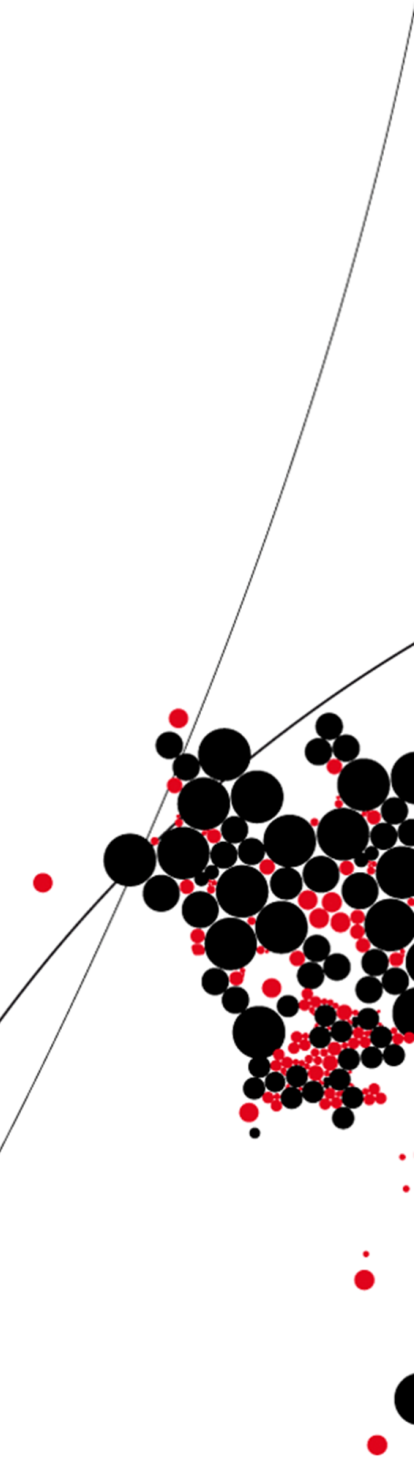
UNIVERSITY OF TWENTE.

Faculty of Electrical Engineering,
Mathematics & Computer Science

Detecting Treatment Effects in Clinical Trials Without a Control Group

Stef Baas

M.Sc. Thesis
November 3, 2019



Supervisors:
Prof. Dr. R.J. Boucherie
Dr. Ir. G.J.A. Fox

Stochastic Operations Research
Faculty of Electrical Engineering,
Mathematics and Computer Science
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

Preface

This report is the result of my final project for the master Applied Mathematics at the University of Twente. The work was performed at the University of Twente from February 2019 until September 2019 under the guidance of Richard Boucherie and Jean-Paul Fox.

I would like to thank Richard Boucherie for his encouragement, guidance and advice throughout this project. The long discussions we had at his office were always insightful. Furthermore, a very special thanks goes out to Jean-Paul Fox. Finding an approach proved very difficult for me at the start of this project and his guidance helped me get on the right path. His support in making me understand concepts in Bayesian statistics (of which I didn't know a lot at the start of this project) and his insights were very fruitful for this project.

After starting with the additional correlation framework previously explored by Jean-Paul, I found previously unseen mathematical results that improved the performance of the inference method greatly. I am proud that my research led to these contributions, which brought us a significant step closer to identifying treatment effects using only the treatment group.

Furthermore, a special thanks goes out to my friends and family for their support during this project.

Stef Baas

Enschede, November 3, 2019

Abstract

The randomized controlled trial has been the golden standard for clinical testing of treatment efficacy for the last 70 years. To determine a treatment effect, patients are randomly assigned to a treatment group or a control group. In the control group, patients sometimes do not receive a treatment, only serving as the statistical controls to determine the treatment effect. This is done such that the average measurement of both groups can be compared, and the statistical significance of the treatment effect can be evaluated. However, it is considered unethical to assign patients to a group who do not receive treatment, while there is already an existing effective therapy. This is especially the case when the placebo group concerns a vulnerable group like children, psychiatric patients, and patients suffering from cancer.

In this research, a statistical method is developed in which the effect of a medical treatment is tested for without a control group. The idea is that groups of patients undergoing effective treatment will show correlated outcomes. The modeling framework considered in this research provides a way to test for this additional correlation in interval-censored survival data. In a simulation study, it is shown that objective Bayesian inference can be efficiently performed on such data, and additional correlation can be tested for.

Keywords: clinical trials, covariance testing, Bayesian statistics, Bayes factors, survival analysis, Markov chain Monte Carlo.

Contents

Chapter 1

Introduction

For the last 70 years, the randomized controlled trial has been the golden standard for statistically assessing the benefits of a new treatment over a standard one (Pocock, 2013). In these trials, patients are randomly assigned to either a control or a treatment group. In cases where e.g. there currently exists no treatment, patients in the control group receive no treatment or only receive a placebo (saline). This is done such that a significant difference in average outcomes can be determined between control group patients and patients in the treatment group(s). The ethical concern with this is however that a group of patients in the trial does not get any treatment, while it is possible to treat them. Especially in cancer research, child care or psychiatric care, clinical trials with a placebo control groups face this criticism.

In (Fox, Mulder, & Sinharay, 2017), Bayesian covariance testing is explored for an equicorrelated multivariate probit model. The explored idea in this research was to apply this to a multivariate survival model. The choice was made to consider the model for (type II) interval censored survival data introduced in X. Lin and Wang (2010). As the underlying latent variables in this model are Gaussian, this survival model can be easily extended to handle more complicated covariance structures.

With the testing procedure considered in this research it is possible to detect treatment effects in clinical data without the need for a control group. Namely, if patients are subjected to an effective treatment, patients will have a (positive or negative) response to this treatment. This change in response will manifest partly in the form of additional covariance in the outcomes of these patients. When there are groups of patients in the trial that have a different response to the treatment, the treatment induced covariance can be tested for, and hence a treatment effect can be determined. Furthermore, in the situation where group differences are detected, personalized medicine might be a viable

option for this treatment.

Testing without the need for a control group has a lot of benefits. If the control group would have gotten a placebo, all patients are now given the treatment. Furthermore, difficulties associated with designing and implementing an RCT are avoided. Finally, the procedure leads to a serious reduction in costs to evaluate the effectiveness of a treatment by only requiring treatment data.

Another manner in which covariance testing might be used is in the case where different versions of a treatment are administered in a trial. Detecting covariance in the outcomes could lead to detection of an optimal version, or could indicate that personalized medicine might be an option.

In the next chapter, a literature study on clinical trials is summarized. After that, an introduction to concepts in Bayesian statistics that are explored in this research will be given. Next, the multivariate survival model considered in this research is introduced. In the chapter that follows, the employed inference method for this model will be explained. As the limits to Bayesian inference are largely determined by computational tractability, a simulation study is performed in Chapter 6 to evaluate whether inference can be performed efficiently and reliably. The final chapter contains a conclusion and discussion.

Chapter 2

Clinical Trials

2.1 Randomized Controlled Trials in medicine

This chapter summarizes a literature study on the design of clinical trials, and methods for sample size reduction. The main sources on clinical trials used here are Friedman et al. (2010) and Pocock (2013).

Following Friedman et al. (2010), a randomized controlled trial (RCT) in medicine can be defined as "*a prospective study comparing the effect and value of intervention(s) against a control in human beings. Subjects are partitioned in groups according to a formal randomization procedure, and subject-linked outcome variables are compared*". RCTs are conducted in medicine, but also increasingly in e.g. business, economy or social sciences (Deaton & Cartwright, 2018). The main difference in medicine is that in many cases the design of the trials has an ethical aspect. In extreme cases, the decision to give a subject an intervention can be the difference between life and death. Another difference between clinical trials and trials in other fields can be the fact that human subjects are considered, hence there is a possibility that subjects do not adhere to the treatment protocol. Finally, double measurements are often not possible, e.g. when patient survival times are measured.

A clinical trial is prospective, which means that subject outcomes can be monitored and analyzed during the trial. Furthermore, subjects do not enroll in the study simultaneously. Due to the prospective nature, intermediate intervention in RCTs is possible. This intermediate intervention can be e.g. to stop the trial prematurely, to adapt the assignment procedure, or to increase the dose of medicine. The prospective aspect leads to flexibility in the design of RCTs, making e.g. online optimization of the trial design possible.

2.1.1 History

According to Pocock (2013), one of the most famous early examples of a modern clinical trial is the study of Lind in 1753, who evaluated treatments for scurvy (Lind, 1757). Procedures to evaluate treatment effect can be traced back to 2000 BC, but Lind's trial was one of the first in which emphasis was placed on keeping all factors other than treatment as comparable as possible.

In the setup by Lind, not much significance was given to the measurement procedure of patient outcomes, deviation from treatment (non-adherence) and the registration of the patient diagnosis at arrival. One of the first proponents of placing emphasis on these factors was Louis (Louis, 1835), stating importance of these factors in 1835 for determining whether bleeding had any effect on the progression of pneumonia. His trial found no significant differences in outcomes for the treatment groups, and led to the eventual decline of bleeding as a treatment.

The first instance of a trial with randomization and single-blinding was reported in 1931 by Amberson Jr (1931). Single-blinding denotes the situation in which patients do not know the groups to which they are assigned. The group allocation in this trial was decided by partitioning the subjects in two groups, and flipping a coin to determine which group gets the new treatment.

Although there were trials in which the treatment effect was obvious, for some trials this was not the case. A formal procedure to determine a significant difference in group outcomes was needed. Such a procedure was introduced by Fisher in his book *design of experiments* (Fisher, 1936). The concept of a Null hypothesis, as well as the Fisher exact test were introduced in this book. The Fisher exact test is used to compare binary outcomes, and is still used to this day in clinical research.

Around the middle of the 20-th century, the randomized clinical trial became the preferred method to evaluate new medical treatments. This development is largely credited to Sir Austin Bradford Hill. Hill introduced the randomized double-blinded controlled trial in the British Medical Research Council's trial of streptomycin for pulmonary tuberculosis (Hill, 1990). In double-blinded RCTs, both the subjects and investigators do not know the group allocation, which is randomized. This double blinding removes a large amount of allocation bias. Since the work of Hill, the design of RCTs has remained relatively unchanged and RCTs remain the golden standard of clinical testing to this day.

2.1.2 Phases of Clinical Research

When RCTs are used to assess the effect of a new treatment, the trial can be classified in one of four *phases* of experimentation (Pocock, 2013):

- **Phase I Trials: Test for Toxicity and Clinical Pharmacology**
  These trials mostly test the safety of a new medicine, not the efficacy, and hence are mostly applied to healthy human test subjects or patients that did not respond to the standard treatment. The objective is often to estimate the maximum tolerated dose in dose-escalation experiments. Other objectives can be e.g. to find the biochemical or psychological effect of the drug on the subject, or to determine the bioavailability of the drug (e.g. how long the drug stays in the body).

- **Phase II Trials: Initial Test of the Clinical effect**
  This phase is reached if the drug has passed phase I. Phase II trials are small scale (100-200 patients) investigations to assess the effects of a number of drugs on patients. These patients are often carefully selected and heavily monitored. Often, phase II trials are used to select the drugs with genuine potential from a larger number of drugs, so that these may continue to phase III.

- **Phase III Trials: Full-scale Evaluation of Treatment**
  When the drug(s) have passed phase II, the new drug(s) are compared to the standard treatment in a larger trial (300 - 3000 patients). A control/reference group is necessary in this phase.

- **Phase IV Trials: Post-marketing surveillance**
  After successful completion of phase III, the drug can be administered to anyone seeking treatment. The physician prescribing the medicine will monitor the long term/large scale effects of the drug.

As phase I and IV trials do not require control groups of patients, the focus in the remainder will lie on phase II-III trials. From this section, it is clear that a clinical trial is often performed in a sequence of trials, and hence does not occur in a vacuum.

2.1.3 Organization of Phase II-III Trials

When a clinical trial is conducted, the research question(s) should be well posed. There should always be a primary question, and possibly some secondary questions. Furthermore, the definition of the study population is an integral part of posing the primary

question. The study population is the part of the total patient population eligible for the trial. In general it is not sufficient to know that a treatment has had an effect, it is also important to know which group of subjects the treatment has effect on. The study population is defined by the inclusion/exclusion criteria of the trial. These criteria are often based on:

1. The potential for the subjects to benefit from the treatment.
2. The possibility to detect the treatment effect in subject outcomes.
3. The possibility that the treatment is harmful for the subjects.
4. Effects of other diseases that could interfere with successful treatment.
5. The probability that the subjects adhere to the treatment protocol.

When defining a study population, it must be kept in mind how this study population relates to the total patient population. In some cases, data (features and outcomes) for the excluded patients are collected. In this case, inference can be made as to what extent trial results can be extrapolated to expected results in the overall population. In some cases, data from excluded patients is not available and some leap of faith, based on expert knowledge, has to be taken to extrapolate trial results to the total population.

From the above discussion, it can be seen that RCTs often only consider a part of the total patient population. How trial results can be extrapolated to expected results for the overall patient population is always something to consider in clinical research.

2.1.4 Randomization Procedure

In clinical trials, the preferred method of assessing a treatment effect is a trial in which patients are randomly allocated to a control or treatment group.

One reason for this is that, in combination with double blinding, it eliminates bias in treatment assignment. An example of this is a physician that always assigns more frail subjects to the experimental/standard treatment because he believes this treatment is superior. With randomization and blinding, this is not possible anymore. Furthermore, randomization is also believed to reduce bias by accounting for unobserved variables having an effect on the outcomes. Under randomization, the same distribution for these *confounding variables* is induced in the control and treatment group.

Lastly, randomization justifies the reasoning that in the case of ineffective treatment, average outcome differences between treatment and control groups are observed by chance. This justifies the use of statistical tests in RCTs.

Different randomization procedures can be used, and in the trial description it should always be clear which one is used:

- **Simple Randomization**
  In simple (fixed) randomization, each patient is assigned to each group k with some fixed probability p_k. In (Friedman et al., 2010), it is advised that allocation should be uniformly distributed in an RCT ($p_k = 1/N$ for N groups). In order to avoid a large difference in the sample sizes of treatment groups, simple randomization can be done according to an accept/reject method with some acceptance criteria (e.g. no more than a difference of 10 patients between all group sizes).

- **Blocked Randomization**
  Another way to avoid serious imbalance in the number of participants assigned to each group is blocked randomization. Subjects are (approximately) divided in K *sampling groups* with size equal to the number treatment groups ($|\mathcal{G}_k| = N$ for each sampling group $\mathcal{G}_k$). Next, members of each sampling group are randomly divided over the treatment groups such that 1 member per sampling group is assigned to each treatment group.

- **Stratified Randomization**
  One of the reasons for randomization is to balance the treatment groups in terms of factors determining the treatment outcomes. In stratified randomization, the subject population is divided in strata (e.g. male/female, age higher/lower than 65). For the arriving subjects, the variables corresponding to the strata are measured. Next, patients in the same strata are randomly divided (by simple/blocked randomization) over treatment groups. A downside to this randomization procedure is that a large number of subjects might be necessary to get a significant amount of subjects per strata. Also, factors thought to be important a priori might turn out to not be important in the outcome analysis, inducing an unnecessary number of strata. According to Friedman et al. (2010), a regression analysis can also be conducted instead of stratification, which results in approximately the same amount of power.

From this section, it is clear that randomization is used in order to reduce allocation bias, and to validate the use of statistical tests. Different methods of randomization are possible.

2.1.5 Outcome Variables and Statistical Tests

In clinical trials, it is often the case that outcomes from one treatment procedure are compared with outcomes from one other treatment procedure in a frequentist hypothesis test. Statistical tests comparing three or more treatment groups, using Bayesian methods or covariates, as well as paired samples are also known in literature (see e.g. Walker and Almond (2010), Armitage, Berry, and Matthews (1971)) but will not be considered in this section, as the most often occurring testing procedures are two-sample frequentist tests.

It is often assumed that the outcomes in the two treatment groups are independent and identically distributed (iid). This assumption is justified by checking that the treatment groups are *balanced*. For this, statistical tests are often performed for assessing differences in the distribution of characteristics between the two treatment groups. The effect of having balanced groups is that all variables having an effect on the comparison are accounted for. When differences between patient outcomes are compared in e.g. a t-test, only the treatment effect will be measured on average.

The main three outcome variables in clinical trials, as well as the most often performed test to assess significant differences are now listed below.

1. **Qualitative responses**
   Qualitative responses are responses that fall in a finite range. Examples of these are e.g. a test results, stages of some disease or the indicator of some symptom. Often used frequentist hypothesis tests on qualitative responses are the *Fisher exact test* (Mehta & Senchaudhuri, 2003) and the Chi-square test (McHugh, 2013). If the qualitative data can be ordered (i.e. is ordinal), the Mann-Whitney U-test can be performed (Mann & Whitney, 1947).

2. **Quantitative responses**
   In the case of quantitative responses, the responses can (approximately) take on any value in $\mathbb{R}$, or a subset of $\mathbb{R}$ with infinite cardinality. Examples of quantitative observations are e.g. concentrations of hormones or tumor size. The most often performed test on this type of data is the independent two sample t-test (Walker & Almond, 2010). Other often performed tests are the Mann-Whitney U-test (again) and Welch's t-test (Pocock, 2013).

3. **Event time responses.**
   Another possible outcome variable in clinical trials can be in the form of event-time responses. This can be the recurrence time of a disease, the time that the patient

comes back to the clinic, or time of death. In chapter 4, where the multivariate survival model is introduced, more information will be given on this type of data. A central object to event-time responses is the survival curve, which for each t returns the probability that the event time is larger than t. The most often performed test for assessing equality of the survival curves based on event-time outcomes is the Logrank test (Korosteleva, 2009).

