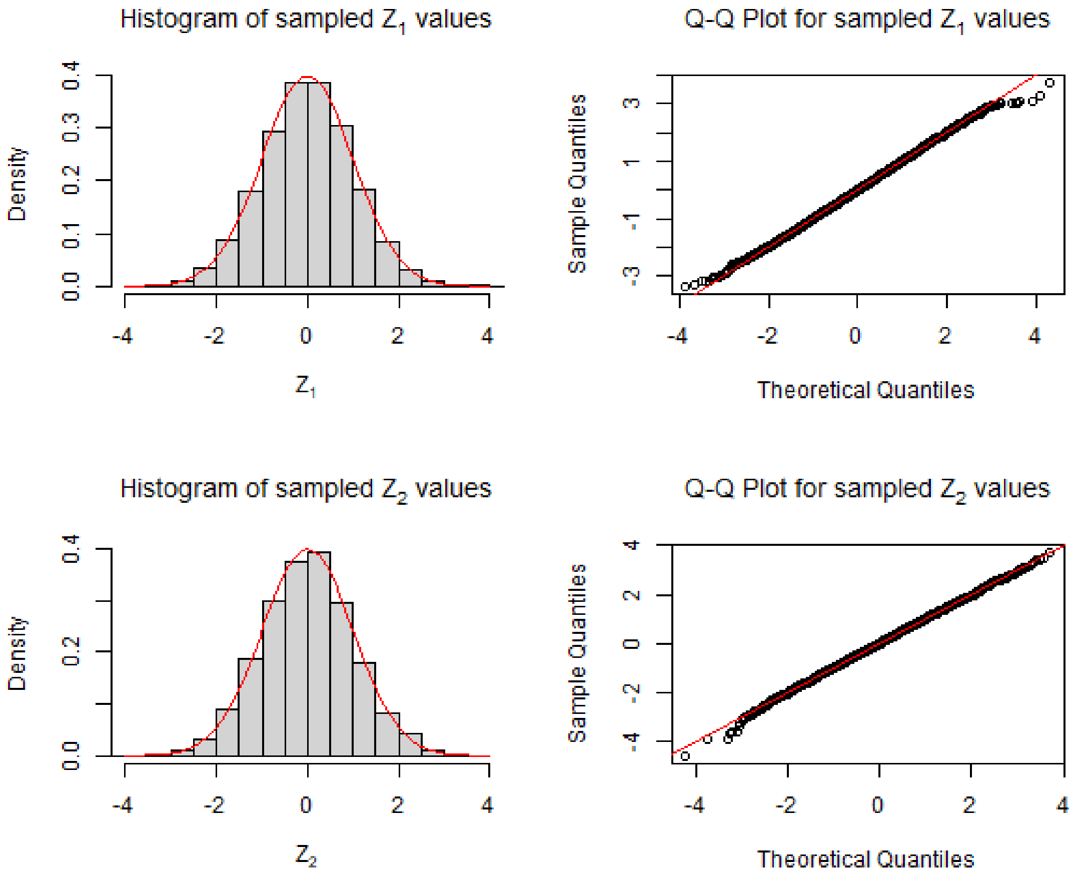


Figure 5.3.6: Histogram and QQ plots for simulated RMST estimates in the joint model under H_0 using $\eta = 0$ and $b_2 = 0$ for each analysis in a group sequential trial using 10^4 replicates.



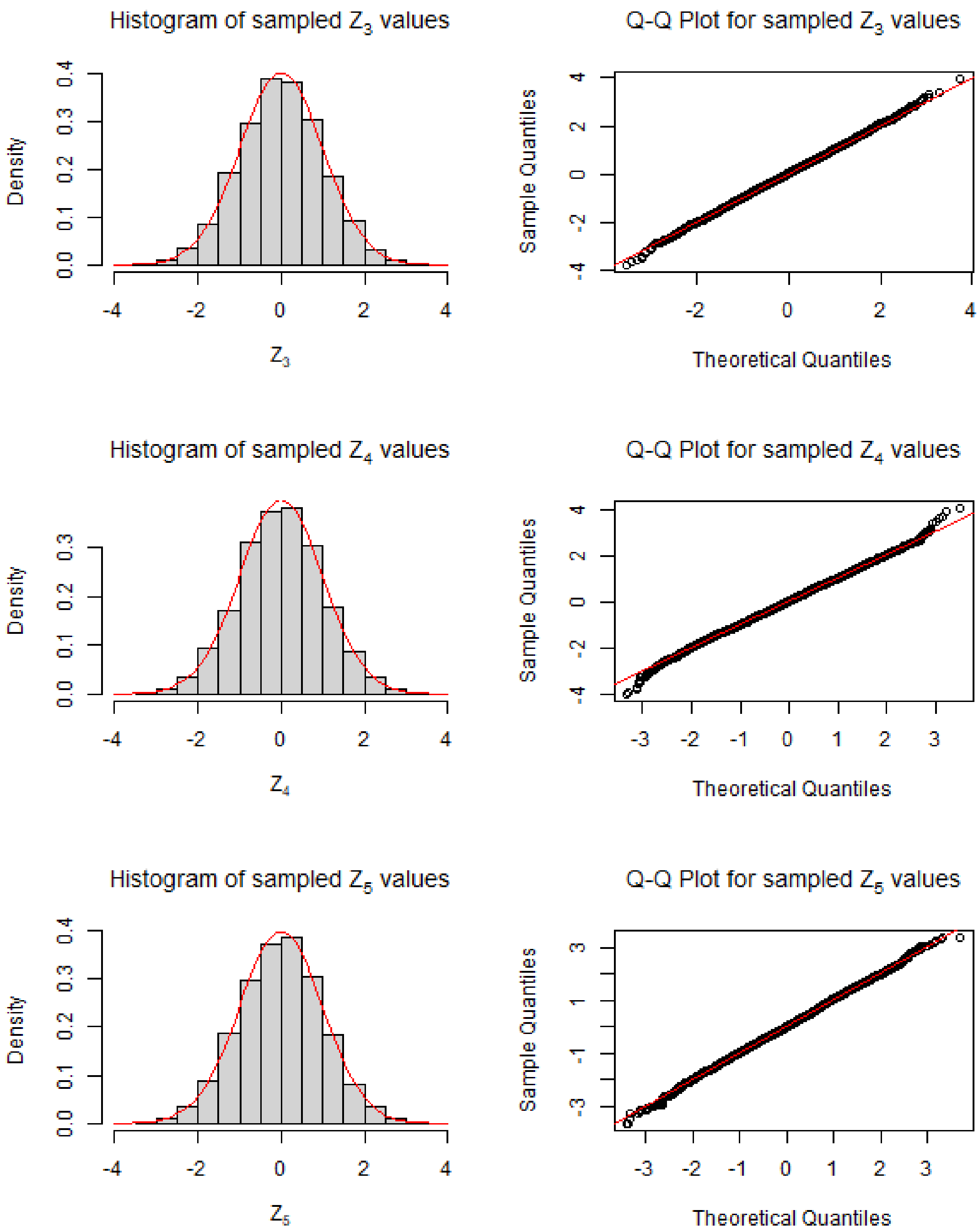


Figure 5.3.7: Histogram and QQ plots for simulated RMST estimates in the joint model under H_A using $\eta = -0.4$ and $b_2 = -0.4$ for each analysis in a group sequential trial using 10^4 replicates.

Further, to assess whether the canonical joint distribution of Definition 2.2 holds, we also assess the covariance structure of the estimates $\Delta(3; \hat{\boldsymbol{\theta}}_1), \ldots, \Delta(3; \hat{\boldsymbol{\theta}}_K)$. Let $\mathcal{I}_k$ be the information level obtained at analysis k and let Z_k be the standardised statistic from analysis k. By the third property of Definition 2.1, for the canonical joint distribution to be true, we should have that

$$Cov(Z_{k_1}, Z_{k_2}) = \sqrt{\frac{\mathcal{I}_{k_1}}{\mathcal{I}_{k_2}}} \qquad \text{for } k_1 \leq k_2.$$

To assess whether this property holds, we calculate $\mathbb{E}(Cov(Z_{k_1}, Z_{k_2}))$ using Monte Carlo methods by estimating $\mathcal{I}_{k_1}$ and $\mathcal{I}_{k_2}$ for each simulation. Then, we can compare this to $\widehat{Cov}(Z_{k_1}, Z_{k_2})$ which is found by taking the covariance of all the Z_{k_1} and Z_{k_2} in the simulations. Tables 5.3.3 and 5.3.6 show the results. We can see that there is little difference in these matrices and the small deviations are consistent with sampling error.

$$\begin{bmatrix} 1.042 & 0.810 & 0.669 & 0.577 & 0.509 \\ 0.810 & 1.040 & 0.852 & 0.730 & 0.652 \\ 0.669 & 0.852 & 1.021 & 0.874 & 0.784 \\ 0.577 & 0.730 & 0.874 & 1.005 & 0.901 \\ 0.509 & 0.652 & 0.784 & 0.901 & 1.006 \end{bmatrix}$$

Table 5.3.1: Under H_0 with $\eta = 0, b_2 = 0$

$$\begin{bmatrix} 1.004 & 0.764 & 0.621 & 0.534 & 0.477 \\ 0.764 & 1.001 & 0.807 & 0.697 & 0.625 \\ 0.621 & 0.807 & 0.984 & 0.852 & 0.766 \\ 0.534 & 0.697 & 0.852 & 0.992 & 0.889 \\ 0.477 & 0.625 & 0.766 & 0.889 & 0.995 \end{bmatrix}$$

Table 5.3.2: Under H_A with $\eta = -0.4, b_2 = -0.4$

Table 5.3.3: Matrix of $\mathbb{E}(Cov(Z_{k1}, Z_{k_2}))$ for group sequential trial with $K = 5$ analyses with 10^4 replicates.

$$\begin{bmatrix} 1.000 & 0.781 & 0.644 & 0.556 & 0.499 \\ 0.781 & 1.000 & 0.824 & 0.712 & 0.638 \\ 0.644 & 0.824 & 1.000 & 0.863 & 0.775 \\ 0.556 & 0.712 & 0.863 & 1.000 & 0.897 \\ 0.499 & 0.638 & 0.775 & 0.897 & 1.000 \end{bmatrix}$$

Table 5.3.4: Under H_0 with $\eta = 0, b_2 = 0$

$$\begin{bmatrix} 1.000 & 0.773 & 0.632 & 0.545 & 0.488 \\ 0.773 & 1.000 & 0.818 & 0.705 & 0.631 \\ 0.632 & 0.818 & 1.000 & 0.861 & 0.772 \\ 0.545 & 0.705 & 0.861 & 1.000 & 0.896 \\ 0.488 & 0.631 & 0.772 & 0.896 & 1.000 \end{bmatrix}$$

Table 5.3.5: Under H_A with $\eta = -0.4, b_2 = -0.4$

Table 5.3.6: Matrix of $\widehat{Cov}(Z_{k_1}, Z_{k_2})$ for group sequential trial with $K = 5$ analyses with 10^4 replicates.

Another check to assess the asymptotic covariance of the sequence of estimates $\Delta(3; \hat{\boldsymbol{\theta}}_1), \ldots, \Delta(3; \hat{\boldsymbol{\theta}}_K)$ is a comparison with the error spending rates for the group sequential trial. Suppose that the canonical joint distribution holds, then the type 1 and type 2 errors will agree with the planned significance and power. This can be expanded further with an error spending test by evaluating the probability of crossing the boundaries at each analysis of the group sequential trial. Under H_0, the amount of type 1 error spent at analysis k is

$$\alpha^{(k)} = \mathbb{P}_{\eta=0,b_2=0}(\text{Continue to analysis } k \text{ and cross the upper boundary at analysis } k).$$

Under H_A, when $\eta = -0.4$ and $b_2 = -0.4$, the amount of type 2 error spent at analysis k is

$$\begin{aligned}\beta^{(k)} = \mathbb{P}_{\eta=-0.4,b_2=-0.4}(&\text{Continue to analysis } k\\ &\text{and cross the upper lower boundary at analysis } k).\end{aligned}$$

Using error spending functions $f(t) = \min\{\alpha t^2, \alpha\}$ and $g(t) = \min\{\beta t^2, \beta\}$, at analysis k, we design the trial with

$$\begin{aligned}
\alpha^{(1)} &= f(\mathcal{I}_1/\mathcal{I}_{max})\\
\beta^{(1)} &= g(\mathcal{I}_1/\mathcal{I}_{max})\\
\alpha^{(k)} &= f(\mathcal{I}_k/\mathcal{I}_{max}) - f(\mathcal{I}_{k-1}/\mathcal{I}_{max}) \qquad \text{for } k = 2, \dots, K\\
\beta^{(k)} &= g(\mathcal{I}_k/\mathcal{I}_{max}) - g(\mathcal{I}_{k-1}/\mathcal{I}_{max}) \qquad \text{for } k = 2, \dots, K.
\end{aligned}$$

Table 5.3.7 shows the probability of crossing each of the boundaries calculated using 10^4 Monte Carlo simulations compared to the expected probability of an error spending test. For 10^4 replicates, we simulate a data set and can calculate $\alpha^{(1)}, \dots, \alpha^{(K)}$ and $\beta^{(1)}, \dots, \beta^{(K)}$ for each replicate. The average values of these over all the simulations are shown in the columns headed "$\mathbb{E}(\alpha^{(k)})$" and "$\mathbb{E}(\beta^{(k)})$".

Analysis	**Simulation type 1 error**	$\mathbb{E}(\boldsymbol{\alpha}^{(\mathbf{k})})$	**Simulation type 2 error**	$\mathbb{E}(\boldsymbol{\beta}^{(\mathbf{k})})$
1	0.0017	0.0013	0.0071	0.0059
2	0.0026	0.0022	0.0089	0.0105
3	0.0045	0.0041	0.0198	0.0203
4	0.0065	0.0061	0.0372	0.0302
5	0.0111	0.0112		
Total	0.0264	0.0250		

Table 5.3.7: Probability of crossing the boundaries in a group sequential trial with $K = 5$ analyses with 10^4 simulated clinical trials, compared to the expected probabilities from an error spending test.