In Appendix G, more information is given on outcome variables in clinical trials, and often used frequentist hypothesis tests.

2.2 Sample Size Reduction in Clinical Trials

Despite the advantages of RCTs, allocating patients randomly in a treatment and control group is unethical in some cases. The main example of this is the case when no (good) treatment is available prior to starting the RCT. Control group patients often only receive a placebo (saline) in this case. Especially in cancer research, child disease or research on psychological diseases, the effect of this is detrimental. Hence, statisticians have been (and are still) trying to redesign RCTs in such a way that the required sample (or control group) size is reduced. This section lists the three main methods found in literature to do this.

2.2.1 History Controlled Trials

In history controlled trials (HCT), control group outcomes are obtained from historical patient data, reducing the minimal required control group size. The main problem with HCT's is the question of how one can decide to what extend the historical data is representative for the current control group. In Pocock (2013), it is stated that the causes for potential incompatibility can be divided into two areas, *patient selection* and the *experimental environment*.

- The incompatibility from patient selection involves the fact that subjects from the historical control group might not adhere to the inclusion criteria of the trial, and it could be impossible to find out what patients would have been included in the trial due to data limitations. Furthermore, a change in patient population between trials might make the results of the former trial not representative for the current one.

- The incompatibility due to experimental environment stems from the fact that e.g. the quality of the historical data might be inferior to the currently collected data and the recording procedure of the trial outcomes may change over time. Furthermore, the overall healthcare procedures for patients may change over time. This could for example be due to doctors leaving the hospital or overall healthcare improvement. Another problem is that non-adherence is often not recorded in historical data. The overall effect is that a historical control group might have entirely different properties as compared to a control group in a clinical trial.

Nevertheless, in Pocock (2013) and Friedman et al. (2010), it is stated that despite the limitations of historical controls, there are cases in which they can be used. In Pocock (2013), it is stated that historical controls from a *previous trial* in the same organization might be of use in a later trial, but he proposes that even then, results should be treated with caution. In the work of Gehan in 1978, it was suggested that historical bias could be overcome by using more complex statistical methods like analysis of covariance (ANCOVA) to allow for difference in patient characteristics. Pocock objects that the possibility of having poor data, a too small amount of features, and environmental changes are then still not accounted for. In Section 2.3, methods in which historical data can be included in a clinical trial are examined in more detail.

2.2.2 Sequential Analysis

Another way to reduce the sample size in a clinical trial is by *sequential analysis*, pioneered by Abraham Wald (Wald, 1945). In sequential analysis, results are analyzed at intermediate time points in the trial. When there is significant evidence that either hypothesis (null/alternative) is false, the trial is stopped with an early conclusion. As already stated, RCTs are prospective, hence intermediate testing is a possibility. When the same test is used repeatedly on an expanding dataset, the type I error rate (chance of rejection under the null) increases, as noted by Pocock (2013).

The reasoning behind this is as follows: Let T_k be the test statistic of the trial when k patient outcomes have been observed, and let $\mathcal{R}_k^\alpha \subset \mathbb{R}$ be the rejection region at level α for T_k under the null hypothesis $\mathbb{P}_0$ (i.e. $\mathbb{P}_0(T_k \in \mathcal{R}_k^\alpha) \leq \alpha$ for all k). Let n be the sample size and $k_1, \ldots, k_m$ be the number of patient outcomes at which the (m) interim analyses take place. Assuming that $T_s \in \mathcal{R}_s^\alpha$ if $T_k \in \mathcal{R}_k^\alpha$ for $s > k$, it then holds that:

$$\mathbb{P}_0(\cup_{i=1}^m \{T_{k_i} \in \mathcal{R}_{k_i}^\alpha\}) = \mathbb{P}_0(T_{k_1} \in \mathcal{R}_{k_1}^\alpha) + \mathbb{P}_0(\cup_{i=2}^m \{T_{k_i} \in \mathcal{R}_{k_i}^\alpha\} \cap \{T_{k_1} \notin \mathcal{R}_{k_1}^\alpha\}) \geq \alpha.$$

It can hence be the case that the type 1 error probability is higher than α, depending on the situation and number of interim tests. Thus, when this strategy is used and the dependence between test statistics is not accounted for, it is advised not to use too many interim analyses and a lower significance level per test than α (the level for the whole sequential testing procedure). In Pocock (2013), guidelines are given on what a good significance level is for a certain group of tests.

Sequential methods became more accepted in clinical trials after the success of the Beta-Blocker Heart Attack Trial (BHAT), which ended in June 1982. The use of a sequential procedure shortened the 4 year trial by 8 months. After this trial, the use of sequential methods in clinical trials increased, as well as research on this topic. More information on sequential analysis can be found e.g. in the second chapter of Lai (2001).

2.2.3 Multi-Armed Bandits

Research has also been done on optimization of the allocation rule in clinical trials. In this case, treatment allocation depends on the already observed outcomes. The situation of allocating the treatments in an optimal way is called a *multi-armed bandit problem* (a type of reinforcement learning). These problems are all analogous to being in a casino with multiple slot machines with different probabilities of success. The bandit, who has to maximize his profit, does not know these probabilities and has to estimate these by pulling different arms. The slot machine situation can be very general, e.g. the machine can give any type of payoff (e.g. real-valued, discrete), can have certain traits/covariates (contextual bandits) and the payoff distribution can change in time (nonstationary bandits). There is an exploration/exploitation tradeoff inherent to these problems. The ”profit” of the ”bandit” in clinical trials can be e.g. the number of patients being cured or the number of ”good” outcomes. In Friedman et al. (2010) and Spiegelhalter, Abrams, and Myles (2004), it is stated that these method are not often used in clinical trials and face a lot of criticism. Based on the latter source, the following objections to adaptive allocation are given:

1. **Multi armed bandits are less robust to model misspecifations as compared to (sequential) RCTs.**
   In multi armed bandit models, one has to make assumptions which then lead to a strategy which is (close to) optimal. The optimal strategy (and hence the trial conclusion) can however depend highly on these assumptions. Think for instance of a contextual (covariate-based) bandit problem where not all important features

are taken into account. RCTs and sequential testing are much more robust in this regard due to randomization/balancing.

2. **It is more difficult to implement a multi-armed bandit based design in practice.**
   A lot of communication is needed between researchers and doctors in order to determine the assigned control group for each new patient, communicate results etc. This additional difficulty in the trial design may make doctors more reluctant on letting their patients participate in the trial.

3. **Multi-armed bandit problems are sensitive to the chosen objective function.**
   The chosen objective function determines the optimal solution, so there is a larger element of choice as compared to the statistical methods, especially when multiple outcomes are involved.

4. **Multi-armed bandit trial designs may induce a larger sample size.**
   In an optimized clinical trial, the idea is that statistical inference is done on two (or more) groups with unequal sample sizes. For statistical tests, it is often seen that significant group outcome differences are observed the earliest when the group sizes are equal. Hence, in clinical trial optimization, a larger sample size is often needed as compared to in standard trials [1]. This larger sample size means that the trial takes longer to complete, hence the total patient population has to wait a longer time for the new medicine.

The importance of the above objections depends on the case at hand. If a trial already has a very large sample size, and the increase due to a multi-armed bandit approach will be negligible, the objection may be rejected.

2.3 Designs of History Controlled Trials

Due to the ethical concerns with randomization in clinical trials, there has been (and still is) an abundance of research on incorporating historical trial data in currently performed clinical trials. One of the earliest articles on inclusion of historical control data in current trials is by Pocock (1976). In this paper, the following six criteria are given for historical control data inclusion:

[1] Note that the objective is not always to minimize trial duration.

1. The treatment for the historical control group must be the same as that for the current control group.
2. The historical control group must have been a part of a recent clinical study which contained the same inclusion criteria as the current trial.
3. The methods of treatment evaluation/analysis must be the same.
4. The distributions of patient characteristics in both groups should be comparable.
5. The historical control group patients should have been treated at the same organization with roughly the same clinical investigators.
6. There must be no indications that factors other than treatment differences will lead to different results in the historical control and current treatment group.

These criteria are often taken as guidelines, as they are quite stringent, reasoning is often given why some of the criteria can be relaxed. In e.g. Lim et al. (2018), van Rosmalen, Dejardin, van Norden, Löwenberg, and Lesaffre (2018), Viele et al. (2014) and in chapter 6.9 of Spiegelhalter et al. (2004), surveys are given of historical control inclusion methods. The methods outlined in these surveys can be classified in five groups:

- Use the historical data as a so called *literature control*.
- Pool the historical control group data with the current data.
- A biased sample approach.
- Use a so called *power prior*.
- Assume a hierarchical model for the current and historical control group. This is also often called a *meta-analytic* approach.

The first approach, using literature controls, corresponds to the often used method employed in clinical research up until the 1950's. It assumes that enough historical data is available to give a reasonable estimate of the control parameter of interest, which will be denoted by θ_c. This parameter can be e.g. the mean of the distribution, the variance, or some quantile. In the case of θ_c being the mean, it is e.g. assumed that the sample mean of the historical data is exactly equal to the true mean of the control group outcomes. In the currently conducted trial, $H_0 : \theta_t = \theta_c$ is then tested against some

alternative hypothesis (e.g. $\theta_t - \theta_c = \delta$, $\theta_t - \theta_c < \delta$ or $\theta_t - \theta_c > \delta$), where θ_t is the same parameter of interest for the treatment group. As already stated, this procedure does not account for changes in the patient population, time dependent effects and change in inclusion/exclusion criteria between trials. In Viele et al. (2014), a simple example is given in which the power and type 1 error rate are seen to be very sensitive to the true parameter θ_c for this type of trial. It is clear that this method is an unreliable way to incorporate historical data in clinical trials, and hence in the following, the focus will lie on the latter four methods.

2.3.1 Pooling of Control Data

When control data is pooled, the historical control group data is pooled with the current control group data. Of course, if the historical control group data is not representative of the current control group data (e.g. due to trends in the data), tests based on this procedure may have less power or a larger probability of a type 1 error, as shown in Viele et al. (2014).
A safer procedure in this regard is often called the test-then-pool procedure. In this procedure, similarity between the historical and current control group data is first tested for. For example, one of the tests in Section 2.1.5 could first be applied to the historical and current control group. If this test rejects, the trial is conducted using only current control group data. If the test doesn't reject, historical and current control data are taken as one sample (pooled). In this way, the amount of historical data included in the trial is decided on in a data-driven way and there is more control on the power and type 1 error of the test than before. The downside to this procedure is that it is an all-or-nothing approach, either all or none of the historic data is included depending on whether the statistic exceeds some threshold value. A way of softening these decision boundaries is by using a Bayesian approach[2]. The *power prior* and *hierarchical modeling* approach are examples of such approaches.

2.3.2 Biased Sample Approach

The first instance of this is the work of Pocock in 1976 (Pocock, 1976). Pocock considered the case of two treatment groups (control and treatment) in a trial with quantitative outcome data. Let Y_i^T be the treatment group outcomes, Y_i^C be the control group out-

[2]For an introduction to Bayesian statistics, see Chapter 3.

comes, and Y_i^H the historical control group outcomes.
The following model was assumed:

$$\begin{aligned} Y_i^T &\overset{iid}{\sim} \mathcal{N}(\mu_T, \sigma_T^2) \\ Y_i^C &\overset{iid}{\sim} \mathcal{N}(\mu_C, \sigma_C^2) \\ Y_i^H &\overset{iid}{\sim} \mathcal{N}(\mu_C + \delta, \sigma_H^2) \\ \delta &\sim \mathcal{N}(0, \sigma_\delta^2). \end{aligned}$$

In the above, $\mathcal{N}(\mu, \sigma^2)$ is the class of normally distributed random variables with mean μ and standard deviation σ. All standard deviations are assumed to be known and all variables are assumed to be independent and identically distributed (which is denoted by $\overset{iid}{\sim}$). The value of interest is now $\mu_T - \mu_C$ (the treatment effect). From the above model, one can see why this approach is called the biased sample approach, the average effect based on Y_i^H alone would give a biased estimator (with bias δ) for the treatment effect.

Using improper uniform priors on μ_T and μ_C, the posterior distribution of $\mu_T - \mu_C$ was derived:

$$\mu_T - \mu_C \sim \mathcal{N}(\overline{Y}_T - \overline{Y}_{C,H},\ \sigma_T^2/N_T + V_{C,H})$$

where

$$\overline{Y}_{C,H} = \frac{(\sigma_H^2/N_H + \sigma_\delta^2)\,\overline{Y}_C + (\sigma_C^2/N_C)\overline{Y}_H}{\sigma_C^2/N_C + \sigma_H^2/N_H + \sigma_\delta^2}$$

$$V_{C,H} = \left((\sigma_C^2/N_C)^{-1} + (\sigma_\delta^2 + \sigma_H^2/N_H)^{-1}\right)^{-1}.$$

In the above, $\overline{Y}_T,\ \overline{Y}_C,\ \overline{Y}_H$ are the sample averages and $N_T,\ , N_C,\ N_H$ the number of patients in the different groups of the trial.

It is seen from the above that the certainty about the difference in means depends on the chosen sample sizes and standard deviations, especially on the chosen value of σ_δ, whose effect doesn't decrease with sample size. It is probably useful to try out different values of σ_δ to get a grip on the robustness w.r.t this parameter. Pocock proposes to set the standard deviations of the observed variables equal to the square root of the sample variance.

2.3.3 Power Prior Approach

Power priors form a way of incorporating historical data in a prior for a Bayesian analysis. The effect of this is that it softens the decision boundary for testing (Viele et al., 2014).

Power priors have emerged in research around the end of the 20th century, and a good summary about them can be found in either (Ibrahim, Chen, Gwon, & Chen, 2015) or (Ibrahim, Chen, et al., 2000). Let D denote the control group data and D_0 the historical control group data, let θ_c denote the current control group parameters and p_{θ_c} be a prior on the control group parameters. Let $f_{\theta_c|D_0}$ and $f_{\theta_c|D}$ be the posterior densities of the control group parameter given the historical data and the current data respectively. In a power prior model, the prior for the Bayesian analysis is given conditional on the historical data D_0. In fact, if one denotes this conditional prior with $p_{\theta_c|D_0}$, for some $a_0 \in [0, 1]$:

$$p_{\theta_c|D_0}(\theta_c) \propto f_{\theta_c|D_0}(\theta_c)^{a_0} p_{\theta_c}(\theta_c).$$

Hence the influence of the historical data on the prior is downweighted by some factor $a_0 \in [0, 1]$. The effect of this is that the posterior density of θ_c based on the historical data is flattened. The justification for this flattening could be for instance that in the current control group, the inclusion/exclusion criteria are different, but it is not known which members in the history control group would have been excluded. The power prior is a way to discount all information equally for the historical control group.
Consider for instance an experiment where for some $\theta \in [0, 1]$

$$I_1, \ldots, I_n \overset{iid}{\sim} Ber(\theta)$$

are recorded. Taking a standard uniform $\mathcal{U}(0, 1)$ on θ and letting $N = \sum_i I_i$, it follows that:

$$f_{\theta|N}(\theta) \propto \theta^N (1-\theta)^{n-N} \mathbb{1}_{[0,1]}(\theta).$$

Here $\mathbb{1}_E(x)$ is the indicator that $x \in E$. Hence, if one was to conduct a new experiment and use a power prior to include historical information on θ, the power prior would be:

$$p_{\theta|N}(\theta) \propto \theta^{a_0 N} (1-\theta)^{a_0(n-N)} \mathbb{1}_{[0,1]}(\theta).$$

Hence, the historical sample size was effectively multiplied with a_0.
From Bayes' rule, it follows that using the power prior, the posterior density of the parameters is given by:

$$f_{\theta_c|D,D_0,a_0}(\theta_c) \propto f_{\theta_c|D}(\theta_c) f_{\theta_c|D_0}(\theta_c)^{a_0} p_{\theta_c}(\theta_c). \tag{2.1}$$

Note that if the posterior was originally well defined (i.e. $f(\theta_c|D)p(\theta_c)$ is integrable for all θ_c), then the posterior following from using the power prior is also well defined. If one takes $a_0 = 1$, this means that the historical control data is pooled with the current control data, and $a_0 = 0$ means that the historical data is thrown away.

In the original definition of the power prior in Ibrahim et al. (2000), a_0 was fixed. In this case, this procedure can be seen as a method that lies between pooling the control data and ignoring it. With fixed a_0, the power prior method does not borrow the historical data dynamically, and hence this procedure still has a quite high risk of having a type 1 error or low power.
An approach to make the borrowing dynamic is to set a prior p_{θ_c,a_0} on (θ_c, a_0) and multiply the conditional densities in (2.1) with p_{θ_c,a_0} instead of p_{θ_c}. However, due to the fact that there is a term $f_{\theta_c|D_0}(\theta)^{a_0}$ in the posterior, it can be the case that the posterior becomes an improper function. A solution for this problem is given in (Neuenschwander, Branson, & Spiegelhalter, 2009), where it is required that $p_{\theta_c,\alpha_0} = p_{\theta_c}(\theta_c)p_{a_0}(a_0)$ and:

$$p_{\theta_c,a_0|D_0} = \frac{f_{\theta_c|D_0}(\theta_c)^{a_0} p_{\theta_c}(\theta_c)}{\int f_{\theta_c|D_0}(\theta_c)^{a_0} p_{\theta_c}(\theta_c) d\theta_c} p_{a_0}(a_0). \tag{2.2}$$

This conditional prior is integrable, which can be seen by integrating over the parameter space of θ first, and then for a_0.
The upside to this method is that the borrowing of historical data is now ”dynamic”. The downside is that the calculation of the numerator in (2.2) is often hard. Furthermore, the posterior distribution of a_0 does not depend on D, as remarked by (Hobbs, Carlin, Mandrekar, & Sargent, 2011). Hence these priors do not measure the so called *commensurability* between D_0 and D. Thus, this method of choosing a_0 is not very dynamic at all, and often overestimates the correspondence between the historical data and current control data. Another problem with power priors is the question on how to combine multiple trials, *hierarchical modeling* provides a better setup for this.