Values for the type 2 error at the final analysis and the total have not been included. This is because at the final analysis, the remaining amount of β is spent in an error spending design and the total β spent will only be equal to 0.9 if $\mathcal{I}_5 = \mathcal{I}_{max}$ exactly. The results show that the probability of crossing the boundaries are very close to the amount of error that is spent at each analysis.

The simulation studies for both the fixed sample and group sequential trials confirm the asymptotic distributional results of the estimate $\Delta(t^*; \hat{\boldsymbol{\theta}})$. Therefore, we may have confidence to perform a clinical trial based on the joint model (5.18). We believe that this simulation study is representative of other values of the parameters $\boldsymbol{\theta}$. Under alternative parameter values we expect to see similar distributional results for the estimate $\Delta(t^*; \hat{\boldsymbol{\theta}})$ so long as the observed number of events between knot points is sufficiently high, for example greater than 30 at the first interim analysis.

5.4 | Comparison of fixed and group sequential trial designs using the restricted mean survival time

With the methodology in place for performing a clinical trial based on this joint model presented in Section 5.2, we compare the results of performing this analysis for a fixed sample versus group sequential design. The main comparison of interest will be the stopping time of the trial, considering the average gain from early stopping in the group sequential trial. We shall also compare the resulting number of hospital visits per person and length of follow-up per person. Each of these outcomes is important for a pharmaceutical company considering the trial design. The first analysis occurs close to the end of recruitment and the probability of stopping before the end of the trial is small. Therefore, the number of patients recruited on completion of the study is not likely to be reduced in the group sequential trial compared to the fixed sample trial and hence, the number of patients enrolled is not considered.

For clarity, the joint model is presented again here. The longitudinal data is assumed to follow a random effects model. For patients $i = 1, \dots, n$ the longitudinal observations are given by

$$W_i(t) = b_{i0} + (b_{1i} + b_2 Z_i)t + \epsilon_i(t)$$

where

- $\begin{bmatrix} b_{i0} \\ b_{i1} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_0 \\ \mu_1 \end{bmatrix}, \begin{bmatrix} \phi_0^2 & 0 \\ 0 & \phi_1^2 \end{bmatrix} \right)$
- $\epsilon_i(t) | \mathbf{b}_i \sim N(0, \sigma^2)$.
- $Z_i = \mathbb{I}\{\text{patient } i \text{ receives the new treatment}\}$.

The hazard function for the survival part of the model is given by

$$h_i(t) = h_0(t) \exp\{\gamma(b_{i0} + (b_{1i} + b_2 Z_i)t) + \eta Z_i\}$$

where the baseline hazard function is

$$h_0(t) = \begin{cases} c_1 & \text{if } t \leq 1 \\ c_2 & \text{if } t > 1 \end{cases} .$$

We set the vector of parameters to be $\boldsymbol{\theta} = (\gamma, \mu_0, \mu_1, b_2, \phi_0, \phi_1, \eta, \sigma^2, c_1, c_2, \lambda)$.

In comparing trial designs, we did the following. For the fixed sample design, the accrual period is fixed at 2 years and follow-up fixed at 3 years. A sample size of n patients is recruited in the first 2 years. For simulation purposes, we ensure that all n patients are recruited within 2 years by simulating arrival times uniformly between 0 and 2 years. In practice, we expect patients to arrive with exponentially distributed waiting times between patients and we can extend recruitment until all n patients have entered the study. The expected power of this simulation study of the fixed sample design is recorded. For the group sequential design, accrual remains the same, with exactly n patients being recruited within the 2 years. We then perform simulation studies of designs with two different calendar times for the final analyses. We shall describe this process in detail shortly. The expected power is recorded for both of the group sequential designs. We can use these two GST observations to find a design with the same power as that of the fixed sample trial. Then the outcomes are compared for fixed and group sequential trial designs with the same power.

In Section 5.3.2 we described how to perform a hypothesis test using the RMST methods and we also give reason for choosing the following values to use in simulation. The common values for simulations under H_0 and H_A are

$$\begin{aligned}(\mu_0, \mu_1) = (3, 1), \phi_0 = 1.2, \phi_1 = 0.25, \sigma^2 = 1,\\ c_1 = 5.32, c_2 = 4.43, t_1 = 1, \lambda = 0.022\end{aligned} \tag{5.24}$$

and we shall investigate properties of the trial for $\gamma = 0, 0.035, 0.07$ and $t^* = 3, 5$. When simulating under H_0 we use $\eta = 0$ and $b_2 = 0$ and when simulating under H_A we use combinations of $\eta = -0.5, -0.4, -0.3$ and $b_2 = -0.45, -0.4, -0.35$. For the longitudinal data, we simulate a trajectory of longitudinal observations by assuming that patient i has a biomarker observation taken at times $t_{i1}, t_{i2}, \ldots, t_{im_i}$. These time points are entry to the study, then until patient i is observed to fail or censored, the measurements are every 2.5 weeks for the first 3 months, then every three months until the study concludes. Therefore, every patient has a biomarker observation at time point $t_{i1} = 0$.

A method for choosing a suitable sample size n is described in Section 5.1.5. We simulate a large data set with 10^4 patients under H_A, and truncate the data set at time 5 years to reflect end of study censoring. The variance-covariance matrix, Σ, for the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$, is estimated using this large data set. Using this estimate for Σ and the choice of t^*, the sample size required to achieve type 1 error $\alpha = 0.025$ and power $1 - \beta = 0.9$ is given by n in Equation (5.4). We have provided the sample sizes and expected power in Table 5.4.1 for an RMST analysis with truncation time $t^* = 3$ years. Power is calculated using 10^4 replicates in a simulation study.

γ	η	b_2	**Sample size**	**Power**
0	-0.3	-0.35	960	0.916
0	-0.3	-0.4	965	0.925
0	-0.3	-0.45	996	0.917
0	-0.4	-0.35	545	0.921
0	-0.4	-0.4	555	0.922
0	-0.4	-0.45	573	0.931
0	-0.5	-0.35	353	0.925
0	-0.5	-0.4	363	0.934
0	-0.5	-0.45	370	0.926
0.035	-0.3	-0.35	798	0.895
0.035	-0.3	-0.4	793	0.890
0.035	-0.3	-0.45	790	0.899
0.035	-0.4	-0.35	461	0.893
0.035	-0.4	-0.4	460	0.894
0.035	-0.4	-0.45	454	0.887
0.035	-0.5	-0.35	299	0.893
0.035	-0.5	-0.4	301	0.896
0.035	-0.5	-0.45	296	0.891
0.07	-0.3	-0.35	736	0.895
0.07	-0.3	-0.4	730	0.901
0.07	-0.3	-0.45	710	0.896
0.07	-0.4	-0.35	434	0.894
0.07	-0.4	-0.4	426	0.892
0.07	-0.4	-0.45	420	0.896
0.07	-0.5	-0.35	287	0.897
0.07	-0.5	-0.4	282	0.898
0.07	-0.5	-0.45	280	0.895

Table 5.4.1: Sample sizes and expected power for model with both longitudinal and survival treatment effects, parameter choices (5.24) and $t^* = 3$ for the RMST analysis based on 10^4 replicates.

We have considered all combinations of parameter values $\gamma = 0, 0.035, 0.07$, $\eta = -0.3, -0.4, -0.5$ and $b_2 = -0.035, -0.4, -0.45$. This table is intended for reference and not for comparison. This is because it is difficult to compare the outcomes of interest (stopping time, number of hospital visits and follow-up time) across different parameter values because the power is not constant. However, we can see roughly how the sample size is affected by these parameters. Larger magnitudes of the two treatment effects η and b_2 require smaller sample sizes, as expected. Further we can see that the sample size calculation from Section 5.1.5 is inaccurate when $\gamma = 0$. The sample size calculation is intended to produce power 0.9 when γ, η and b_2 are the values in the Table, however the simulated power is higher in each case. This is likely to be due to approximation error in the Delta method which makes a linear approximation to the RMST function. For the case $\gamma = 0$, this linear approximation is not accurate.

In Section 5.3.3 we described a method for determining the times of interim analyses. We shall now recap this method. Let $\mathcal{I}_f$ be the information level required for a fixed sample test to attain type 1 error α and power $1 - \beta$ when $\Delta(t^*; \boldsymbol{\theta}) = \delta$. For a group sequential test with K analyses and error spending functions $f(t) = \min\{\alpha t^2, \alpha\}$ and $g(t) = \min\{\beta t^2, \beta\}$, calculate $\mathcal{I}_{max}$ such that information levels $\mathcal{I}_1, \ldots, \mathcal{I}_K$ are equally spaced and the boundaries at the final analysis are such that $a_K = b_K$. This process is described in Section 2.1.3. Then the relationship between information and calendar time is determined. To do so, a large data set is simulated under H_A and at a selection of time points, this data set is truncated and the information levels at these time points are calculated. A *log*-relationship is fit to this data set of analysis times and information levels as in Figure 5.3.5. The times of interim analyses $\tau_1, \ldots, \tau_K$ are then chosen so that $\mathcal{I}_1, \ldots, \mathcal{I}_K$ are equally spaced and that $\mathcal{I}_K = \mathcal{I}_{max}$. This method is governed by a single design parameter $\mathcal{I}_{max}$. Hence, by altering the value of the design parameter $\mathcal{I}_{max}$, and subsequently the analysis times, we can match the power of the group sequential trial to the fixed sample trial. Further, by altering only one parameter, all other design aspects remain fixed so that the change in power arises solely from the times of the interim analyses.

To compare outcomes, we observe simulation studies at different values of $\mathcal{I}_{max}$ and interpolate to match the fixed sample trial. Table 5.4.2 gives the notation for a fixed sample trial design, two group sequential trial designs and the outcomes for each. The final row of Table 5.4.2 will be described shortly.

Design	Planned final information	Expected power	Expected stopping time	Expected number of hospital visits per patient	Expected follow-up per patient
Fixed	$\mathcal{I}_f$	P_f	s_f	v_f	F_f
GST 1	$\mathcal{I}_{max}(1)$	$P(1)$	$s(1)$	$v(1)$	$F(1)$
GST 2	$\mathcal{I}_{max}(2)$	$P(2)$	$s(2)$	$v(2)$	$F(2)$
Final GST	$\mathcal{I}_{max}$	P_g	s_g	v_g	F_g

Table 5.4.2: Notation for outcome variables for different trial designs.

We seek the group sequential trial design with final information $\mathcal{I}_{max}$ with expected power P_f. This is found by interpolating between the points $(\mathcal{I}(1), P(1))$ and $(\mathcal{I}(2), P(2))$.The interpolated maximum information level is given by

$$\mathcal{I}_{max} = \frac{\mathcal{I}_{max}(1)[P_f - P(2)] - \mathcal{I}_{max}(2)[P_f - P(1)]}{P(1) - P(2)}.$$

We further determine the other outcome variables at this value $\mathcal{I}_{max}$. Therefore, the values s_g, v_g and F_g are given by

$$\begin{aligned} s_g &= \frac{s(1)[\mathcal{I}_{max} - \mathcal{I}_{max}(2)] - s(2)[\mathcal{I}_{max} - \mathcal{I}_{max}(1)]}{\mathcal{I}_{max}(1) - \mathcal{I}_{max}(2)} \\ v_g &= \frac{v(1)[\mathcal{I}_{max} - \mathcal{I}_{max}(2)] - v(2)[\mathcal{I}_{max} - \mathcal{I}_{max}(1)]}{\mathcal{I}_{max}(1) - \mathcal{I}_{max}(2)} \\ F_g &= \frac{F(1)[\mathcal{I}_{max} - \mathcal{I}_{max}(2)] - F(2)[\mathcal{I}_{max} - \mathcal{I}_{max}(1)]}{\mathcal{I}_{max}(1) - \mathcal{I}_{max}(2)}. \end{aligned}$$

Figure 5.4.1 shows a visual interpretation for calculating $\mathcal{I}_{max}$ and s_g. For this example, the parameter values for simulation are given by (5.24) and we are simulating under H_A where $\eta = b_2 = -0.4$. The outcome interpolated design has $\mathcal{I}_{max} = 198.5$ and $s_g = 3.54$.

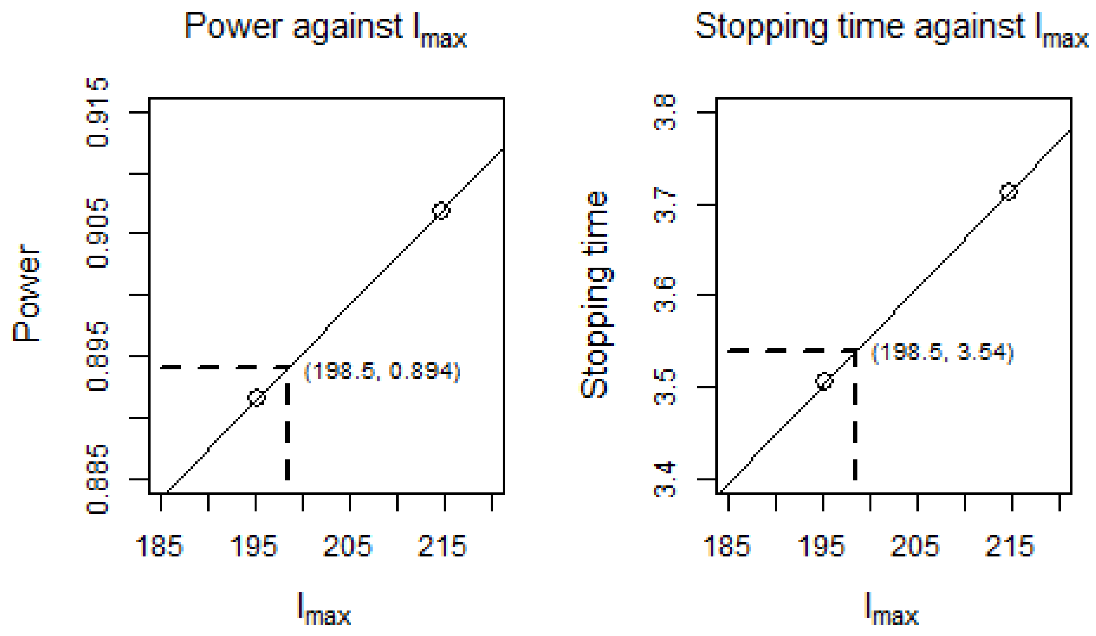


Figure 5.4.1: Interpolation between two GST designs to calculate $\mathcal{I}_{max}$ and s_g.