2.3.4 Hierarchical Modeling

The hierarchical model or meta-analytic approach was proposed in Neuenschwander, Capkun-Niggli, Branson, and Spiegelhalter (2010). It can deal with multiple historical trials, and assumes that the parameters $\theta_c^1, \ldots, \theta_c^K$ for these (K) trials are iid drawn from the same normal distribution:

$$\theta_c^1, \ldots, \theta_c^K \overset{iid}{\sim} \mathcal{N}(\mu_c, \sigma_c^2).$$

Often, the assumption of continuous support can be justified by making a transformation on model parameters that have bounded support (e.g. assuming a generalized linear model).

The data $D_1, \ldots, D_K$ corresponding to these trials are then sampled from some parametrized probability measures $\mathbb{P}_{\theta_c^1}, \ldots, \mathbb{P}_{\theta_c^K}$. When a prior on μ_c (often non-informative) and σ_c^2 are chosen, prediction of θ_c^{K+1} (the control parameter for the current trial) can be done using $D_0, \ldots, D_K$. If this prior for θ_c^{K+1} is then used for improving the trial design, this procedure is called *Meta Analytic Prediction*.

If after the trial, information $D_0, \ldots, D_{K+1}$ are used to strengthen the conclusion of the trial, this is called the *Meta Analytic Combined method*. Tests based on the meta analytic procedure have a smooth decision boundary where historical data is dynamically included. In terms of power and type 1 error probability, they perform better than the power prior in that both measures can assume a supremum/infimum between the curves given by pooling and exclusion of historical data. Furthermore, examples can be given in which these suprema/infima are less extreme as compared to the test-then-pool procedure (see (Viele et al., 2014)).

Sometimes, to account for the fact that there is a possibility that the historical and current controls do not correspond, the meta analytic posterior of θ_{K+1}^c is mixed with a (preferably) conjugate prior $p_{\theta_{K+1}^c}$ to construct the meta analytic prior for θ_{K+1}^c. For some chosen $w \in [0, 1]$ the following definition holds for this prior:

$$p_{\theta_{K+1}^c|(D_1,\ldots,D_K)} = w f_{\theta_{K+1}^c|(D_1,\ldots,D_K)} + (1-w) p_{\theta_{K+1}^c}.$$

This approach was introduced in (Schmidli et al., 2014), the effect is that the prior flattens, just as in the power prior approach. The parameter w can be based on expert opinion, or multiple values could be tried out.

In the next chapter, an introduction is given to the concepts in Bayesian statistics used in the remainder of this thesis.

Chapter 3

Principles of Bayesian Statistics

Bayesian statistics is a way to couple prior belief with inference from observations in a way that is based on Bayes' rule. This well-known rule is basically a one-step derivation from the definition of the conditional probability density.
Given two real valued random variables X, Y having a joint probability density denoted by $f_{(X,Y)}$, and marginal densities f_X and f_Y, the conditional density $f_{X|Y}$ of X given Y is defined for all $x, y \in \mathbb{R}$ as :

$$f_{X|Y}(x,y) := \begin{cases} \frac{f_{(X,Y)}(x,y)}{f_Y(y)} & \text{if } f_Y(y) > 0, \\ 0 & \text{else.} \end{cases}$$

Hence, using that $f_{Y|X}$ is similarly defined with the same joint density, one obtains *Bayes' rule*:

$$f_{Y|X}(x,y) = \begin{cases} \frac{f_{X|Y}(x,y) f_Y(y)}{f_X(x)} & \text{if } f_X(x) > 0, \\ 0 & \text{else.} \end{cases}$$

In Bayesian statistics, it is often the case that X is observed, and Y is unknown. A well-known result from probability theory states that $f_{Y|X}(X, y)$ (i.e. the conditional density of Y with a random first argument distributed as X) induces a conditional probability measure for Y. This basically states that when $X = x$ is observed, $f_{Y|X}(x, \cdot)$[1] is a probability density function.

Note that as $X = x$ was observed, it must be that $f_X(x) > 0$ and hence

$$f_{Y|X}(x,\cdot) = \frac{f_{X|Y}(x,\cdot) f_Y(\cdot)}{f_X(x)} \propto f_{X|Y}(x,\cdot) f_Y(\cdot). \tag{3.1}$$

[1] The $\cdot$ denotes the free variable.

The function on the left is often called the *posterior density* of Y given X. The $\propto$ sign above denotes that the latter function is merely a scaling of the former function. In Bayesian inference, the scaling can be taken out of consideration as it is known that $f_{Y|X}$ is a probability density and hence integrates to 1. In the remainder, this notation will be used often as it makes the derivations shorter. By convention, the random variable corresponding to this posterior density will be denoted by $Y|X$. Bayesian inference revolves around the posterior density $f_{Y|X}$, as it contains all information (e.g. the mean, variance, credible intervals) of Y given X.
In Bayesian statistics, it is often the case that (X, Y) follows a model, where X are the observations and Y are parameters of the model. An example of this would be the case where X is a vector of n independent $\mathcal{N}(\mu, 1)-$distributed random variables. In the Bayesian setting, Y would represent μ and inference is done on $\mu|X$. Note that from relation (3.1) it follows that one has to define a *prior (density)* p_μ on μ in order to do inference on μ[2]. When comparing this method to classical frequentist statistics it induces an additional modeling choice, namely specifying the prior. Furthermore, it seems quite counterintuïtive to treat the parameter μ as a random variable. However, there are two ways in which it can still be justified to do this:

- One way is to see the model as a *belief model* instead of a probability model, this correspond to *subjective Bayesian inference*. The prior now corresponds to the prior belief of the conductor of the experiment. In the previous example, the researcher might have a strong belief that μ lies in the interval $[100, 200]$ and hence might select a $\mathcal{N}(150, 400)$ prior on μ where almost 99% of the probability mass lies in $[100, 200]$. In cases where $\mu \in [100, 200]$, this can lead to quicker contraction of the posterior density for μ and hence a smaller sample size is required in order to do reliable inference on μ. The tradeoff here is that when $\mu \notin [100, 200]$, posterior contraction might be slower than in a setup where a less informative prior is used. One hence has to really think before making such *a priori* assumptions.

- In many cases, a frequentist-matching prior can be used (Mukerjee, 2003) to make the inference procedures of Bayesian and frequentist statistics agree (often to the extent of some approximation). This corresponds to *objective Bayesian inference*. The chosen priors in objective Bayesian inference are often *improper*, which means that they are not integrable, but can lead to integrable posteriors. In the normal variables example, a matching prior would be to take $p_\mu \propto 1$. As $f_{X|\mu}$

[2] In the following, to denote that a probability density is a prior, priors will be denoted by p.

is a product of normal densities, it is easy to see that

$$f_{\mu|X} \propto e^{-\frac{n(\mu-\overline{X}_n)^2}{2}}, \qquad \left(\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i\right). \tag{3.2}$$

Hence $\mu|X \sim \mathcal{N}(\overline{X}_n, 1/n)$ which can be compared to the frequentist situation, where we would *estimate* μ with $\overline{X}_n \sim \mathcal{N}(\mu, 1/n)$. [3]

In a sense, frequentist-matching priors do not give information on Y, which is why they are often more attractive from a frequentist point of view. Furthermore, Bayesian analysis can still be conducted, which also brings some advantages (e.g. in the case of a latent variable model).

As already stated after the first bullet point above, the choice of prior has a large influence on the contraction rate of the posterior. For computational purposes, the choice of prior can also matter a lot. For a given model, a special class of priors are the *conjugate priors* for that model. If a model $\mathcal{M}$ for (X, Y) induces the *likelihood function* $f_{X|Y}$, the class of conjugate prior distributions $\mathcal{F}_{\mathcal{M}}^{CP}$ for that model is given by all prior densities p_Y such that $f_{X|Y} \cdot p_Y \in \mathcal{F}_{\mathcal{M}}^{CP}$, and hence $f_{Y|X} \in \mathcal{F}_{\mathcal{M}}^{CP}$.

Consider for example the independent standard normal model. As $f_{X|\mu} \propto f_{\mu|X}$ in Relation (3.2), it follows that for a $\mathcal{N}(\mu_0, \sigma_0^2)$ prior on μ, the following relation holds for the posterior density:

$$\begin{aligned} f_{\mu|X}(\mu) &\propto e^{-\frac{1}{2}\left(\frac{(\mu-\overline{X}_n)^2}{1/n} + \frac{(\mu-\mu_0)^2}{\sigma_0^2}\right)} \propto e^{-\frac{1}{2}\left((n+1/\sigma_0^2)\mu^2 - 2\mu(n\overline{X}_n + \mu_0/\sigma_0^2)\right)} \\ &\propto e^{-\frac{1}{2}\frac{\left(\mu - \frac{n\overline{X}_n + \mu_0/\sigma_0^2}{n+1/\sigma_0^2}\right)^2}{1/(n+1/\sigma_0^2)}}. \end{aligned} \tag{3.3}$$

Hence for this prior, it holds that $\mu|X \sim \mathcal{N}\left(\frac{n\overline{X}_n + \mu_0/\sigma_0^2}{n+1/\sigma_0^2}, \frac{1}{n+1/\sigma_0^2}\right)$ and thus the class of normal distributions forms a class of conjugate priors for this model.

In the case that the class of conjugate priors is parametrized (i.e. there exists an isomorphism between $\mathcal{F}_{\mathcal{M}}^{CP}$ and $\mathbb{R}^d$ for some d), this is of particular use. In this case, calculating the posterior density of $Y|X$ amounts simply to updating the parameters of the prior. One example where this is useful is in on-line Bayesian estimation, where X is frequently updated. Bayesian inference with conjugate priors makes that the posterior density of Y can be quickly updated in this case.

[3] Note the difference in interpretation above, the statement $\mu|X \sim \mathcal{N}(\overline{X}_n, 1/n)$ says that if we consider the data X as if it came from the described normal model, the posterior density of the mean given that model is a $\mathcal{N}(\overline{X}_n, 1/n)-$distribution. In the latter case ($\overline{X}_n \sim \mathcal{N}(\mu, 1/n)$), the statement is much weaker, it says that *if* the data follows the described model, the sample mean has a $\mathcal{N}(\mu, 1/n)-$distribution. In the former case, there are no restrictions on the underlying distribution of the observations.

3.1 P values and Bayes Factors

3.1.1 p-values

As already stated in chapter 2, frequentist hypothesis tests are the most often performed tests in clinical trials. When a frequentist hypothesis test has been rejected, a corresponding p-value is often reported. First, a formal definition of a p-value is given, which can be found in e.g. Shafer, Shen, Vereshchagin, Vovk, et al. (2011) or Grünwald (2016):

Definition 1. A *p-value* under the *null hypothesis probability measure* $\mathbb{P}_0$ is defined as a random variable P such that for all $\alpha \in [0, 1]$:

$$\mathbb{P}_0(P \leq \alpha) \leq \alpha. \tag{3.4}$$

The above definition basically states that P is a good representation of the probability of attaining at most its own value. A p-value is called *precise* if the second inequality in Equation (3.4) is an equality. Often, in clinical research, a null hypothesis is rejected (and trial results are deemed significant) when $P \leq 0.05$.
In frequentist hypothesis testing, the p-value is often based on the test statistic T, which is a real-valued function of the data X. If Y is an iid sample of the data under $\mathbb{P}_0$, it follows from the above definition that:

$$P_T := \mathbb{P}_0(T(X) < T(Y)|X)$$

is a p-value based on T (see e.g. Grünwald (2016)).

P-values based on test statistics (denoted P_T) have faced a lot of criticism in the last two decades (see e.g. Wagenmakers (2007), Assaf and Tsionas (2018), Wasserstein, Lazar, et al. (2016) or Grünwald (2016)). The main points of critique are:

1. P_T **is hard to interpret**
   Most clinical researchers have a hard time interpreting p-values. In a survey published in JAMA (Windish, Huot, & Green, 2007), medical residents were asked to answer multiple-choice questions about subjects in statistics. One of these subjects was about interpreting p-values. Of all residents, only 58.8% was able to correctly answer questions about interpreting a P-value. Furthermore, while 88% of the residents expressed fair/complete confidence in their knowledge about P-values, only 62.8% of these residents could answer elementary p-value interpretation questions correctly.

Often, a p-value is misunderstood as the probability that the null hypothesis H_0 is true. This is not the case as from Definition 3.4, it is seen that P only has an interpretation under $\mathbb{P}_0$. Hence, a P value can only be interpreted assuming that the null hypothesis is true.
A further problem with interpretability exists when $P > 0.05$. From Definition 3.4, it can be seen that, if the P-value is not precise, the probability under the null of P being e.g. less than 0.8 can still be smaller than 0.05. However, many researchers interpret a large P value as a large amount of evidence for the null hypothesis.
Furthermore, from Definition 3.4, it can be seen that no alternative hypothesis is considered. Hence it is also not the case that a large p-value indicates a lot of evidence for any alternative hypothesis, which is also sometimes assumed.

2. **Tests based on P_T will always reject with increasing sample size.**
Note that as "all models are wrong" (Box, Luceno, & del Carmen Paniagua-Quinones, 2011), the probability of rejection based on the p-value will tend to 1 with increasing sample size. This is a fallacy of the p-value, but can also be seen as a fallacy of frequentist hypothesis testing as a whole. This phenomenon is described in e.g. M. Lin, Lucas Jr, and Shmueli (2013).

3. **P_T is sensitive to the sampling procedure**
Here the example of optional continuation of experiments can help to give a clarification. Suppose that a researcher is conducting an experiment and after it is completed he obtains a p-value of 0.06. As this p-value is very close to 0.05, he decides to restart the experiment in the hope that the p-value will become lower than 0.05 If this researcher calculates the p-value as if he had done the whole experiment without an intermediate observation, he is in the wrong. He namely has to calculate the p-value conditional on the already observed data. Calculating these conditional p-values can be difficult in some cases.

4. **The value of P_T is often based on asymptotic results.**
Many tests, such as e.g. the chi-squared test or the independent two-samples t-test are based on asymptotic results. These tests form a good approximation when the sample size is large. However, in clinical trials, large sample sizes are often not observed.

Due to these criticisms, alternative hypothesis testing procedures, such as Bayes factor testing are often proposed, which will now be introduced.

3.1.2 Bayes Factors

For a dataset D, let a model $\mathcal{M}$ for this data denote a probability density $f_{D|\mathcal{M}}$ on the support of the data. The posterior evidence for a model $\mathcal{M}$ given the data can now be defined as:

$$f_{\mathcal{M}|D} := \frac{f_{D|\mathcal{M}} p_{\mathcal{M}}}{p_D}.$$

In the above, $p_{\mathcal{M}}$ is the prior probability of model $\mathcal{M}$ being true and p_D is the prior for the data D.

Now, let H_0 and H_1 be models for D. If $p_{H_0|D} = 0$ iff $p_{H_1|D} = 0$, the Bayes factor for testing H_0 vs. H_1 is now defined as the ratio between the posterior evidence for the models:

$$B(D) = \frac{f_{H_0|D}(D)}{f_{H_1|D}(D)} = \frac{f_{D|H_0}(D)p_{H_0}}{f_{D|H_1}(D)p_{H_1}} \frac{p_D(D)}{p_D(D)} = \frac{f_{D|H_0}(D)p_{H_0}}{f_{D|H_1}(D)p_{H_1}}.$$

The prior probabilities for the two models are often taken equal ($p_{H_0} = p_{H_1}$) and in this case the Bayes factor is equal to the likelihood ratio of the data under the two models:

$$B(D) = \frac{f_{D|H_0}(D)}{f_{D|H_1}(D)}. \tag{3.5}$$

This gives the Bayes factor a clear interpretation, when $B(D) = \alpha$, the data is α times more likely to have been generated under model H_0 as compared to under H_1.

Let $\mathbb{P}_i$ be the probability measure under H_i for $i \in \{0, 1\}$, and denote with $\mathbb{E}_i$ the corresponding expectation. If D lies in $\mathbb{R}^n$ for some n, it follows from Markov's inequality, as $B(D)$ is nonnegative, that for all $\alpha \in (0, 1)$:

$$\mathbb{P}_1(B(D) \geq 1/\alpha) \leq \alpha \mathbb{E}_1 [B(D)] = \alpha \int_{\mathbb{R}^n} \frac{f_{D|H_0}(x) f_{D|H_1}(x)}{f_{D|H_1}(x)} dx = \alpha \int_{\mathbb{R}^n} f_{D|H_0}(x) dx = \alpha. \tag{3.6}$$

Hence, under H_1, large values of $B(D)$ are unlikely.

Similarly, for $\alpha \in (0, 1)$:

$$\mathbb{P}_0(B(D) \leq \alpha) = \mathbb{P}_0(1/B(D) \geq 1/\alpha) \leq \alpha \mathbb{E}_0 [1/B(D)] = \alpha \int_{\mathbb{R}^n} \frac{f_{D|H_1}(x) f_{D|H_0}(x)}{f_{D|H_0}(x)} dx$$

$$= \alpha \int_{\mathbb{R}^n} f_{D|H_1}(x) dx = \alpha.$$

Thus, under H_0, small values of $B(D)$ are unlikely.

These properties link Bayes factors to p-values. Indeed, from the derivation above and Relation (3.4), it can be seen that $B(D)$ is a p-value under H_0. Furthermore, by writing $\mathbb{P}_1(B(D) \geq 1/\alpha)$ as $\mathbb{P}_1(1/B(D) \leq \alpha)$, it can be seen that $1/B(D)$ is a p-value

under H_1. Hence, H_0 and H_1 can be rejected (but never simultaneously) based on $B(D)$.

Notice that a lot of downsides of using p-values based on test statistics do not play a role with Bayes factors. When a Bayes factor is large/small, this indeed means that the probability of the null hypothesis is large/small and the inverse relation holds for the alternative hypothesis, hence downside 1 for p-values on Page 26 is taken out of consideration.

Furthermore, inference in Bayesian statistics does not revolve around the assumption that the chosen model is true. Hence, the problem with rejection of the null hypothesis with large sample sizes (issue 2 on Page 27) is no issue anymore.

Also, Bayes factors are also useful for optional continuation. Following e.g. Ly, Etz, Marsman, and Wagenmakers (2018), if a researcher intermediately looks at the results of the test and decides to continue, the Bayes factor at the endpoint will just be:

$$\frac{f_{D_{(2)}|(D_{(1)},H_0)}}{f_{D_{(2)}|(D_{(1)},H_1)}} = \frac{f_{D|H_0}}{f_{D|H_1}} \cdot \frac{f_{D_{(1)}|H_1}}{f_{D_{(1)}|H_0}}. \tag{3.7}$$

In the above, $D_{(1)}$ denotes the observed data prior to the intermediate decision, and $D_{(2)}$ denotes the data after the decision was made. Note that in order to make the decision, the researcher has already calculated $p_{D_{(1)}|H_i}$ and hence at the end of the experiment only has to calculate $p_{D|H_i}$ to determine the conditional Bayes factor. The equation above also states that the Bayes factor for the total dataset D is just the multiplication of the Bayes factor for $D_{(1)}$ and the conditional Bayes factor for $D_{(2)}|D_{(1)}$.

Lastly, most results in Bayesian statistics hold with finite sample size. Often, no asymptotic approximations need to be made. Hence the last issue on Page 27 is also solved.

In a sequential analysis, the Bayes factor can be sequentially updated based on new information. The result of this can be viewed as a discrete time stochastic process. Quite recently, a class of stochastic processes has been investigated that encompasses this process. This class is called the *test martingales* (Grünwald, 2016). Studying these types of processes could produce some results that apply to a situation more general than just clinical trials. In Appendix F, some (known) results about test martingales are stated.

In this research, the Bayes factor is approximated using samples from a Markov Chain Monte Carlo (MCMC) algorithm, hence these class of algorithms, in particular the Metropolis-Hastings algorithm and Gibbs sampling, are now elaborated on.

3.2 Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) techniques are often used for sampling from $f_{Y|X}$ when the unobserved variable Y in Bayesian analysis is multidimensional and/or no analytical expression can be found for the joint posterior density of $Y|X$. The idea in MCMC is to create an ergodic, discrete time, (often) continuous state Markov Chain with stationary distribution equal to the probability measure induced by the density $f_{Y|X}$. One of the techniques to construct such a Markov Chain is the so called Metropolis-Hastings algorithm.

3.2.1 The Metropolis-Hastings Algorithm

First, define $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ to be a transition kernel iff $g(x, \cdot)$ is a probability density on $\mathbb{R}^d$ for all $x \in \mathbb{R}^d$. Denote with $\mathbb{P}_x^g$ the induced probability measure at x. In the Metropolis-Hastings algorithm, as a first step a transition kernel g supporting Y (i.e. $g(z, y) > 0$ for all $z, y \in \text{supp}(Y)$) is chosen. Now, for all z, y in the support of Y, the acceptance ratio is defined as follows:

$$a(z, y) = \min\left(1, \frac{f_{Y|X}(y)}{f_{Y|X}(z)} \frac{g(y,\ z)}{g(z,\ y)}\right) = \min\left(1, \frac{f_{X|Y}(x, y) p_Y(y)}{f_{X|Y}(x, z) p_Y(z)} \frac{g(y,\ z)}{g(z,\ y)}\right). \tag{3.8}$$

Note that the second equality above only holds because of proportionality, the normalization constants divide out. Let $\mathbb{P}_0$ be an (initial) probability measure on the support of Y, the Metropolis-Hastings algorithm can now be written as follows:

Algorithm 1 Metropolis-Hastings

Inputs:
  Observed data X;
Initialize:
  draw $Y_0 \sim \mathbb{P}_0$;
for $k \in \{1, \ldots, N\}$ **do**
  draw $Y^* \sim \mathbb{P}^g_{Y_{k-1}}$;
  draw $U_k \sim \mathcal{U}(0,1)$;
  if $U_k \leq a(Y_{k-1}, Y^*)$ **then**
    set $Y_k := Y^*$;
  else
    set $Y_k := Y_{k-1}$;
  end if
end for
Outputs:
  $Y_1, \ldots, Y_N$.

In the above, $\mathcal{U}(a,b)$ is the class of uniformly distributed random variables on the interval $[a,b]$ for $a < b$ in $\mathbb{R}$.
The following well-known result now follows:

Theorem 1. *The Markov chain induced by the Metropolis-Hastings algorithm has a stationary distribution equal to the probability distribution induced by $f_{Y|X}$.*

Proof. As the (would be) stationary probability density is $f_{Y|X}$ and the kernel for the Markov chain is

$$g'(z,y) = g(z,y)a(z,y) = g(z,y)\min\left(1, \frac{f_{Y|X}(y)}{f_{Y|X}(z)}\frac{g(y,z)}{g(z,y)}\right),$$

it follows that when $f_{Y|X}(z)$ and $f_{Y|Z}(y)$ are positive:

$$\begin{aligned} f_{Y|X}(z)g'(z,y) &= f_{Y|X}(z)g(z,y)\min\left(1, \frac{f_{Y|X}(y)}{f_{Y|X}(z)}\frac{g(y,z)}{g(z,y)}\right) \\ &= \min\left(f_{Y|X}(z)g(z,y), f_{Y|X}(y)g(y,z)\right) \\ &= f_{Y|X}(y)g(y,z)\min\left(1, \frac{f_{Y|X}(z)}{f_{Y|X}(y)}\frac{g(z,y)}{g(y,z)}\right) = f_{Y|X}(y)g'(y,z). \end{aligned}$$

Note that the above equations do not hold when Y is not supported by g. If one of the conditional densities is zero, the acceptance ratio is zero and as a result, the above

equality also holds. Hence detailed balance holds and according to Kelly's Lemma (Kelly, 2011)[4], the Markov chain is ergodic with a stationary distribution induced by probability density $f_{Y|X}$. □

Note that the above theorem does not give any guarantees that the induced chain is **ergodic** (in short, that in some sense convergence to the stationary distribution indeed occurs). For this, stronger conditions about the generating chain and model are needed (see e.g. Roberts and Smith (1994)). The rate of convergence to the stationary distribution is determined by the choice of transition kernel g. An alternative MCMC method, the *Gibbs sampling algorithm*, is now described in the next section.

3.2.2 Gibbs Sampling

In the Gibbs sampling algorithm, Y is marginalized into smaller parts $Y = [Y_1, \ldots, Y_k]'$ which may or may not be univariate. The transition kernel is now chosen such that sequentially $Y_i|Y_{(-i),X}$ is sampled for $i \in \{1, \ldots, k\}$ (here $Y_{(-i)}$ is Y with the part corresponding to Y_i removed).
Denote with $d_i(\cdot)$ the projection of a vector onto the same vector with the part corresponding to Y_i removed (e.g. $d_i(Y) = Y_{(-i)}$).
The global transition kernel is now equal to a kernel where sampling from the transition kernels $g_1, \ldots, g_k$ is sequentially performed, where:

$$g_i(z, y) := \frac{f_{Y|X}(y)}{f_{Y_{(-i)}|X}(d_i(y))} \mathbb{I}(d_i(z) = d_i(y)).$$

Here, $\mathbb{I}(F)$ is the indicator that the *statement/event* F holds.

Now, it is seen that when sampling according to g_i is performed, the corresponding rejection rate α_i is given by:

$$a_i(z, y) = \min\left(1, \frac{f_{Y|X}(y)}{f_{Y|X}(z)} \frac{g_i(y, z)}{g_i(z, y)}\right) = \min\left(1, \frac{f_{Y|X}(y)}{f_{Y|X}(z)} \frac{f_{Y|X}(z) f_{Y_{(-i)}|X}(d_i(y)) \mathbb{I}(d_i(z) = d_i(y))}{f_{Y|X}(y) f_{Y_{(-i)}|X}(d_i(z)) \mathbb{I}(d_i(z) = d_i(y))}\right)$$
$$= \min\left(1, \frac{f_{Y_{(-i)}|X}(d_i(y)) \mathbb{I}(d_i(z) = d_i(y))}{f_{Y_{(-i)}|X}(d_i(z)) \mathbb{I}(d_i(z) = d_i(y))}\right).$$

[4] Originally, Kelly's lemma holds for a *discrete state* Markov chain with *transition/stationary probabilities*, but by considering the limit for the result on Markov chains on finer and finer partitions of the state space, one can show that the result also holds with a *continuous state space* when using the *transition/stationary densities*.

Due to the fact that new samples Y are generated according to the kernel g_i, it is always the case that $d_i(Y_{k-1}) = d_i(Y^*)$ in the Markov chain. Hence, the above term is always equal to 1 in the Gibbs sampling algorithm. Hence, under some additional conditions, the Gibbs sampling algorithm is equivalent to the Metropolis-Hastings algorithm with acceptance ratio equal to 1. This does not mean however that Gibbs samplers are optimal in the sense of convergence. The acceptance ratio is 1 but it can still be the case that the sampled process shows a lot of autocorrelation and is thus very sensitive to the initial state. An advantage of the Gibbs sampler is that one can for instance make use of conjugacy in each Gibbs sampling step, resulting in a fast evaluation of posterior marginal densities, especially when the posterior distribution is parametrized.

3.3 Laplace-Metropolis Approximation

Returning to Bayes factors, in elaborate models, it is hard to analytically evaluate $f_{X|H_i}$ from Equation 3.5. However, if the hypotheses (or models) describe a joint distribution of X and some latent variable $Y \in \mathbb{R}^d$, the Bayes factor can be written as follows:

$$B(D) = \frac{\int_{\mathbb{R}^d} f_{D|(Y,H_0)}(D,y) f_{Y|H_0}(y) dy}{\int_{\mathbb{R}^d} f_{D|(Y,H_1)}(D,y) f_{Y|H_1}(y) dy}. \tag{3.9}$$

Both integrals above can be approximated with the so called *Laplace-Metropolis* estimator (Lewis & Raftery, 1997). This estimator revolves around the Laplace approximation, which is based on a second order Taylor approximation for a special class of functions. Namely, for a bounded unimodal function $h \in \mathcal{C}^2(\mathbb{R}^p)$:

$$\int_{\mathbb{R}^p} e^{h(u)} du \approx (2\pi)^{p/2} |\mathbf{H}^*|^{1/2} e^{h(\mathbf{u}^*)}.$$

In the above, $\mathbf{u}^* = \underset{u \in \mathbb{R}^p}{\arg\max}\, h(u)$ and $\mathbf{H}^* = -\left[\text{Hess}(h)(\mathbf{u}^*)\right]^{-1}$.
In the above, Hess stands for the Hessian operator.
Taking $h_i(u) = \log(f_{D|(Y,H_i)}(D,u) f_{Y|H_i}(u))$ and d_i the number of incorporated latent variables under H_i for $i \in \{0,1\}$, it follows that when h is bounded, twice continuously differentiable and unimodal that:

$$\int_{\mathbb{R}^d} f_{D|(Y,H_i)}(D,y) f_{Y|H_i}(y) dy \approx (2\pi)^{d_i/2} |\mathbf{H}_i^*|^{1/2} f_{D|(Y,H_i)}(D,\mathbf{u}_i^*) f_{Y|H_i}(\mathbf{u}_i^*).$$

In the above, $\mathbf{u}_i^*$ is equal to $\underset{u \in \mathbb{R}^d}{\arg\max} \log(f_{D|(Y,H_i)}(D,u) f_{Y|H_i}(u))$, note that this is exactly the *maximum a posteriori estimator* of the parameters under H_i. Furthermore, $\mathbf{H}_i^*$ is the

negative inverse Hessian of h at $\mathbf{u}^*$. The expectation of the inverse of this is exactly the so called Fisher information matrix (see e.g. Van der Vaart (2000)). The inverse of the Fisher information matrix is (under regularity conditions) the asymptotic covariance matrix in a central limit theorem for the maximum likelihood estimator. Hence, one could state that $|\mathbf{H}_i^*|$ represents the variance in the posterior density of the parameters.

The formula above leads to the following approximation for the logarithm of the Bayes factor:

$$\log(B(D)) \approx \frac{d_0 - d_1}{2}\log(2\pi) + \frac{1}{2}\left(\log\left(|\mathbf{H}_0^*|\right) - \log\left(|\mathbf{H}_1^*|\right)\right) + (l_0(X, \mathbf{u}_0^*) - l_1(X, \mathbf{u}_1^*)) + (\lambda_0(\mathbf{u}_0^*) - \lambda_1(\mathbf{u}_1^*)). \tag{3.10}$$

In the above, $l_i(u) = \log(f_{D|(Y,H_i)}(D, u))$ and $\lambda_i(u) = \log(f_{Y|H_i}(u))$ for all u.
Approximation 3.10 can be divided in four interpretable parts. The part $(d_0 - d_1)\log(2\pi)/2$ corresponds to the difference in dimensions of the models H_0 and H_1. From this, it is seen that the Lapplace approximated Bayes factor gives more evidence toward models with more parameters. The part $\left(\log\left(|\mathbf{H}_0^*|\right) - \log\left(|\mathbf{H}_1^*|\right)\right)/2$ evaluates the difference in covariance/spread in the posterior densities. Curiously, the Laplace Metropolis Bayes factor gives more evidence toward the model with the most amount of posterior variance. This seems counterintuïtive as more posterior covariance implies more posterior uncertainty about the parameter values. No intuïtive explanation for this was given in (Lewis & Raftery, 1997) or found elsewhere.

The part corresponding to $(l_0(X, \mathbf{u}_0^*) - l_1(X, \mathbf{u}_1^*))$ gives more evidence to the estimate that gives the highest posterior density and the part corresponding to $(\lambda_0(\mathbf{u}_0^*) - \lambda_1(\mathbf{u}_1^*))$ gives more evidence to the estimate that has the highest prior density.
In Lewis and Raftery (1997), it is advised to estimate $\mathbf{u}^*$ with the multivariate posterior median, and to estimate $\mathbf{H}^*$ with the weighted variance matrix estimate described in Rousseeuw and Van Zomeren (1990). The former is also done in this research, however it took too much time in practice to compute the weighted variance matrix, hence the standard covariance estimator was used.

Chapter 4

Treatment Induced Correlation in a Survival Model

In this chapter, the multivariate survival model combining the research of X. Lin and Wang (2010) and Fox et al. (2017) will be introduced. This model was not yet seen in literature. In the first section, an introduction to survival analysis is given, after which the model for censored survival data proposed in X. Lin and Wang (2010) is introduced. In the last two sections, the multivariate extension of the survival model is defined, after which it is explained how the baseline is modeled.

4.1 Survival Analysis

Time to event or survival responses are analyzed using so called *survival analysis* (or reliability/duration analysis in engineering/economics). The events of interest are e.g. recurrence to the clinic, time until equipment breaks down, or time to default in economics. The subjects arrive to the trial at some time point, and are followed until the event of interest has been observed, or the event-time is *right censored*. This censoring basically entails all events that cause the time-to-event measurement to stop early. Examples of this are e.g. patients leaving the trial, stopping of the trial or death due to other causes than the disease. Next to right censoring, left censoring and interval censoring can also occur. Left censoring occurs when only an upper bound on the event time is known. This can happen for instance when patients are added to a clinical trial retrospectively (e.g. when the inclusion criteria are redefined) and the event has already occurred before the patients have entered the trial. Interval censoring occurs when only the interval in which the event time lies is known. This can happen when monitoring is not done continuously but only between certain time points.

Denote the event times for subject i with T_i. The *observations* for subject i in survival analysis are now time intervals $[L_i, R_i)$ in which T_i lies. When no censoring occurs, this interval becomes degenerate (where the interval is now taken to be closed), but in the case of left/right/interval censoring, it contains more than one point. In addition to time intervals, the observations for a patient can also contain some covariates, denoted by X_i for patient i. These covariates can be e.g. age, gender, and length.

4.2 The Survival Model Introduced By Lin and Wang

Let $\mathcal{T} = \{-\infty, 0, t_1, \ldots, t_l, \infty\}$ denote a set of time points, where $t_i \in \mathbb{R}$ for all i. In the model proposed in X. Lin and Wang (2010), it is assumed that right and left censoring can occur. Furthermore, it is assumed that the event-times are interval censored with endpoints in $\mathcal{T}$. Hence, the observations are nondegenerate intervals $[L_i, R_i)$, with $L_i, R_i \in \mathcal{T}$ such that $R_i > L_i$. In the case $L_i = -\infty$, the event-time is left censored and when $R_i = \infty$ the event-time is right censored. It is assumed that event-times for patients cannot be both left and right censored. Furthermore, p subject-linked covariates are observed, and the vector of covariates for patient i is denoted $X_i \in \mathbb{R}^p$. Let there be N patients under consideration (i.e. let the sample size be N).