All designs (fixed sample and the two GST designs) are assessed using the same set of patient outcomes. This is possible as n is fixed and follow-up is varied. This leads to correlated values of the properties of designs, which is beneficial. For example, consider the stopping time outcome. The random variables $s(1)$ and $s(2)$ are correlated and, although we are not sure by how much, the final design will have $Var(s_g) < 2Var(s(1))$. Further, computation is expensive so 10^4 replicates is used in the simulation studies. Interpolating between GST designs means that we need not run the analysis at the interpolated $\mathcal{I}_{max}$ value, and we can trust that the outcomes P_g, s_g, v_g and F_g will be very close to the values which would have occurred if this simulation was performed, due to correlation between designs.

A comparison of the fixed versus group sequential trial designs for an RMST analysis with truncation time $t^* = 3$ years is given in Table 5.4.3. All simulations are performed under H_A. The outcomes (stopping time, number of hospital visits and follow-up time) are the averages using 10^4 simulations. We have considered the stopping time when the trial stops for efficacy and futility separately as the decision upon stopping will affect whether the drug is taken to market or not.

γ	η	$\mathbf{b_2}$	**Expected Stopping time**			**Expected visits per patient**		**Expected follow-up per patient**	
			Fixed	**GST**		**Fixed**	**GST**	**Fixed**	**GST**
			s_f	s_g		v_f	v_g	F_f	V_g
				Reject $\mathbf{H_0}$	**Accept $\mathbf{H_0}$**				
0	-0.3	-0.35	5	3.54	4.04	14.9	11.3	2.62	1.74
0	-0.3	-0.4	5	3.57	4.02	14.9	11.8	2.62	1.86
0	-0.3	-0.45	5	3.58	3.98	14.9	11.6	2.62	1.81
0	-0.4	-0.35	5	3.59	4.08	14.9	11.8	2.62	1.87
0	-0.4	-0.4	5	3.59	4.08	14.9	11.9	2.62	1.88
0	-0.4	-0.45	5	3.61	4.13	14.9	11.9	2.62	1.89
0	-0.5	-0.35	5	3.64	4.22	14.9	12.0	2.61	1.90
0	-0.5	-0.4	5	3.61	4.11	14.9	11.9	2.61	1.89
0	-0.5	-0.45	5	3.51	3.90	14.9	11.7	2.61	1.83
0.035	-0.3	-0.35	5	3.47	4.34	15.0	11.8	2.65	1.87
0.035	-0.3	-0.4	5	3.48	4.24	15.0	11.9	2.65	1.87
0.035	-0.3	-0.45	5	3.49	4.23	15.0	11.9	2.65	1.87
0.035	-0.4	-0.35	5	3.48	4.32	15.0	11.9	2.64	1.87
0.035	-0.4	-0.4	5	3.48	4.28	15.0	11.9	2.64	1.87
0.035	-0.4	-0.45	5	3.47	4.27	15.0	11.9	2.64	1.87
0.035	-0.5	-0.35	5	3.41	4.25	15.0	11.7	2.64	1.84
0.035	-0.5	-0.4	5	3.48	4.27	15.0	11.8	2.64	1.86
0.035	-0.5	-0.45	5	3.44	4.20	15.0	11.8	2.64	1.85
0.07	-0.3	-0.35	5	3.44	4.27	15.1	11.9	2.67	1.88
0.07	-0.3	-0.4	5	3.52	4.39	15.1	12.0	2.67	1.91
0.07	-0.3	-0.45	5	3.52	4.30	15.1	12.0	2.67	1.91
0.07	-0.4	-0.35	5	3.51	4.29	15.1	12.0	2.67	1.90
0.07	-0.4	-0.4	5	3.53	4.29	15.1	12.0	2.67	1.91
0.07	-0.4	-0.45	5	3.50	4.27	15.1	11.9	2.67	1.89
0.07	-0.5	-0.35	5	3.48	4.29	15.1	11.9	2.66	1.88
0.07	-0.5	-0.4	5	3.48	4.26	15.1	11.9	2.66	1.88
0.07	-0.5	-0.45	5	3.47	4.22	15.1	11.9	2.66	1.87

Table 5.4.3: Fixed vs Group Sequential design comparison for model with both longitudinal and survival treatment effects, parameter $t^* = 3$ for the RMST analysis and 10^4 replicates.

Clearly, the group sequential design is much more efficient than the fixed sample design since on average the trial stops roughly 1.5 years earlier than the fixed sample trial when it stops for efficacy and roughly 0.8 years early when it stops for futility. The probability of stopping for efficacy is 0.9 so this is the key case. This benefit means the drug can be taken to market sooner and patients receive an effective treatment early. Further, the number of hospital visits per patient and patient-level follow-up times are dramatically reduced when we use the group sequential design rather than the fixed sample design.

We have previously discussed the importance of the choice of the truncation time t^* for the RMST analysis in Section 5.1.6. The truncation time should be chosen based on clinical meaning and so far we have shown the results for the analysis with $t^* = 3$. We now give a subset of results for the case $t^* = 5$ to assess the differences in the trial when the value of t^* is changed. Note that the value $t^* = 5$ is greater than the maximum follow-up time and hence, this analysis is not possible when the non-parametric RMST estimate is used. Table 5.4.4 gives the sample size calculations using the methods discussed in Section 5.1.5 and the resulting power estimates for a fixed sample clinical trial using simulation with 10^4 Monte Carlo estimates. The sample sizes and power values of Table 5.4.1 are included here for reference.

γ	η	$\mathbf{b_2}$	**Sample size**		**Power**	
			$\mathbf{t^* = 3}$	$\mathbf{t^* = 5}$	$\mathbf{t^* = 3}$	$\mathbf{t^* = 5}$
0	-0.3	-0.35	960	958	0.916	0.922
0	-0.3	-0.45	996	990	0.917	0.918
0	-0.5	-0.35	353	352	0.925	0.923
0	-0.5	-0.45	370	367	0.926	0.936
0.07	-0.3	-0.35	736	671	0.895	0.897
0.07	-0.3	-0.45	710	629	0.896	0.902
0.07	-0.5	-0.35	287	273	0.897	0.903
0.07	-0.5	-0.45	280	262	0.895	0.904

Table 5.4.4: Sample sizes and expected power for model with both longitudinal and survival treatment effects, parameter $t^* = 5$ for the RMST analysis and 10^4 replicates.