It is assumed that the latent (unobserved) event-times T_i are conditionally independently distributed. This means that given every parameter in the model, the event times are independently distributed. Furthermore, it is assumed that there exists $\beta \in \mathbb{R}^p$ and a continuous nondecreasing *baseline hazard function* $\alpha(\cdot)$ such that for all patients i:

$$T_i = \alpha^{-1}(X_i\beta + \epsilon_i). \tag{4.1}$$

In the above equation, $\{\epsilon_1, \ldots, \epsilon_N\}$ is a set of iid random variables. The above model can be seen as a linear regression model, but with a possibly nonlinear increasing *transformation function* α^{-1}, which is equal to the inverse of the baseline hazard function. The term $X_i\beta$ basically shifts the argument of the inverse baseline to the left/right and hence decreases/increases the expected event time respectively (note that α^{-1} has to be an increasing function).

By choosing ϵ_i to have a certain distribution, some often used models in survival analysis can be obtained. When the standard logistic distribution is assumed for ϵ_i, one obtains the logistic survival model (X. Lin & Wang, 2010). Furthermore, when α is assumed to be differentiable and when the standard extreme value distribution is assumed, one obtains the same partial likelihood as in the often used Cox proportional hazards model

(Pettitt, 1984). Lastly, when the distribution of ϵ_i is assumed to be standard normal, one obtains the survival model introduced in X. Lin and Wang (2010).

To further highlight the generality of model 4.1 further, note that model 4.1 equals a linear regression model when T_i is observed and α equals identity. Furthermore, when α is allowed to be discontinuous, the model also contains ordinal regression models when α is taken to be a certain step function with jumps at the thresholds. This situation can of course be seen as a limiting case for model 4.1, as a combination of step functions can be approximated arbitrarily well by functions in the class of increasing continuous functions.

From now on, the model introduced in (X. Lin & Wang, 2010) is considered, hence, it is assumed that $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0,1)$. In the article, the latent event-times T_i are transformed by the baseline α, resulting in latent normally distributed random variables Z_i:

$$Z_i := \alpha(T_i) \overset{iid}{\sim} \mathcal{N}(X_i\beta, 1).$$

Note that the following relation now holds between the observed intervals and the latent variables Z_i:

$$\{L_i = L,\ R_i = R\} \iff \{Z_i \in [\alpha(L), \alpha(R))\}.$$

Note that the above model looks a lot like an ordinal regression model. However, the *thresholds* $\alpha(T)$ for $T \in \mathcal{T}$ are restricted to be in the range of the baseline function α. Furthermore, due to left or right censoring, both L and R need to be specified, as the left survival endpoint does not determine the right survival endpoint. This model is now extended to a multivariate framework in the next section.

4.3 The Multivariate Survival Model

In this section, a multivariate extension of the model proposed by (X. Lin & Wang, 2010) will be introduced, which has strong links to the model in (Wu & Wang, 2019). The extension is based around the assumption that there now exist groups of patients for which a treatment effect is drawn independently from the same normal distribution. This idea has links to the hierarchical or meta-analytic models considered in section 2.3.

Let there be n patient groups with m_j patients in group j. The group number is indicated with a subscript j for all variables (e.g. T_i is now denoted T_{ij}). It is now assumed that:

$$Z_{ij} = \alpha(T_{ij}) = X_{ij}\beta + \theta_j + \epsilon_{ij}, \qquad \theta_j \overset{iid}{\sim} \mathcal{N}(\mu, \tau), \qquad \epsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0,1). \tag{4.2}$$

In the above, the treatment effect for group j is θ_j. One can now define $\beta^{(2)} := [\mu, \beta]'$ and $X_{ij}^{(2)} = [1, X_{ij}]'$ to arrive at another linear combination $(X_{ij}^{(2)})'\beta^{(2)}$ defining the expected value of Z_{ij}. Furthermore, the term $\theta_j - \mu$ can be integrated out to obtain a multivariate distribution for the variables Z_{ij} in each group j. Define $Z_j = [Z_{1j}, \ldots, Z_{m_j}]$, and $X_j^{(2)}$ the matrix with row i equal to $X_{ij}^{(2)}$ it then follows that, in vector form:

$$Z_j = X_j^{(2)}\beta^{(2)} + E_j, \qquad E_j \sim \mathcal{N}_{m_j}(0, I_{m_j} + \tau J_{m_j}). \tag{4.3}$$

In the above, J_{m_j} is the square matrix in $\mathbb{R}^{m_j \times m_j}$ with every element equal to 1.

When examining Relation (4.3), it is seen that it allows τ to take on more values than in Relation (4.2), where $\tau \geq 0$ is required. To see this, let $\mathbf{1}_{m_j}$ denote the vector in $\mathbb{R}^{m_j}$ with every element equal to 1 and let e_i^j denote the i-th unit vector in $\mathbb{R}^{m_j}$, it holds that:

$$\begin{aligned}(I_{m_j} + \tau J_m)\mathbf{1}_{m_j} &= (1 + m_j\tau)\mathbf{1}_{m_j} \\ (I_{m_j} + \tau J_m)(e_i^j - \mathbf{1}_{m_j}/m_j) &= (e_i^j - \mathbf{1}_{m_j}/m_j).\end{aligned}$$

hence, considering the eigenvalue decomposition, the matrix $(I_{m_j} + \tau J_{m_j})$ is positive definite iff $\tau > -1/m_j$ for all j. Hence, under (4.2), $\tau < 0$ is not possible but under (4.3) it is. Due to this property, the choice was made to henceforth consider 4.3 as the model for the latent vectors Z_j. The reason for this will be explained shortly.

It is seen from Equation 4.3 that besides the often assumed difference in average outcomes μ due to treatment, it is also assumed that there are groups of patients in the trial that have *correlated outcomes* due to the treatment procedure. A situation in which this can happen is when no standard treatment is currently available. Namely, in this case, when patients are treated, suddenly some aspects of the treatment procedure have an effect on the outcomes (e.g. the surgeon or physician), inducing correlated outcomes in patient groups. Detecting these correlated groups should then mean that the treatment indeed has had an effect, otherwise these factors shouldn't matter. The assumption here is that all other effects (e.g. age, gender) are accounted for by the regression $X_{ij}\beta$.

The other case in which correlation is induced is when slightly different versions of the same treatment are applied in a clinical trial. If it matters which version is administered to patients, slight differences in performance in patient groups should be detectable. If there is evidence for $\tau > 0$, this indicates that there could possibly be one best (or worst) version of the treatment. If negative correlation is detected ($\tau < 0$), it would mean that a version works well for some patients, but less well for others. Personalized medicine might be an option in this case. Lastly, when there is evidence for

$\tau = 0$ in this case, it means that it doesn't matter which version is administered. If one version is cheaper than the others, this means that the medicine can be sold for a lower price. Hence, following Fox et al. (2017), the idea is now to construct a Bayes factor for comparing the evidence of $\tau = 0$ against $\tau \neq 0$.

The main advantage of taking 4.3 to be the model for Z_j is now that it ensures that under the null hypothesis ($H_0 : \tau = 0$) the parameter τ lies in the interior of the parameter set. This is an advantage as Bayes factors can break down when parameters are on the boundary of the parameter set (Pauler, Wakefield, & Kass, 1999).

Rewriting $X := X^{(2)}$ and $\beta := \beta^{(2)}$, the model considered in the remainder is hence:

Definition 2. *The Survival Model with Treatment Induced Covariance*
For all $j \in \{1, \ldots, n\}$, $L, R \in \mathcal{T}$ such that $R > L$ and $m \in \{1, \ldots, m_j\}$:

$$\begin{aligned} &\{L_{ij} = L,\ R_{ij} = R\} \iff \{Z_{ij} \in [\alpha(L), \alpha(R))\}, \\ &Z_j \sim \mathcal{N}_{m_j}(X_j\beta, I_{m_j} + \tau J_{m_j}). \end{aligned}$$

To reiterate, the observations in the above model are the *events* $\{L_{ij} = L,\ R_{ij} = R\}$. Note that when $\tau \geq 0$ in the above model, the second line above is still equivalent to Equation 4.2. Hence, the model in Definition 2 can also be rewritten as:

Definition 3. *Survival Model with Treatment Induced Covariance (Alternative definition)*
For all $j \in \{1, \ldots, n\}$, $L, R \in \mathcal{T}$ such that $R > L$ and $m \in \{1, \ldots, m_j\}$:

$$\begin{aligned} &\{L_{ij} = L,\ R_{ij} = R\} \iff \{Z_{ij} \in [\alpha(L), \alpha(R))\}, \\ &Z_j \sim \begin{cases} \mathcal{N}_{m_j}(X_j\beta + \theta_j \mathbf{1}_{m_j}, I_{m_j}) & \text{for } \tau \geq 0, \\ \mathcal{N}_{m_j}(X_j\beta, I_{m_j} + \tau J_{m_j}) & \text{for } \tau < 0. \end{cases} \end{aligned}$$

In this research, it was found that making this case distinction based on τ can reduce the computational effort for inference under this model. Hence, in some sampling steps in Chapter 5, the choice will be made to sample some parameters under the above definition of the model. In the sections of Chapter 5, it will always be explained under which model specification the parameters are sampled.
For readers of this thesis, might be jarring to go from the model described by Equation 4.2 to Defintion 2 to Definition 3. At first, the plan in this research was actually to do inference strictly under Definition 2. However, it was seen that it takes a very long time to do inference on the baseline function α in this case. At first, an extension of the method outlined in X. Lin and Wang (2010) was investigated, which could be efficiently

performed under 2. However, in this research it was falsely assumed that the inference method outlined in the article was wrong (see Appendix D for more information) and hence other methods of inference were investigated. In hindsight however (after completion of the thesis), the inference method of Lin and Wang turned out to be correct and hence other inference methods can be investigated. A last thing of note is the work in (Wu & Wang, 2019), where this approach is taken for positive correlation only.

4.4 Modeling the Baseline as a Combination of Integrated Splines

As in X. Lin and Wang (2010), the choice is made to model the continuous nondecreasing function α with a translated linear combination of I-splines $B_l^{d,\mathbf{t}}$:

$$\alpha(t) = \gamma_0 + \sum_{l=1}^{k} \gamma_l B_l^{d,\mathbf{t}}(t) \qquad \forall t \in \mathbb{R}. \tag{4.4}$$

The functions $B_l^{d,\mathbf{t}}$, as well as their parameters, are defined in/around Equation (4.5), but first the M-splines will be elaborated on.

The family of M-splines is often used for function estimation as it contains the solution to a general penalized regression problem (see Theorem 2 below). Furthermore, it will be seen that the M-splines are only nonzero on a closed interval. The effect of this in e.g. density estimation problems is that for the cases where there are no observations, the approximation will be equal to 0, which is often desired.

The family $\mathcal{M}_{d,\mathbf{t}}$ of M-spline functions is defined by a degree $d \in \mathbb{N}$ and a sequence of $k + d$ knots $\mathbf{t} = (t_1, \ldots, t_{k+d})$ such that:

- $t_1 = t_2 = \cdots = t_d$
- $t_{k+1} = t_{n+2} = \cdots = t_{k+d}$
- $t_l < t_{l+d}$ for $l \in \{1, \ldots, k\}$

The set $\{t_{d+1}, t_{d+2}, \ldots, t_k\}$ are called the *interior knots* for the M-splines $\mathcal{M}_{d,\mathbf{t}}$.
Given the sequence of knots $\mathbf{t}$ and the degree d, the set of corresponding M-splines

$\mathcal{M}_{d,\mathbf{t}} = \{M_1^{d,\mathbf{t}}, \ldots, M_k^{d,\mathbf{t}}\}$ can be recursively defined for all $l \in \{1, \ldots, k\}$ and $t \in \mathbb{R}$:

$$M_l^{1,\mathbf{t}}(t) = \frac{\mathbb{1}_{[t_l, t_{l+1})}(t)}{t_{l+1} - t_l},$$
$$M_l^{d,\mathbf{t}}(t) = \frac{d}{d-1} \cdot \frac{(t - t_l) M_l^{d-1,\mathbf{t}}(t) + (t_{l+d} - t) M_{l+1}^{d-1,\mathbf{t}}(t)}{(t_{l+d} - t_l)} \qquad \forall d \in \mathbb{N}.$$

The M-splines for the sequence of knots $\mathbf{t}$ and degree d have the following characterizing properties:

- $M_l^{d,\mathbf{t}}(t) \geq 0$ with equality if $t \notin [t_l, t_{l+k}]$,
- The domain of every function $M_l^{d,\mathbf{t}}$ is $\mathbb{R}$ and $\int_{\mathbb{R}} M_l^{d,\mathbf{t}}(t)dt = 1$,
- $M_l^{d,\mathbf{t}} \in \mathcal{C}^{d-2}(\mathbb{R})$ for $d \geq 2$, where $\mathcal{C}^q(S)$ is the set of all q-times continuously differentiable functions on the set S,
- When restricted to $[t_l, t_{l+d})$, $M_l^{d,\mathbf{t}}$ is a polynomial of degree $d-1$,
- For $d \geq 2$, any function $f \in \mathcal{C}^{d-2}(\mathbb{R})$ that, when restricted to any of the intervals $[t_l, t_{l+1})$ is a polynomial of degree $d-1$, can be expressed as a linear combination of $M_1^{d,\mathbf{t}}, \ldots, M_k^{d,\mathbf{t}}$.

The latter property hence implies that $\mathcal{M}^{d,\mathbf{t}}$ forms a basis for the vector space $\mathcal{P}^{d,\mathbf{t}}$ of $\mathcal{C}^{d-2}(\mathbb{R})$-functions that are polynomials of degree $d-1$ when restricted to the sets $[t_l, t_{l+d})$ for all l ($d \geq 2$). This property is particularly of use when considering the next Theorem.

Theorem 2. *Consider data points* $(t_1, y_1), \ldots, (t_n, y_n)$, $\lambda \in (0, \infty)$ *and*

$$\mathcal{F} = \left\{ f \in \mathcal{C}^2(\mathbb{R}) \mid \int_{\mathbb{R}} \left(\frac{\partial^2}{\partial t^2} f(t) \right)^2 dt < \infty \right\}.$$

It holds that:

$$\underset{f \in \mathcal{F}}{\arg\min} \left(\left[\sum_{l=1}^{n} (y_l - f(t_l))^2 \right] + \lambda \int_{\mathbb{R}} \left(\frac{\partial^2}{\partial t^2} f(t) \right)^2 dt \right) \in \mathcal{P}^{4,\mathbf{t}}.$$

In the above, the internal knots of $\mathbf{t}$ *are* $t_1, \ldots, t_n$*.*

Proof. See (Wahba, 1990). □

This result hence that under a regularity penalty (the second derivative term), the so called *cubic*[1] splines contain the best approximation to the data out of a large class of functions. This makes the M-spline basis very appealing for interpolation, and due to them integrating to 1, the M-splines are often used for probability density estimation.
The corresponding cumulative probability functions are the I-splines (integrated splines):

$$B_l^{d,\mathbf{t}}(t) = \int_{-\infty}^{t} M_l^{d,\mathbf{t}}(s)ds. \tag{4.5}$$

The d-th degree $I-$splines with knots $\mathbf{t}$ are denoted by $B_l^{d,\mathbf{t}}$ instead of the usual $I_l^{d,\mathbf{t}}$ to prevent confusion with the identity matrix in the remainder.

[1] Note that a function in $\mathcal{P}^{4,\mathbf{t}}$ is a third order polynomial between the knots.

Chapter 5

Inference for the Survival Model with Additional Correlation

In this section, it is explained how samples are generated from the posterior density $f_{(\beta,\gamma,\tau)|(L,R,X)}$ under Model 2 and the specification for α in the last section. The posterior samples are generated with a Metropolis-Hastings algorithm. Data augmentation is used, meaning that as an intermediate step, the latent variables Z from the last chapter are sampled. Each group of parameters Z, τ, β, γ is sampled sequentially, and if possible, a Gibbs sampling step is performed to do so. The main purpose of the sampling procedure in this research is to estimate the Bayes factor for testing $H_0 : \tau = 0$ vs. $H_1 : \tau \neq 0$ with the Laplace-Metropolis estimator described in Section 3.3.

In the following, denote with subscript j the corresponding matrix for group j (e.g. $X_j \in \mathbb{R}^{m_j \times p}$) and with two subscripts the patient-specific vector (e.g. $X_{ij} \in \mathbb{R}^p$ for patient i in group j). Furthermore, no subscript indicates the set of all group matrices (e.g. $X = \{X_1, \ldots, X_n\}$). The degree d and knots $\mathbf{t}$ of the I-splines are fixed in the algorithm. Furthermore, it is assumed that the patient groups have been determined beforehand, and are hence also fixed.

5.1 Initialization

In this section, it is described how the Metropolis-Hastings algorithm is initialized. This corresponds to determining $\mathbb{P}_0$ in Algorithm 1. The choice of the initial distribution has a large influence on the speed of convergence of the Metropolis-Hastings algorithm. If the starting point is a point with very small posterior density, the induced Markov process can take a long time to converge to the limit distribution.
The initial distribution is determined by maximum likelihood estimation under the survival

model where $\tau = 0$ (i.e. $p_\tau = \delta_0$, the Dirac measure at 0), as in this case one can efficiently evaluate the parameter likelihood. As τ is fixed at zero, the estimator for (β, γ) remains to be found. Maximum likelihood estimation is performed with the nonlinear minimizer `nlm` in the software package `R`.
Let $\vec{L}_{ij} \in \mathbb{R}^{k+1}$ be the vector of B-spline values such that $\vec{L}_{ij}\gamma = \alpha(L_{ij})$, and similarly define $\vec{R}_{ij}$, the likelihood to maximize can now be written as:

$$f_{(L,R)|(X,\beta,\gamma)}(\beta,\gamma) = \prod_{i=1}^{n}\prod_{j=1}^{m_j}(\Phi(\vec{R}_{ij}\gamma - X_{ij}\beta) - \Phi(\vec{L}_{ij}\gamma - X_{ij}\beta)).$$

For numerical stability, the negative log likelihood $-\log(f_{(L,R)|(X,\beta,\gamma)})$ is minimized.