A comparison between the sample sizes is difficult to make because the power is not the same in each case. However, it seems as though the difference in t^* has a small effect on the sample size. All sample sizes are within 10% of each other. The reduction in sample size should be set against the disadvantage that occurs from relying on model assumptions. These assumptions cannot be checked with the data available at an interim analysis when t^* is large.

Table 5.4.5 shows a comparison of the fixed versus group sequential trial designs for an RMST analysis with truncation time $t^* = 5$ years. Similarly to the results for $t^* = 3$, the group sequential design out-performs the fixed sample design since stopping times, number of hospital visits and patient-level follow-up times are all dramatically reduced. These reductions are roughly equal to the reductions in Table 5.4.3 when $t^* = 3$, the benefit for using a larger t^* is the reduction in sample size for both fixed and group sequential designs.

γ	η	$\mathbf{b_2}$	**Expected Stopping time**			**Expected visits per patient**		**Expected follow-up per patient**	
			Fixed s_f	**GST** s_g		**Fixed** v_f	**GST** v_g	**Fixed** F_f	**GST** V_g
				Reject $\mathbf{H_0}$	**Accept $\mathbf{H_0}$**				
0	-0.3	-0.35	5	3.52	4.09	14.9	11.7	2.62	1.85
0	-0.3	-0.45	5	3.57	3.97	14.9	11.8	2.62	1.87
0	-0.5	-0.35	5	3.46	4.03	14.9	11.6	2.61	1.81
0	-0.5	-0.45	5	3.56	3.96	14.9	11.7	2.62	1.85
0.07	-0.3	-0.35	5	3.54	4.35	15.1	12.1	2.67	1.92
0.07	-0.3	-0.45	5	3.54	4.29	15.1	12	2.67	1.92
0.07	-0.5	-0.35	5	3.49	4.21	15.1	11.9	2.67	1.88
0.07	-0.5	-0.45	5	3.48	4.23	15.1	11.9	2.66	1.88

Table 5.4.5: Fixed vs Group Sequential design comparison for model with both longitudinal and survival treatment effects, parameter $t^* = 5$ for the RMST analysis and 10^4 replicates.

To complete this comparison, we consider the outcomes when we simulate data under H_0. The parameters η and b_2 describe the two treatment effects and hence, H_0 is described by the scenario where $\eta = 0$ and $b_2 = 0$, or equivalently where the difference in RMST is $\Delta(t^*; \gamma, \eta, b_2) = 0$. Table 5.4.6 shows the type 1 error simulation results each with a Monte Carlo sample size of 10^4. The sample sizes used are those that correspond to $\eta = -0.5, b_2 = -0.45$ in the power calculations. For reference, these are included in the table. The results show that type 1 error is close to $\alpha = 0.025$ in each case.

t*	γ	Sample size	Type 1 error	
			Fixed	GST
3	0	370	0.025	0.025
3	0.035	296	0.026	0.024
3	0.07	280	0.024	0.024
5	0	367	0.026	0.025
5	0.07	262	0.025	0.028

Table 5.4.6: Fixed vs Group Sequential type 1 error rates for model with both longitudinal and survival treatment effects with 10^4 replicates.

We assess the difference in outcomes when data is simulated under H_0. Table 5.4.7 shows the stopping times, number of hospital visits per patient and patient-level follow-up times for both fixed sample trials and group sequential designs. The probability of stopping for efficacy is 0.025 so this a rare event, but we see that the GST design stops early for efficacy roughly 0.5 years earlier than the fixed design. The GST design stops roughly 1.8 years earlier than the fixed sample design when we stop for futility, which means that the trial can be stopped and the resources can be used for something else. Further, number of hospital visits and length of follow-up per patient are also dramatically reduced for the GST in comparison to the fixed sample design.

t*	γ	Expected Stopping time			Expected visits per patient		Expected follow-up per patient	
		Fixed s_f	GST s_g		Fixed v_f	GST v_g	Fixed F_f	GST V_g
			Reject $\mathbf{H_0}$	Accept $\mathbf{H_0}$				
3	0	5	4.68	3.05	14.1	10.8	2.44	1.62
3	0.035	5	4.48	3.26	14.2	10.8	2.46	1.61
3	0.07	5	4.41	3.27	14.3	11.0	2.48	1.66
5	0	5	4.36	2.96	13.9	10.0	2.39	1.43
5	0.07	5	4.29	3.11	14.1	10.3	2.43	1.49

Table 5.4.7: Fixed vs Group Sequential stopping time comparison under H_0 for model with both longitudinal and survival treatment effects with 10^4 replicates.

We have shown comparative results for fixed and group sequential trials for a range of parameter values t^*, γ, η and b_2. We have considered how the trials compare when we simulate data under the alternative hypothesis, stopping for efficacy and also how the trials compare when we simulate data under the null hypothesis and stop for futility. In all cases, the group sequential trial stops early on average by values in the range 1.3-2 years compared to the analysis time at 5 years in the fixed sample trial. Further, the number of hospital visits and the patient-level follow-up times are reduced significantly when a group sequential design is chosen over a fixed sample design.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

6.1 | Conclusions

Two joint models have been proposed for joint modelling of longitudinal and time-to-event data. These models differ by the causal pathway of the treatment. We have shown that it is possible to create a group sequential trial based on each of these joint models and through simulation, we have shown the benefits of these approaches.

Joint model 1 has a single treatment effect that acts directly on the survival endpoint and this model is motivated by literature. The conditional score method is used to find a treatment effect estimate for this model and we have displayed new theoretical results for the distribution of the sequence of treatment effect estimates $\hat{\eta}_1, \dots, \hat{\eta}_K$ found using the conditional score method in a group sequential trial. Although the canonical joint distribution for the sequence $\hat{\eta}_1, \dots, \hat{\eta}_K$ does not hold, we show that it is sensible and practical to proceed assuming that the canonical joint distribution holds anyway. In particular, we have proven that by assuming the canonical joint distribution holds, and using a non-binding futility boundary, the trial is conservative with respect to type 1 error rates. We believe this non-binding case is popular in practice and also presents good evidence that the trial with a binding futility boundary preserves type 1 error conservatively. Finally, using simulation studies we have seen that the deviations from planned type 1 error α are minimal.

Section 4.5 displays the results for this joint model. We show that by including the longitudinal data, compared to the case where the longitudinal data is observed but left out of the analysis, we can greatly improve the efficiency of the trial with respect to sample size. In some cases, 1.67 times as many patients are required to achieve the same power in the analysis where the longitudinal data is left out. These results are seen in both the fixed sample trial and the group sequential trial.

Other benefits for using this joint model are:

- No distributional assumptions are required for the random effects of the longitudinal data.
- The conditional score method is computationally efficient.