As the minimization method `nlm` uses a Newton-type algorithm, the Hessian is calculated at each point, and the Hessian at the optimum is an optional output value for this algorithm. This is convenient because the Hessian at the optimum also provides an estimate of the Fisher information matrix $I_{(\beta,\gamma)}$ at n observations for this model. The inverse of this matrix is the covariance matrix for the MLE in a central limit theorem (see (Van der Vaart, 2000), page 65).

Hence, if the MLE is denoted by $\hat{\theta}$ and the estimated Fisher information matrix at n observations with $\hat{I}_\theta$, the choice is now made to sample the starting values of $[\beta, \gamma]$ in the Metropolis-Hastings algorithm as:

$$[\beta_0, \gamma_0] \sim \mathcal{N}_{p+k+1}\left(\hat{\theta}, \hat{I_\theta}^{-1}\right). \tag{5.1}$$

Furthermore, the initial value of τ, and (hence) also all treatment effects θ_j, are set to 0. The choice to sample the Metropolis-Hastings starting point from a distribution instead of setting it equal to the MLE is made because sometimes, one wants to diagnose convergence of the Markov Chain with e.g. the Gelman-Rubin diagnostic. In this case it is often advised to initialize the different Markov chains at different points (Gelman, Rubin, et al., 1992).

After initializing the Metropolis-Hastings algorithm, the parameter groups $Z, [\beta, \theta], \tau, \gamma$ are sampled sequentially on the other parameter groups and the observations. Theorem (1) then dictates that asymptotically, the sampled values of (β, γ, τ) have to converge to the joint posterior distribution.

5.2 Sampling $Z|(\tau, \gamma, \beta, X, L, R)$

In this section, it is described how $Z|(\tau, \gamma, \beta, X, L, R)$ is sampled according to Definition 2. Z is sampled according to this definition instead of Definition 3 because it can be

done more efficiently as no case distinction for τ is needed. Given $(\tau, \gamma, \beta, X_j, L_j, R_j)$, each Z_j is a multivariate truncated normally distributed vector. The truncation interval for element Z_{ij} is $[\alpha(L_{ij}), \alpha(R_{ij}))$. Truncated multivariate normal vectors cannot be sampled jointly in an efficient manner. Instead, the often used solution is to marginally sample Z_{ij} conditional on $(Z_j^{(-i)}, \tau, \gamma, \beta, X_j, L_j, R_j)$ in a Gibbs sampling step.
Following Corollary 4.1 with $a = 1,\ b = \tau$, it holds that for all i, j:

$$Z_{ij}|(Z_j^{(-i)}, \gamma, \beta, X, L, R) \sim \mathcal{N}(\mu_{ij}, \sigma_j^2)|_{[\alpha(L_{ij}),\alpha(R_{ij}))}, \tag{5.2}$$

where

$$\mu_{ij} = (X_j\beta)_i + \frac{\tau}{(m_j - 1)\tau + 1} \sum_{k \neq i} (Z_{kj} - (X_j\beta)_k),$$

$$\sigma_j^2 = \frac{m_j\tau + 1}{(m_j - 1)\tau + 1}.$$

Now, for all $\mu, a, b \in \mathbb{R}$, $\sigma \in (0, \infty)$ such that $a < b$, $X \sim \mathcal{N}(\mu, \sigma^2)|_{[a,b)}$ can be efficiently sampled as:

$$X = \Phi^{-1}\left(\Phi\left(\frac{a-\mu}{\sigma}\right) + U \cdot \left(\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)\right)\right)\sigma + \mu, \qquad U \sim \mathcal{U}(0, 1).$$

5.3 Sampling $\tau|(\beta, X, Z)$

As Z is a sufficient statistic for τ, the parameter τ is conditioned only on (β, X, Z) in this step of the Metropolis-Hastings algorithm. The choice was made to sample τ under Definition 2 of the survival model as this can be done more efficiently. For this, an approach similar to the one described in Fox et al. (2017) is followed.

Define $\epsilon_j = Z_j - X_j\beta \sim N_{m_j}(0, I_{m_j} + \tau J_{m_j})$ and consider the orthonormal Helmert matrices

$$H_j = \begin{bmatrix} \frac{1}{\sqrt{m_j}} & \frac{1}{\sqrt{m_j}} & \frac{1}{\sqrt{m_j}} & \cdots & \frac{1}{\sqrt{m_j}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 & \cdots & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \frac{1}{\sqrt{m_j(m_j-1)}} & \frac{1}{\sqrt{m_j(m_j-1)}} & \frac{1}{\sqrt{m_j(m_j-1)}} & \cdots & -\frac{m_j-1}{\sqrt{m_j(m_j-1)}} \end{bmatrix} \qquad \forall j.$$

It now follows that

$$V_j = \frac{H_j\epsilon_j}{\sqrt{m_j}} \sim \mathcal{N}_{m_j}\left(0,\ \frac{I_{m_j}}{m_j} + \tau e_1^j \left(e_1^j\right)'\right).$$

This means that only the first coefficient of V_j depends on τ and one hence only has to condition on $V_{1j} \sim \mathcal{N}(0, 1/m_j + \tau)$ for all j to do inference on τ.
If the m_j are all the same (assume equal to m), and if one takes an inverse gamma (IG(α_0, β_0)) prior on $\tau + 1/m$, it holds that τ has the *shifted inverse gamma distribution* as described in Fox et al. (2017):

$$(\tau + 1/m)|(V_{11}, \ldots, V_{1n}) \sim \text{IG}\left(\alpha_0 + n/2,\ \beta_0 + \frac{\sum_{j=1}^{n} V_{1j}^2}{2}\right).$$

Which means that τ can be sampled efficiently a posteriori by sampling $1/m + \tau$ as above, and subtracting $1/m$ from it.
When considering the case of clinical trials however, it is not realistic to assume that the group sizes are all equal.
In the case of unequal group sizes the following relation holds for the posterior density of τ :

$$f_{\tau|(V_{11},\ldots,V_{1n})} \propto p_\tau(\tau) \prod_{j=1}^{n} \frac{1}{\sqrt{1/m_j + \tau}} e^{-\frac{V_{1j}^2}{2(1/m_j+\tau)}} \mathbb{I}(\tau \geq -1/m_j).$$

Define $m^* := \max(\{m_j : j \in \{1, \ldots, n\}\})$ and note that the posterior density of τ can only be positive on $[-1/m^*, \infty)$.

The proposed method for sampling from this density is with a Metropolis-Hastings step with a bounded Gaussian random walk transition kernel:

$$g_\sigma(\tau^{(1)}, \tau^{(2)}) \propto e^{-\frac{\left(\tau^{(1)}-\tau^{(2)}\right)^2}{2\sigma^2}} \mathbb{I}\left(\tau^{(1)}, \tau^{(2)} \geq -1/m^*\right) \qquad \forall \tau^{(1)}, \tau^{(2)} \in \mathbb{R}. \tag{5.3}$$

The above kernel is chosen as the acceptance rate for the proposals can be tuned with the step size parameter σ. Similar to the procedure described in (Fox, 2010), page 84, σ is updated dynamically for a fixed number of times (20). The initial value of σ is 1. To get the fraction of accepted proposals in the Metropolis-Hastings algorithm to roughly 50%, σ is multiplied by 2 if the fraction of accepted proposals in the last 50 iterations is higher than 50% and it is divided by 2 otherwise. Note that an increase of σ induces that the proposals deviate more from the current value, and hence a lower amount of accepted proposals can be expected. Note that only after the last time σ was updated. This ”delay of Markovness” was seen to provide no serious consequences for convergence of the Markov chain.
The transition kernel in Relation (5.3) is symmetric, hence when dropping the indicators (which will always equal 1 by virtue of the chosen kernel) and using an independent (of

β, γ) prior p_τ on τ, the acceptance ratio becomes:

$$a(\tau^{(1)}, \tau^{(2)}) = \min\left(1, \frac{p_\tau\left(\tau^{(2)}\right)}{p_\tau\left(\tau^{(1)}\right)}\left[\prod_{j=1}^{n}\frac{\sqrt{1/m_j+\tau^{(1)}}}{\sqrt{1/m_j+\tau^{(2)}}}e^{-\frac{V_{1j}^2}{2}\left(\frac{1}{1/m_j+\tau^{(2)}}-\frac{1}{1/m_j+\tau^{(1)}}\right)}\right]\right). \quad (5.4)$$

5.4 Sampling $\beta|(\tau, X, Z)$

As Z is a sufficient statistic for β, the parameter β is conditioned only on (β, X, Z) in this step of the Metropolis-Hastings algorithm. The choice is made to sample β under Definition 3, hence a case distinction is made based on τ. The choice was made to simulate β under this definition because when $\tau > 0$, both θ and β can be sampled simultaneously. This is more efficient compared to the case where $\beta|(\tau, X, Z)$ and $\theta|(\beta, \tau, X, Z)$ are sampled separately.

Define $G \in \mathbb{R}^{s\times n}$, where $s = \sum_{j=1}^{n} m_j$ to be the group indicator matrix such that

$$G_{ij} = \mathbb{I}(\text{patient } i \text{ is in group } j).$$

Denoting with $\vec{Z}$ the concatenation of all vectors Z_j and $\vec{X}$ the vertical concatenation of all matrices X_j. Now, when the last sampled value of τ is nonnegative, it holds that:

$$\vec{Z} = [\vec{X}, G]\begin{bmatrix}\beta\\ \theta\end{bmatrix} + \vec{\epsilon}, \qquad \vec{\epsilon} \sim \mathcal{N}_s(0, I_s)$$

which is a standard linear regression model.

Now, as a priori $\theta_j \sim \mathcal{N}_j(0, \tau)$ when $\tau \geq 0$ in Model 3, a $\mathcal{N}_n(0, \tau I_n)$-prior is taken for each θ_j. Combining this prior on θ with an independent (of θ, γ, τ) multivariate $\mathcal{N}_p(\mu_0, \Sigma_0)$ prior on β, it is now a well-known result (see Gelman et al. (2013) page 356) that the posterior distribution of $[\beta, \theta]$ is as follows:

$$\begin{aligned}
&(\beta, \theta)|(\tau, X, Z) \sim \mathcal{N}_{p+n}(\mu, \Sigma), \quad (5.5)\\
&\Sigma = \begin{bmatrix}\vec{X}'\vec{X} + \Sigma_0^{-1} & \vec{X}'G\\ G'\vec{X} & G'G + \tau^{-1}I_n\end{bmatrix}^{-1},\\
&\mu = \Sigma\begin{bmatrix}\Sigma_0^{-1}\beta_0 + \vec{X}'\vec{Z}\\ G'\vec{Z}\end{bmatrix}.
\end{aligned}$$

When $\tau < 0$, the second line in Definition 3 holds. Let $\tilde{Z}_j = (I_{m_j} + \tau J_{m_j})^{-1/2}Z_j$ and $\tilde{X}_j = (I_{m_j} + \tau J_{m_j})^{-1/2}X_j$. Now denoting with $\vec{\tilde{Z}}$ the concatenation of $\tilde{Z}_j$ and with $\vec{\tilde{X}}$ the concatenation of $\tilde{X}_j$, it then holds that:

$$\vec{\tilde{Z}} = \vec{\tilde{X}}\beta + \vec{\epsilon}, \qquad \vec{\epsilon} \sim \mathcal{N}_s(0, I_s).$$

Now, with the same prior on β, it holds again from Gelman et al. (2013) that:

$$\begin{aligned}&\beta|(\tau, X, Z) \sim \mathcal{N}_p(\mu, \Sigma),\\ &\Sigma = (\vec{X}'\vec{X} + \Sigma_0^{-1})^{-1},\\ &\mu = \Sigma(\Sigma_0^{-1}\beta_0 + \vec{X}'\vec{Z}).\end{aligned} \tag{5.6}$$

5.5 Sampling $\gamma|(\beta, \theta, X, \tau, L, R)$

The choice was made to do inference on γ according to Definition 3. This decision was made because γ is sampled according to a Metropolis-Hastings step, and when $\tau > 0$ the acceptance ratio can be evaluated more efficiently under Definition 3 as compared to under Definition 2. Furthermore, the choice was made to condition γ only on (β, X, τ, L, R) and possibly also θ (if the last sampled τ is nonnegative), and hence not Z. This is allowed (as Z is latent) and this is often seen to speed up the Markov chain convergence (see e.g. Fox (2010), page 84).

The transition kernel chosen for this step is a partially nonnegative Gaussian random walk, for all $\gamma^{(1)}, \gamma^{(2)} \in \mathbb{R}^{k+1}$:

$$g_\sigma(\gamma^{(1)}, \gamma^{(2)}) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{\left(\gamma_1^{(1)} - \gamma_1^{(2)}\right)^2}{2\sigma_1^2}} \prod_{i=2}^{k+1} \sqrt{\frac{2}{\pi\sigma_i^2}} e^{-\frac{\left(\gamma_i^{(1)} - \gamma_i^{(2)}\right)^2}{2\sigma_i^2}} \mathbb{I}(\gamma_i^{(1)}, \gamma_i^{(2)} \geq 0). \tag{5.7}$$

Note that the first element of γ is sampled with a Gaussian random walk, and hence can be negative. The reason for choosing this kernel is the same as the reasoning on Page 46, the variance parameters $\sigma_1, \ldots, \sigma_{k+1}$ are updated in the same manner as σ in the last section.

Note that the chosen transition kernel g_σ is symmetric, hence from Equation (3.8), when taking independent (of β, τ) priors p_γ on γ, it follows that the acceptance ratio is equal to the (capped) ratio of the posterior densities for the old and new γ values:

$$a(\gamma^{(1)}, \gamma^{(2)}) = \min\left(1, \frac{f_{\gamma|(\beta,X,\tau,L,R)}(\gamma^{(2)})}{f_{\gamma|(\beta,X,\tau,L,R)}(\gamma^{(1)})}\right) = \min\left(1, \frac{f_{(L,R)|(\beta,X,\tau,\gamma)}(\gamma^{(2)})p_\gamma(\gamma^{(2)})}{f_{(L,R)|(\beta,X,\tau,\gamma)}(\gamma^{(1)})p_\gamma(\gamma^{(1)})}\right). \tag{5.8}$$

Now, as the likelihood of (L, R) only depends on γ through α, it follows that

$$f_{(L,R)|(\beta,X,\tau,\gamma)} = f_{(L,R)|(\beta,X,\tau,\alpha)}.$$

Let $\Phi_n(E; \mu, \Sigma)$ denote the probability that a multivariate normal vector with mean μ and covariance matrix Σ lies in $E \in \mathcal{B}(\mathbb{R}^n)$. It now holds that under Model 2:

$$f_{(L,R)|(\beta,X,\tau,\alpha)} = \prod_{j=1}^{n} \Phi_{m_j} \left(\prod_{i=1}^{m_j} [\alpha(L_{ij}), \alpha(R_{ij})); X_j\beta, I_{m_j} + \tau J_{m_j} \right). \tag{5.9}$$

When the last sampled τ is nonnegative, the treatment effects θ_j are used to determine $f_{(L,R)|(\beta,\theta,X,\alpha)}$. Namely, in this case, following Definition 3:

$$f_{(L,R)|(\beta,\theta,X,\alpha)} = \prod_{j=1}^{n} \prod_{i=1}^{m_j} (\Phi(\alpha(R_{ij}) - X_{ij}\beta - \theta_j) - \Phi(\alpha(L_{ij}) - X_{ij}\beta - \theta_j)). \tag{5.10}$$

When the last sampled value of τ is negative, another technique is used. The following result namely holds:

Theorem 3. *For all $j \in \{1, \ldots, n\}$ and $\tau \in (-1/m_j, \infty)$:*

$$\begin{aligned} &\Phi_{m_j} \left(\prod_{i=1}^{m_j} [\alpha(L_{ij}), \alpha(R_{ij})); \; X_j\beta; \; I_{m_j} + \tau J_{m_j} \right) \\ &= \mathbb{E} \left[\frac{e^{\frac{\tau \left(\sum_{i=1}^{m_j} V_{ij} \right)^2}{2(1+m_j\tau)}}}{\sqrt{1+m_j\tau}} \right] \prod_{i=1}^{m_j} (\Phi(\alpha(R_{ij}) - X_{ij}\beta) - \Phi(\alpha(L_{ij}) - X_{ij}\beta)). \end{aligned} \tag{5.11}$$

In the above, $V_{ij} \sim \mathcal{N}(0,1)|_{[\alpha(L_{ij})-X_{ij}\beta,\, \alpha(R_{ij})-X_{ij}\beta)}$ for all i.