The US Food and Drug Administration (2019) encourages the identification of covariates expected to have an important influence on the primary outcome and also discuss accounting for these covariates in the analysis. This motivates the use of Joint model 2. This model includes a treatment effect acting upon the longitudinal data and a second treatment effect acting directly on the survival endpoint. Therefore, we have adjusted for any confounding that may occur. The need for two treatment effects motivates use of the Restricted Mean Survival Time (RMST) as a test statistic and we have proven that the canonical joint distribution holds for the sequence of RMST differences $\Delta(t^*; \hat{\boldsymbol{\theta}}_1), \ldots, \Delta(t^*; \hat{\boldsymbol{\theta}}_K)$ obtained at the interim analyses of a group sequential trial. We have shown that the parametric RMST estimate is favourable since there could be a penalty for using the Kaplan-Meier RMST estimate, that is, a large sample size and/or long study duration are required for many choices of t^*.

In Section 5.4 we have shown comparative results for fixed versus group sequential trials using this joint model. The benefits of GSTs for this simulation study were overwhelming, with the GST stopping roughly 1.5 years early in all cases. Further, the number of hospital visits per patient and average follow-up time per patient are dramatically reduced for the GST compared with the fixed sample trial.

A decision between the methods of Chapters 4 and 5 should be primarily based on the beliefs about the model. Suppose that we truly expect the biomarker to be influenced by treatment, then the RMST analysis of Chapter 5 should be used to analyse the data. It is not possible to use the conditional score analysis of Chapter 4 in this case because we cannot include the treatment effect on the biomarker in the hypothesis test. However, there are advantages to the conditional score method. Most notably, the conditional score estimator is semi-parametric and there is no requirement to specify the baseline hazard function. Therefore by using the conditional score method, we can avoid the complications of parameter identifiability that arise as a result of specifying knot points in the baseline hazard function and the design of the trial is made simpler. Further, when using the conditional score method, we do not need to specify the distribution of the random effects. Therefore, if we believe that the effect of treatment on the biomarker is small or zero, then the conditional score method is the preferred analysis choice.

6.2 | Further work

The findings of this thesis present many avenues for further research. The first is to consider different types of group sequential boundaries. Suppose that the drug regulatory agencies are not convinced by the proposed joint model and are not confident to use this model for the efficacy analysis; they request that the log-rank statistic be used to define the efficacy upper boundary $b_1, \ldots, b_K$. However, the investigator has particular interest in using the joint model to define the futility lower boundary $a_1, \ldots, a_K$. To calculate this set of boundary points, we must therefore determine the joint distribution of the log-rank statistic and the statistic from the joint model ($\hat{\eta}$ for joint model 1 and $\Delta(t^*;\hat{\boldsymbol{\theta}})$ for joint model 2). Further, it may be of interest to the investigator to consider a non-binding futility function. Interest then lies in the potential efficiency gain for this clinical trial design.

Royston and Parmar (2013) suggest that during calculation of the RMST estimate, extrapolation of parameter estimates should be avoided. Taking this into account, in Section 5.3.1 we discuss designing fixed sample clinical trials with emphasis on avoiding extrapolation. However, this becomes more of a challenge when designing a group sequential clinical trial, and there is a compromise occurring between extrapolation and early stopping. In all of our simulations, we have fit the data to the same model it was simulated from and hence, extrapolation is not an issue here since all parameter estimates are unbiased. It is of particular interest to investigate the affect of extrapolating parameter estimates for a misspecified model and determine a suitable limit for the time between interim analyses $\tau_1, \ldots, \tau_K$ and truncation time t^*.

Along a similar avenue, we would like to investigate the robustness of the model when certain aspects are misspecified. We first consider the second joint model which has the limitation that a baseline hazard function must be specified in advance so that fully parametric analyses can been performed. We would like to know the implications of using an incorrect baseline hazard function and if this results in an inflated type 1 error. In this case, an incorrect baseline hazard function might be as minor as misspecifying knot points in the piecewise constant baseline hazard function or it could imply that the entire functional form is wrong. Also, we would like to know how misspecifying the distribution of the random effects of the longitudinal data affect the analysis. For example, suppose that we have fitted the data to a normal distribution but the true underlying distribution for the random effects is a student-t distribution. How would this affect the overall type 1 error for the trial. Thus far, we have assumed that the random effects are normally distributed. This results in computational efficiency as Gauss-Hermite integration can be employed during calculation of the likelihood function. Some consideration needs to be given to the computation for a model where random effects are not normally distributed.

The limitations regarding the baseline hazard function and distribution of the random effects are bypassed when using the conditional score method. However, we do have the disadvantage that we cannot be sure that the canonical joint distribution holds for the sequence of treatment effect estimates. In Section 4.4.3 we gave some evidence showing that the trial is likely to be conservative with respect to type 1 error and we suggest that during planning, a single, large clinical trial should be simulated to check that the correlations are in the direction that would give rise to type 1 error less that or equal to α. That is, we suggest estimating Σ using a simulation with 4800 patients then checking that $\rho \geq \rho^*$. The difficulty with this approach is that we require knowledge of the model parameters before commencing the trial. Therefore, it is of interest to assess the effects of planning a trial with parameter values that are incorrect.

Another limitation that is common across the two joint models is that the biomarker is assumed to follow a linear trajectory. In practice, this is uncommon since for biomarkers which are made up of count data such as circulating tumour DNA (Rothwell et al. (2019)), we impose a non-negative constraint. It is therefore important to assess how well these non-linear functions can be captured by a simple linear model and whether the conditional score and RMST analyses are robust to this misspecification. It may then be necessary to develop theory that allows for generalised linear mixed models for the longitudinal data.

Finally, it is of high interest to apply these methods to a real clinical trial data set. So far, we have presented results for joint models with simple linear longitudinal trajectories and a treatment indicator as the only covariate, and we have described how these models can be generalised to include more complex trajectories and more covariates. In practice, calculation of the trial statistic may be computationally expensive and require further consideration. Taylor et al. (2013) present a joint model for clinical recurrence of prostate tumours, a time-to-event outcome, and prostate specific antigen (PSA), a longitudinal data measurement. Using the specified form for the joint model, we could design a group sequential trial to implement the methods of this thesis. Bikdeli et al. (2017) present a literature review for 220 surrogate endpoint trials within cardiovascular disease. The primary endpoint for each of these trials was a time-to-event outcome, at least 42 trials used longitudinal data as a surrogate endpoint and these trials frequently show superiority of the treatment intervention. This indicates a need for including biomarker observations when available in time-to-event clinical trials. Further, under the joint modelling framework, there is no loss for collecting time-to-event observations for inclusion in the joint model compared to only collecting biomarker observations and performing a surrogate endpoint clinical trial.

CHAPTER 7

BIBLIOGRAPHY

M. Akacha, F. Bretz, D. Ohlssen, G. Rosenkranz, and H. Schmidli. Estimands and their role in clinical trials. *Statistics in Biopharmaceutical Research*, 9(3):268–271, 2017. (Cited on page 152.)

P. K. Andersen and R. D. Gill. Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, pages 1100–1120, 1982. (Cited on pages 35, 43, 46, 49, and 72.)

P. K. Andersen, O. Borgan, R. D. Gill, and N. Keiding. *Statistical Models Based on Counting Processes*. New York: Springer Science & Business Media, 2012. (Cited on pages 35, 36, 43, 44, and 45.)

B. Bikdeli, N. Punnanithinont, Y. Akram, I. Lee, N. R. Desai, J. S. Ross, and H. M. Krumholz. Two decades of cardiovascular trials with primary surrogate endpoints: 1990–2011. *Journal of the American Heart Association*, 6(3):e005285, 2017. (Cited on page 207.)

R. J. Carroll, D. Ruppert, C. M. Crainiceanu, and L. A. Stefanski. *Measurement Error in Nonlinear Models: A Modern Perspective*. London: Chapman and Hall/CRC, 2006. (Cited on pages 73 and 84.)

P.-Y. Chen and A. A. Tsiatis. Causal inference on the difference of the restricted mean lifetime between two groups. *Biometrics*, 57(4):1030–1038, 2001. (Cited on page 152.)

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society*, 34(2):187–202, 1972. (Cited on page 32.)

D. R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975. (Cited on pages 32 and 33.)

D. R. Cox and D. V. Hinkley. *Theoretical Statistics*. London: Chapman and Hall/CRC, 1979. (Cited on page 22.)

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–22, 1977. (Cited on page 168.)

J. L. Doob. The limiting distributions of certain statistics. *The Annals of Mathematical Statistics*, 6(3):160–169, 1935. (Cited on pages 154 and 171.)

A. Genz. Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1(2):141–149, 1992. (Cited on page 120.)

A. Genz, F. Bretz, T. Miwa, X. Mi, F. Leisch, F. Scheipl, B. Bornkamp, M. Maechler, T. Hothorn, and M. T. Hothorn. Package 'mvtnorm'. *Journal of Computational and Graphical Statistics*, 11:950–971, 2020. (Cited on pages 108 and 120.)

A. I. Goldman, B. P. Carlin, L. R. Crane, C. Launer, J. A. Korvick, L. Deyton, and D. I. Abrams. Response of CD4 lymphocytes and clinical consequences of treatment using ddI or ddC in patients with advanced HIV infection. *Journal of Acquired Immune Deficiency Syndromes*, 11(2):161–169, 1996. (Cited on page 54.)

J. Haybittle. Repeated assessment of results in clinical trials of cancer treatment. *The British Journal of Radiology*, 44(526):793–797, 1971. (Cited on page 113.)

C. Jennison and B. W. Turnbull. Interim analyses: the repeated confidence interval approach. *Journal of the Royal Statistical Society*, 51(3):305–334, 1989. (Cited on pages 182 and 184.)

C. Jennison and B. W. Turnbull. Group-sequential analysis incorporating covariate information. *Journal of the American Statistical Association*, 92(440):1330–1341, 1997. (Cited on pages 6, 22, 47, 61, 122, and 124.)

C. Jennison and B. W. Turnbull. *Group Sequential Methods with Applications to Clinical Trials*. London: Chapman and Hall/CRC, 2000. (Cited on pages 4, 91, and 136.)

E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958. (Cited on page 13.)

S. W. Lagakos. General right censoring and its impact on the analysis of survival data. *Biometrics*, pages 139–156, 1979. (Cited on page 10.)

Q. Liu and D. A. Pierce. A note on Gauss Hermite quadrature. *Biometrika*, 81(3): 624–629, 1994. (Cited on page 18.)

X. Lu and A. A. Tsiatis. Improving the efficiency of the log-rank test using auxiliary covariates. *Biometrika*, 95(3):679–694, 2008. (Cited on pages 54 and 55.)

Y. Lu and L. Tian. Statistical considerations for sequential analysis of the restricted mean survival time for randomized clinical trials. *Statistics in Biopharmaceutical Research*, pages 1–9, 2020. (Cited on page 152.)

S. Murray and A. A. Tsiatis. Sequential methods for comparing years of life saved in the two-sample censored data problem. *Biometrics*, 55(4):1085–1092, 1999. (Cited on page 152.)

National Research Council. *The Prevention and Treatment of Missing Data in Clinical Trials.* Washington, D.C: National Academies Press, 2010. (Cited on page 152.)

P. C. O'Brien and T. R. Fleming. A multiple testing procedure for clinical trials. *Biometrics*, pages 549–556, 1979. (Cited on page 113.)

S. Pampallona and A. A. Tsiatis. Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *Journal of Statistical Planning and Inference*, 42(1-2):19–35, 1994. (Cited on page 7.)

S. J. Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199, 1977. (Cited on pages 7 and 113.)

D. Rizopoulos. Jm: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, 35(9):1–33, 2010. (Cited on pages 89, 174, and 177.)

D. Rizopoulos. *Joint Models for Longitudinal and Time-to-event Data: With Applications in R.* London: Chapman and Hall/CRC, 2012. (Cited on pages 54 and 168.)

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951. (Cited on page 138.)

D. G. Rothwell, M. Ayub, N. Cook, F. Thistlethwaite, L. Carter, E. Dean, N. Smith, S. Villa, J. Dransfield, A. Clipson, et al. Utility of ctdna to support patient selection for early phase clinical trials: the target study. *Nature medicine*, 25(5): 738–743, 2019. (Cited on page 207.)

P. Royston and M. K. Parmar. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Medical Research Methodology*, 13(1):152, 2013. (Cited on pages 150, 151, 155, and 206.)

T. Shao, T. C. Chen, and R. Frank. Tables of zeros and Gaussian weights of certain associated Laguerre polynomials and the related generalized Hermite polynomials. *Mathematics of Computation*, 18(88):598–616, 1964. (Cited on page 18.)

J. M. Taylor, Y. Park, D. P. Ankerst, C. Proust-Lima, S. Williams, L. Kestin, K. Bae, T. Pickles, and H. Sandler. Real-time individual predictions of prostate cancer recurrence using joint models. *Biometrics*, 69(1):206–213, 2013. (Cited on pages 54 and 207.)

A. A. Tsiatis and M. Davidian. A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika*, 88 (2):447–458, 2001. (Cited on pages 54, 55, 58, 59, 60, 61, 62, 63, 67, 69, and 73.)

A. A. Tsiatis and M. Davidian. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, pages 809–834, 2004. (Cited on pages 54, 56, and 168.)

A. A. Tsiatis, H. Boucher, and K. Kim. Sequential methods for parametric survival models. *Biometrika*, 82(1):165–173, 1995. (Cited on pages 49 and 51.)

US Food and Drug Administration. *Guidance for Industry: Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biologics with Continuous Outcomes [Draft Guidance]*. Silver Spring: US Food and Drug Administration, 2019. (Cited on pages 150 and 204.)

A. W. Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge: University Press, 2000. (Cited on page 69.)

J. Wakefield. *Bayesian and Frequentist Regression Methods*. Berlin: Springer Science & Business Media, 2013. (Cited on pages 22 and 31.)

L. Zhao, B. Claggett, L. Tian, H. Uno, M. A. Pfeffer, S. D. Solomon, L. Trippa, and L. Wei. On the restricted mean survival time curve in survival analysis. *Biometrics*, 72(1):215–221, 2016. (Cited on page 152.)