Proof. See Appendix C. □

The above integral could not be simplified any further, however, a lower and upper bound were found for $\tau < 0$, namely:

Corollary 3.1. For all $j \in \{1, \ldots, n\}$ and $\tau \in (-1/m_j, 0)$, let

$$P_j = \Phi_{m_j} \left(\prod_{i=1}^{m_j} [\alpha(L_{ij}), \alpha(R_{ij})); \; X_j\beta; \; I_{m_j} + \tau J_{m_j} \right).$$

If $\tilde{L}_{ij} = \alpha(L_{ij}) - X_{ij}\beta$, and $\tilde{R}_{ij}$ is defined similarly, it then holds that:

$$\frac{e^{\frac{\tau \left(I_j^{(1)} + I_j^{(2)} \right)}{2(m_j\tau+1)}}}{\sqrt{m_j\tau+1}} \prod_{i=1}^{m_j} (\Phi(\tilde{R}_{ij}) - \Phi(\tilde{L}_{ij})) \leq P_j \leq \frac{1}{\sqrt{m_j\tau+1}} \prod_{i=1}^{m_j} (\Phi(\tilde{R}_{ij}) - \Phi(\tilde{L}_{ij})) \tag{5.12}$$

where

$$I_j^{(1)} = \left[\sum_{i=1}^{m_j} \frac{\tilde{L}_{ij} e^{-\frac{\tilde{L}_{ij}^2}{2}} - \tilde{R}_{ij} e^{-\frac{\tilde{R}_{ij}^2}{2}}}{\sqrt{2\pi}\left(\Phi(\tilde{R}_{ij}) - \Phi(\tilde{L}_{ij})\right)} \right] + m_j,$$

$$I_j^{(2)} = \sum_{i=1}^{m_j} \sum_{k \neq i} \left(\frac{e^{-\frac{\tilde{L}_{ij}^2}{2}} - e^{-\frac{\tilde{R}_{ij}^2}{2}}}{\sqrt{2\pi}\left(\Phi(\tilde{R}_{ij}) - \Phi(\tilde{L}_{ij})\right)} \right) \left(\frac{e^{-\frac{\tilde{L}_{kj}^2}{2}} - e^{-\frac{\tilde{R}_{kj}^2}{2}}}{\sqrt{2\pi}\left(\Phi(\tilde{R}_{kj}) - \Phi(\tilde{L}_{kj})\right)} \right).$$

Proof. Again, see Appendix C. □

These bounds are used to construct lower/upper bounds for the acceptance ratio in Equation 5.8. If the sampled uniform random variable lies above/under the upper/lower bound, the new value of γ is rejected/accepted without having to determine 5.9 twice. Whenever the sampled uniform random variable is between the bounds for the acceptance ratio, 5.8 is evaluated using 5.11 where the expectation is evaluated with Monte Carlo integration.

In the case that τ has been sampled nonnegatively, the likelihood of the data (given θ) can be evaluated quite efficiently, hence the choice is made to accept or reject each new element γ_i of γ drawn under 5.7 with 5.8. In the case that τ is sampled negatively, evaluating the likelihood takes a longer time, hence the whole vector γ, drawn under 5.7 is evaluated at once with 5.8.

Developing the sampling procedure for γ in this chapter provided the largest challenge in the methodological part of this research. First, an extension of the method proposed in X. Lin and Wang (2010) was explored. It was however (falsely) assumed during this research that the inference method proposed in this article contained an error (see Appendix D).

Another considered technique was to sample γ uniformly in the domain of the spline matrix such that the induced *thresholds* correctly separate the latent variables Z. However, it was seen that while the sampling procedure for this is fast, the correlation between Z and γ becomes very high. The effect is that convergence of the Markov chain to the limit distribution takes too long.

It now became clear that γ had to be sampled according to a Metropolis-Hastings step, and the acceptance ratio 5.8 had to be calculated efficiently in some way. For this, an efficient way had to be found to compute 5.9. Due to the fact that the number of variables n can be high, the integral $\Phi_n(E; \mu; \Sigma)$ cannot be computed efficiently in general (see e.g. Gassmann, Deák, and Szántai (2002)). However, by making use of the simple correlation structure, Formula 5.11 was derived, which can be used to

evaluate the probability more efficiently with Monte Carlo techniques. This method was also chosen as it was also seen to induce quick Markov Chain convergence in practice, by independence of Z. This procedure of sampling spline coefficients has not yet been seen in literature.

5.6 Summary of Inference Method

To summarize the previous sections, the steps to sample $(\beta, \gamma, \tau)|(L, R, X)$ can be summarized in the following algorithm:

Algorithm 2 Metropolis-Hastings sampler for the survival model with treatment-induced correlation (Definition 2)

Inputs:
Observed data (L, R, X), number of iterations M, N_σ, priors p_γ, p_τ and regression coefficient prior parameters μ_0, Σ_0;
Initialize:
Sample $(\beta_0, \gamma_0) \sim \mathcal{N}_{p+k+1}\left(\hat{\theta}, \hat{I}_\theta^{-1}\right)$ as in (5.1);
Set $\tau_0 := 0$ and $\theta_i := 0$ for all $i \in \{1, \ldots, n\}$;
Sample $Z_{ij} \sim \mathcal{N}(X_{ij}\beta, 1)|_{[\alpha(L_{ij}), \alpha(R_{ij}))}$ for all i, j;
Set $c_i := 0$ and $\sigma_i := 1$ for all $i \in \{1, \ldots, k+2\}$;
for $l \in \{1, \ldots, M\}$ **do**
    for $j \in \{1, \ldots, n\}$ **do**
        for $i \in \{1, \ldots, m_j\}$ **do**
            Sample $Z_{ij}|(Z_j^{(-i)}, \gamma_{l-1}, \beta_{l-1}, X, L, R, \tau_{l-1}) \sim \mathcal{N}(\mu_{ij}, \sigma_j^2)|_{[\alpha(L_{ij}), \alpha(R_{ij})}$
                as in (5.2) ;
        end for
    end for
    Sample τ^* according to the Markov chain described by g_σ in (5.3) with $\sigma = \sigma_1$;
    Sample $U \sim \mathcal{U}(0, 1)$ and let a be as in (5.4);
    if $U \leq a(\tau_{l-1}, \tau^*)$ **then**
        Set $\tau_l := \tau^*$;
        Set $c_1 := c_1 + 1$;
    else
        Set $\tau_l := \tau_{l-1}$;
    end if

if $\tau_l \geq 0$ **then**
  Sample $(\beta_l, \theta)|(\tau, X, Z)$ as in (5.5);
else
  Sample $\beta_l|(\tau, X, Z)$ as in (5.6);
end if
if $\tau_l \geq 0$ **then**
  Set $\gamma := \gamma_{l-1}$;
  for $i \in \{1, \ldots, k+1\}$ **do**
    Sample γ_i^* univariate according to the Markov chain described by g_σ in (5.7) with $\sigma = [\sigma_2, \ldots, \sigma_{k+2}]$;
    Let γ^* be such that $\gamma_j^* = \gamma_j$ for $j \neq i$ and $\gamma_j^* = \gamma_i^*$ if $i = j$;
    Sample $U \sim \mathcal{U}(0,1)$ and let a be as in (5.8) where $f_{(LR)|(\beta,\theta,X,\tau,\alpha)}$ is calculated with Equation 5.10.
    if $U \leq a(\gamma, \gamma^*)$ **then**
      Set $\gamma := \gamma^*$;
      Set $c_{i+1} : c_{i+1} + 1$;
    end if
  end for
  Set $\gamma_l = \gamma$;
else
  Sample γ_i^* multivariate according to g_σ in Equation 5.7 with $\sigma = [\sigma_2, \ldots, \sigma_{k+2}]$;
  Sample $U \sim \mathcal{U}(0,1)$ and let a be as in (5.8), where $f_{(L,R)|(\beta,X\tau,\alpha)}$ from Equation 5.9 is approximated by Monte Carlo integration (100 samples);
  if $U \leq a(\gamma, \gamma^*)$ **then**
    Set $\gamma := \gamma^*$;
    Set $c_{i+1} : c_{i+1} + 1$ for all $i \in \{1, \ldots, k+1\}$;
  end if
end if
if $\mod(l, 50) = 0$ and $l < 20N_\sigma$ **then**
  for $i \in \{1, \ldots, k+2\}$ **do**
    if $c_i/50 > 0.5$ **then**
      $\sigma_i := 2\sigma_i$;
    else
      $\sigma_i := \sigma_i/2$;
    end if
  end for
end if
end for
Outputs:
$\beta_1, \ldots, \beta_M, \gamma_1, \ldots, \gamma_M$ and $\tau_1, \ldots, \tau_M$.

Chapter 6

Simulation Study

In this chapter, the inference method of the previous chapter will be tested according to a simulation study. The focus will be on the average performance of the method on a set of relatively small datasets and a set of relatively large datasets. The performance on small datasets should give a feeling on how the algorithm performs in practice, while the performance on the larger datasets highlights more of the large sample performance of the algorithm. The method to simulate the datasets is explained in the next section.

6.1 Simulation Procedure

The differences between the two evaluated datasets lies in the number of groups, the expected number of patients per group, and the baseline function for each scenario (each sampled instance). In the smaller datasets, similar to the first simulation study of X. Lin and Wang (2010), the baseline function α is fixed per scenario and taken to be:

$$\alpha(t) = -3 + 2t \qquad \forall t \in \mathbb{R}.$$

In the case of the larger datasets, the baseline function is not fixed per scenario. Instead, the baseline is sampled as follows:

$$\begin{aligned} &d = 5, k = 6, \mathbf{t} = (0, 0, 0, 0, 0, 9.5, 19, 19, 19, 19, 19), \\ &\alpha(t) = \gamma_0 + \sum_{l=1}^{k} \gamma_l B_l^{d,\mathbf{t}}(t) \qquad \forall t \in \mathbb{R}, \\ &\gamma_0 \sim \mathcal{N}(0, 3),\ \gamma_i \sim \mathcal{N}(0, 1)|_{[0,\infty)} \qquad \forall i \in 1, \ldots, k. \end{aligned} \tag{6.1}$$

In Figure 6.1, one can see an example of a baseline function on the interval $[0, 19]$, sampled according to the procedure described above.

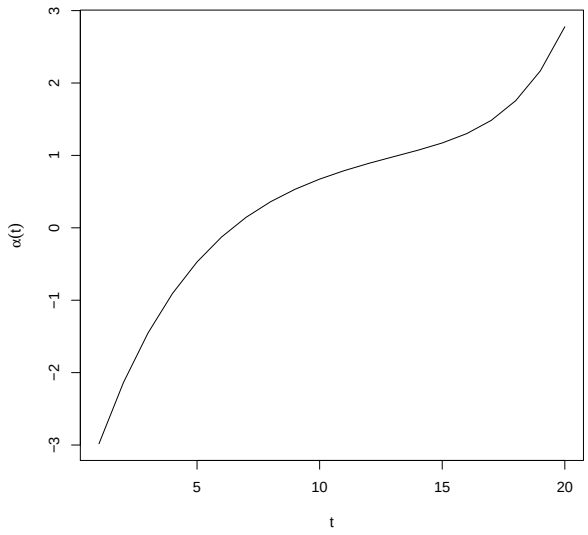


Figure 6.1: Example of a baseline function α sampled according to (6.1).

In the smaller datasets, the number of groups is set to 10, while in the larger datasets, the number of groups is set to 100. The expected number of patients per group in the smaller datasets is set to 50 while in the larger datasets, 100 patients are expected to be in a group. The number of patients for each individual group is independently sampled according to a Poisson distribution restricted to $\{0, \dots, m^*\}$ for some $m^* \in \mathbb{N}$. The two datasets, each containing 100 scenarios of clinical trials, are independently sampled for each value of τ in the set $\{-1/(2m^*), 0, 0.1, \dots, 1\}$.

For both datasets, the number of covariates per patient is 10 ($p = 10$), and the regression coefficients are sampled independently according to a $\mathcal{U}(-1, 1)$-distribution. The possible values of the left and right endpoints L_{ij} and R_{ij} are taken to be $\mathcal{T} = \{-\infty, 0, 1, 2, \dots, 19, \infty\}$. Hence, every patient has an event time between two values in this set.

In order to get a more or less balanced distribution of patients over the endpoints, each patient is assigned one finite value $\kappa_{ij}^{(1)}$ in $\mathcal{T}$. Next, $\kappa_{ij}^{(2)}$ is defined as the value in $\mathcal{T}$ consecutive to $\kappa_{ij}^{(1)}$. In the case that $\kappa_{ij}^{(1)}$ is the maximum finite value in $\mathcal{T}$, the smallest possible value for $\kappa_{ij}^{(2)}$ is ∞. This leads to a problem in the simulation procedure, and to circumvent this, $\kappa_{ij}^{(2)}$ is set to $\kappa_{ij}^{(1)} + 2$. The assignment of patients to intervals is done in such a way that every such interval is assigned roughly the same amount of patients.

One thing of note is that this is an idealized situation. In practice, there will probably be more observations of events at the earlier time intervals than in the later intervals. Hence, providing a good estimate for γ at the later time points will probably be more difficult for real-life datasets.
After assigning the intervals, the regression means μ_{ij}, which will eventually correspond to $X_{ij}\beta$ for patient i in group j are sampled uniformly inside the interval $\left[\alpha\left(\kappa_{ij}^{(1)}\right), \alpha\left(\kappa_{ij}^{(2)}\right)\right)$. Next, the covariates X_{ij} are sampled as follows. The first $p-1$ covariates are sampled uniformly in the interval $[-10, 10]$, and the last covariate $[X_{ij}]_p$ is set equal to

$$[X_{ij}]_p := \frac{\mu_{ij} - \sum_{k=1}^{p-1}[X_{ij}]_k\beta_k}{\beta_p}, \qquad \text{enforcing } \mu_{ij} = X_{ij}\beta.$$

After this, the latent vectors Z_j are sampled as:

$$Z_j \sim \mathcal{N}_{m_j}(X_j\beta, I_{m_j} + \tau J_{m_j}).$$

Now, the left and right endpoints of the observed interval are determined as the tuple (L_{ij}, R_{ij}) such that R_{ij} is the smallest value in $\mathcal{T}$ higher than L_{ij} and $Z_{ij} \in [\alpha(L_{ij}), \alpha(R_{ij}))$. The last step is now to include some additional right and left censoring. Of the patients for which L_{ij} is finite, 5% is drawn at random and R_{ij} is set to ∞. Similarly, 5% of the patients for which the right endpoint is finite is selected for left censoring (L_{ij} is set to $-\infty$). The method to simulate the datasets is summarized in Algorithm 3.

A last thing of note is the chosen priors for the parameters. For γ_0 a $\mathcal{N}(0, 10^{10})$ prior was chosen, for γ_i where $i > 0$, a $\mathcal{N}(0, 10^{10})_{[0,\infty)}$ prior was chosen. For $\tau + 1/m^*$, a IG$(10^{-10}, 10^{-10})-$prior was chosen. For β, the choice was made to set $\mu_0 = 0$ and $\Sigma_0^{-1} = \mathcal{O}_p$ (the zero matrix) . This corresponds to taking the improper prior $p_\beta \propto 1$ for β.

Thus, where possible, a non-informative prior was chosen, and if this wasn't possible, a nearly non-informative prior was chosen. Hence, the objective Bayesian approach is taken. After performing a few test simulations, the choice was made to set the number of Markov chain iterations for the smaller set to 5000. Similarly, for the larger datasets, the number of Markov chain iterations was set to 3000. The burn-in period for both these datasets was set to 1000. The simulation study took about a week to complete on a standard laptop using the software package `R`. One last thing of note is that no `R` packages for MCMC (such as `RStan`) were used in this simulation study.

Algorithm 3 Method to simulate a set of clinical trials with data according to Model 2.

Inputs:
Number of groups n, the number of covariates p, the spline degree d and knots $\mathbf{t}$, the set of possible endpoints $\mathcal{T}$, the censoring frequency q and the expected number of patients per group, denoted λ. Additionally, the baseline function α can also be given as input.
Initialize:
Define $B_1^{d,\mathbf{t}}, \ldots, B_k^{d,\mathbf{t}}$ to be the I-splines of order d with knots $\mathbf{t}$.
Sample group sizes $m_i \overset{iid}{\sim}$ Pois(λ);
Sample $\beta_i \overset{iid}{\sim} \mathcal{U}(-1, 1)$;
if α is undefined **then**
    Sample $\gamma_0 \sim \mathcal{N}(0, 3)$ and $\gamma_i \sim \mathcal{N}(0, 1)|_{[0,\infty)}$ for $i \in \{1, \ldots, k\}$;
    Define $\alpha := \gamma_0 + \sum_{i=1}^{k} B_i^{d,\mathbf{t}} \gamma_i$
end if
for $j \in \{1, \ldots, n\}$ **do**
    for $i \in \{1, \ldots, m_j\}$ **do**
        Set $\kappa_{ij}^{(1)} := T_{l_{ij}}$, where $l_{ij} = \mod\left(\sum_{l=1}^{j-1} m_l + i,\ |\mathcal{T}| - 2\right) + 1$
        and T_i is the i-th order statistic in $\mathcal{T}$;
        Set $\kappa_{ij}^{(2)} := T_{k_{ij}+1}$;
        if $\kappa_{ij}^{(2)} = \infty$ **then**
            Set $\kappa_{ij}^{(2)} := \kappa_{ij}^{(1)} + 2$;
        end if
        Sample $\mu_{ij} \sim \mathcal{U}\left(\alpha\left(\kappa_{ij}^{(1)}\right), \alpha\left(\kappa_{ij}^{(2)}\right)\right)$;
        Sample $(X_{ij})_1, \ldots, (X_{ij})_{p-1} \overset{iid}{\sim} \mathcal{U}(-10, 10)$;
        Set $(X_{ij})_p := \frac{\mu_{ij} - \sum_{l=1}^{p-1} (X_{ij})_l \beta_l}{\beta_p}$;
    end for
    Sample $Z_j \sim \mathcal{N}_{m_j}(X_j\beta,\ I_{m_j} + \tau J_{m_j})$;
    for $i \in \{1, \ldots, m_j\}$ **do**
        Set $L_{ij} := \max\{t \in \mathcal{T} :\ \alpha(t) \leq Z_{ij}\}$;
        Set $R_{ij} := \min\{t \in \mathcal{T} :\ \alpha(t) > Z_{ij}\}$;
    end for
end for
Set $\mathcal{I}_L = \{(i, j) :\ j \in \{1, \ldots, n\},\ i \in \{1, \ldots, m_j\},\ L_{ij} > -\infty\}$;
Choose $\lceil q \cdot |\mathcal{I}_L| \rceil$ elements (i, j) of $\mathcal{I}_L$ and set the corresponding R_{ij} to ∞;
Set $\mathcal{I}_R = \{(i, j) :\ j \in \{1, \ldots, n\},\ i \in \{1, \ldots, m_j\},\ R_{ij} < \infty\}$;
Choose $\lceil q \cdot |\mathcal{I}_R| \rceil$ elements (i, j) of $\mathcal{I}_R$ and set the corresponding L_{ij} to $-\infty$;
Outputs:
    $L, R, X, m_1, \ldots, m_n$

6.2 Parameter Recovery

In Table 6.1, one can find *parameter recovery* results for the inference method on the smaller dataset (10 groups, 50 patients on average). The estimator for each parameter is the posterior median. A distinction in the performance evaluation is made between the estimator for the baseline translation coefficient γ_0 and the estimators for the other baseline coefficients $\gamma_{(-0)}$. For the regression and baseline coefficient estimators, the relative bias (RB) and relative precision (RP) measures, defined as the bias and standard deviation divided by the absolute value of the true parameter, are calculated in percentage points for each value of τ. The median of these performance measures over all simulated scenarios was then calculated. The median was taken because with only 100 simulated scenario's per τ value, it was seen that the outliers in performance effected the estimated sample mean too much. The median over medians and median over standard deviations of the posterior samples of τ are also displayed in the table. Note that the median posterior standard deviation for τ is in absolute percentage points, not relative percentage points.

No significant relation can be seen between the value of τ and the relative bias/precision of the posterior medians for β and γ. Almost no bias is seen for the regression parameters, while the baseline coefficients show more bias. The median relative precisions for the regression coefficients are also smaller than those for the spline coefficients. This could be due to the fact that the spline functions act more locally than the regression parameters. The information for a spline coefficient is only provided by patients having the event-time in a certain time interval.

The posterior bias and variance of τ are seen to increase with τ. This relationship also holds in other models. Consider for instance inference on the variance in iid normal data. The posterior variance has an inverse gamma distribution. For this distribution, the variance is proportional to the mean.

As the median posterior standard deviation in τ is in many cases almost equal to the true value of τ in the table below, it cannot be said that the posterior density of τ is very informative about the true value of τ. This is mainly due to the fact that effectively, as there are 10 groups, there are only 10 observations of τ for each scenario.

In Table 6.2, one can find parameter recovery results for the inference method on the larger dataset (100 groups, on average 100 patients). It is seen that the relative bias and precision for the regression coefficients has decreased. As the baseline in the larger dataset is not taken to be fixed, it is not fair to compare differences in the relative bias

Table 6.1: Parameter recovery results for $n = 10$, expected group size $\lambda = 50$, 100 simulations and 5000 Markov chain iterations with a burn-in period of 1000.

	$\hat{\beta}$		$\hat{\gamma}_0$		$\hat{\gamma}_{(-0)}$		$\hat{\tau}$	
τ	RB (%)	RP (%)	RB (%)	RP (%)	RB (%)	RP (%)	Median	SD (%)
-0.0071	-0.12	6.35	3.57	13.55	-3.73	9.21	-0.0069	1.60
0	0.23	6.36	4.47	13.96	-5.52	8.76	0.0094	2.83
0.1	0.24	5.95	1.34	13.72	-0.49	8.87	0.09	7.55
0.2	-0.06	5.91	2.46	13.57	-5.25	9.96	0.22	17.34
0.3	0.10	5.78	1.54	13.75	-1.37	10.33	0.27	20.87
0.4	-0.51	6.39	2.51	13.75	-3.66	10.47	0.37	27.46
0.5	0.11	6.04	3.55	13.36	-3.51	11.18	0.52	39.33
0.6	0.26	6.03	1.74	13.50	-2.48	12.26	0.65	47.27
0.7	-0.05	5.91	3.19	13.61	-2.33	11.88	0.74	55.16
0.8	0.39	5.70	3.33	13.66	-1.09	12.62	0.85	62.81
0.9	0.00	6.05	2.01	12.77	-2.54	12.91	0.94	65.04
1	0.22	6.88	2.46	13.98	2.00	14.19	1.11	78.76

and precision of the spline coefficients to the same measures for the small dataset. It is seen that the performance measures for the spline coefficients are approximately the same. It is seen that the medians for τ lie closer to the true value as compared to in the smaller dataset, and the posterior standard deviation in τ is now also lower. The median is very close (exact up to two decimals) to the true value of τ for $\tau < 0.5$, after which some bias can still be seen. It can now be said that the posterior density of τ is quite informative for the true value of τ.

When compared to the regression coefficients, the spline coefficients showed more posterior variation. To get a better view of the effect of this larger variation, the estimated baseline is compared to the true baseline in the next part. In Figure 6.2, the true baseline is shown for the smaller dataset, as well as the mean of all estimated (median) baselines and the 5% and 95% pointwise quantiles over all estimated baselines. Note that in this set of estimated baselines no distinction between values of τ is made.

Table 6.2: Parameter recovery results for $n = 100$, expected group size $\lambda = 100$, 100 simulations and 3000 Markov chain iterations with a burn-in period of 1000.

	$\hat{\beta}$		$\hat{\gamma}_0$		$\hat{\gamma}_{(-0)}$		$\hat{\tau}$	
τ	RB (%)	RP (%)	RB (%)	RP (%)	RB (%)	RP (%)	Median	SD (%)
-0.0036	0.00	1.51	0.63	8.33	1.43	1.81	−0.0035	0.12
0	−0.07	1.55	0.04	18.28	2.37	4.33	0.0011	0.20
0.1	0.16	1.58	−0.05	15.95	2.75	4.63	0.10	1.68
0.2	−0.02	1.56	−2.09	16.20	2.79	4.59	0.20	3.16
0.3	−0.02	1.63	−2.09	14.84	3.03	7.62	0.30	4.85
0.4	0.07	1.70	−1.32	15.60	4.81	7.04	0.40	6.12
0.5	0.01	1.62	0.22	16.50	2.95	8.18	0.53	8.07
0.6	0.03	1.55	−1.47	15.03	1.18	5.40	0.61	9.20
0.7	−0.05	1.55	0.19	17.23	6.22	9.92	0.76	12.56
0.8	−0.06	1.56	−2.63	15.19	2.11	10.36	0.83	13.06
0.9	−0.05	1.40	−2.03	15.55	2.63	8.92	0.93	14.47
1	−0.05	1.56	−0.94	16.45	2.74	10.29	1.01	15.99

It is seen that the mean baseline is quite close to the true baseline, and that the radius of the confidence interval increases with the event-time. This mainly has to do with the fact that the baseline is modeled as a linear combination of I-spline functions, and these functions are all nonzero at the larger event-times. Hence the uncertainty in the spline coefficients has most of its effect on the uncertainty in the baseline at the larger event-times. The maximum radius of the 90% confidence interval lies around 10%. It can be concluded from this that the baseline can be determined with quite a lot of certainty from the scenarios in the small dataset.

In Figure 6.3, the true baseline is plotted against the mean, 95% and 5% baseline quantiles for one scenario in the large dataset. With this larger sample size, the variance is not seen to noticeably increase anymore with t this time. The baselines sampled from the posterior distribution lie very close to the true value in this scenario. It is however seen that the true baseline slightly misses the 90% confidence interval at lower/larger event times. This could be due to the fact that the quantiles are determined pointwise instead of e.g. over the whole vector γ (which should increase the radius).

Another way to evaluate the performance of the Metropolis-Hastings algorithm is to check whether estimated *credible intervals* for parameters really give the desired *coverage* for that parameter. The credible interval of probability p is defined as the region between the $(1-p)/2$-th and the $(1+p)/2$-th quantile of the posterior distribution for that parameter (equal tails). It is hence an interval in the support of the random variable that contains a posterior probability of p. The *coverage* of a credible interval is the expected frequency of times that the true parameter lies in that credible interval. The choice is made to evaluate coverage of the credible intervals for τ. For each scenario in the small and large dataset, credible intervals of probability $p \in \{0.6, 0.65, \ldots, 0.9, 0.95, 0.99\}$ are estimated by taking the empirical $(1-p)/2$-th and $(1+p)/2$-th quantiles of the posterior samples. For each p and τ, the frequency of scenario's where the value of τ was in the credible interval was calculated. In Figure 6.4 and 6.5, for both the small and large dataset respectively, the average, minimum and maximum coverage over all values of τ are plotted vs. the credibility value p.

It is seen that the average coverage increases with the credibility value, and often lies close to the theoretical coverage, as expected. It can be seen that there is quite a lot of spread in the coverage values over the values of τ. This might have to do with the small (100) number of scenarios per dataset. The two figures show approximately the same behavior. This is expected, the credible intervals for the larger dataset should be tighter, however the performance of these intervals should be approximately the same as the (wider) credible intervals for the smaller dataset.

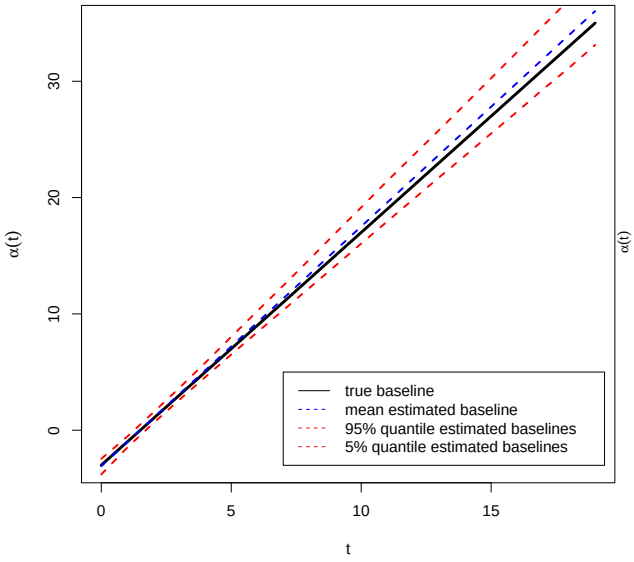


Figure 6.2: True baseline vs. the average estimated baseline as well as the 5% and 95% pointwise quantiles over all small sample scenario's.

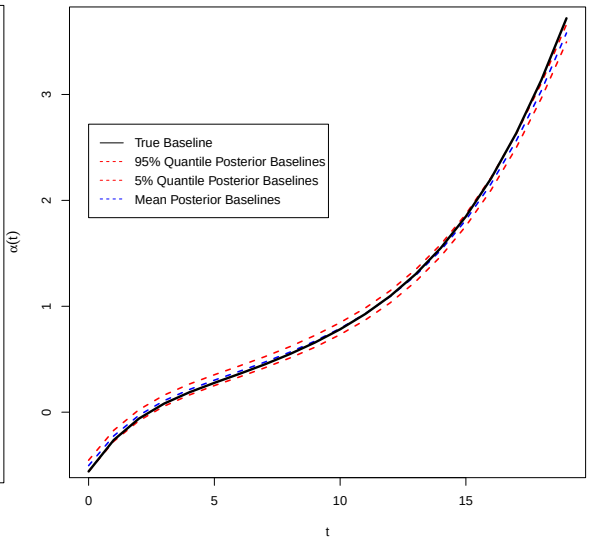


Figure 6.3: True baseline vs. the average estimated baseline as well as the 5% and 95% pointwise quantiles for one scenario in the large dataset.

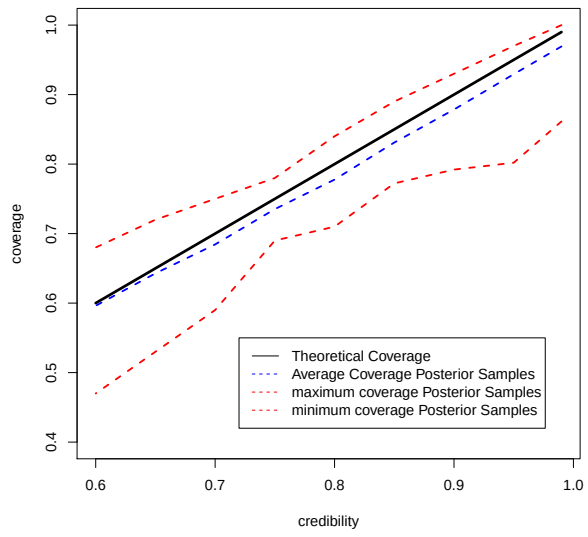


Figure 6.4: Estimated vs. Theoretical coverage of credible intervals for τ in the smaller dataset. The mean coverage over all values of τ is shown, as well as the pointwise minimum and maximum over all values of τ.

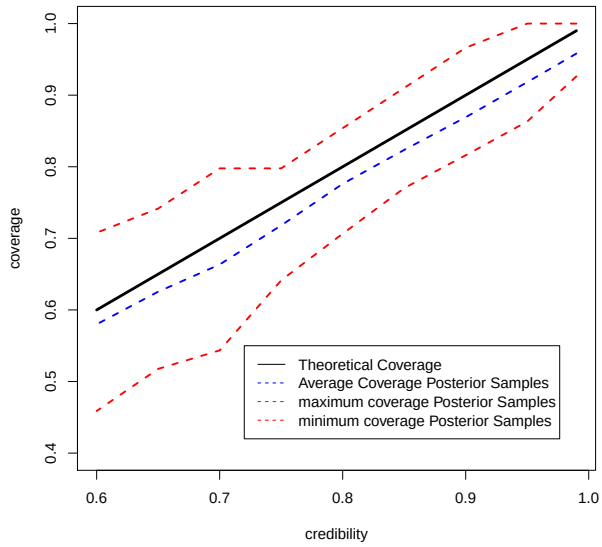


Figure 6.5: Estimated vs. Theoretical coverage of credible intervals for τ in the larger dataset. The mean coverage over all values of τ is shown, as well as the pointwise minimum and maximum over all values of τ.

6.3 Bayes Factor Evaluation

In this section, the Laplace-Metropolis approximated Bayes factors are calculated and evaluated for the scenario's in the small and large dataset. Remember from Section 3.3 that the Bayes factor is approximated with:

$$\frac{d_0 - d_1}{2}\log(2\pi) + \frac{1}{2}\left(|\mathbf{H}_0^*| - |\mathbf{H}_1^*|\right) + \left(l_0(D, \mathbf{u}_0^*) - l_1(D, \mathbf{u}_1^*)\right) - \left(\lambda_0(\mathbf{u}_0^*) - \lambda_1(\mathbf{u}_1^*)\right).$$

In this case, $d_0 - d_1 = -1$ as model H_1 also contains parameter τ. Furthermore, $u = (\beta, \gamma, \tau)$. Now, let $\vec{L}, \vec{R}$ be as in Section 5.1, and $D = (X, L, R)$. When looking at the priors described on page 55, it holds that:

$$\begin{aligned}
l_0(X, L, R, \beta, \gamma, \tau) &= \log\left(\prod_{j=1}^{n}\prod_{i=1}^{m_j}\left(\Phi(\vec{R}_{ij}\gamma - X_{ij}\beta) - \Phi(\vec{L}_{ij}\gamma - X_{ij}\beta)\right)\right), \\
l_1(X, L, R, \beta, \gamma, \tau) &= \log\left(\prod_{j=1}^{n}\Phi_{m_j}\left(\prod_{i=1}^{m_j}[\vec{L}_{ij}\gamma, \vec{R}_{ij}\gamma);\ X_j\beta;\ I_{m_j} + \tau J_{m_j}\right)\right), \\
\lambda_0(\beta, \gamma) &= \log\left(\phi\left(\frac{\gamma_0}{10^5}\right)\prod_{l=1}^{k} 2\cdot\phi\left(\frac{\gamma_i}{10^5}\right)\mathbb{1}_{[0,\infty)}(\gamma_i)\right) + C, \\
\lambda_1(\beta, \gamma, \tau) &= \log\left(p_\tau(\tau)\cdot\phi\left(\frac{\gamma_0}{10^5}\right)\prod_{l=1}^{k} 2\phi\left(\frac{\gamma_i}{10^5}\right)\mathbb{1}_{[0,\infty)}(\gamma_i)\right) + C.
\end{aligned}$$

In the above, C is the term induced by the constant improper prior on β, notice that in the Bayes factor approximation, this constant vanishes assuming that the same prior is taken in both models. Furthermore, p_τ is the shifted inverse gamma density with scale and shape parameters equal to 10^{-10}.

In Figure 6.6 and 6.7, the pointwise average over the scenarios, as well as the 99.75% and 0.25% quantiles of logarithm of the Laplace-Metropolis approximated Bayes factor are plotted vs. the true value of τ for the smaller and larger dataset respectively. When looking at the results, a negative relationship can be seen between the average logarithm and τ. The negative relation is as expected, with a larger value of τ one expects a larger evidence in favor of the hypothesis $H_1 : \tau \neq 0$ as compared to $H_0 : \tau = 0$, resulting in a lower Bayes factor.

For the smaller dataset it is seen that for $\tau \approx -0.007$ and $\tau \in \{0, 0.1\}$, the average log-Bayes factor is positive and for $\tau \geq 0.2$ the average log-Bayes factor is negative. When looking at the quantiles, it is seen that the log Bayes factor is positive for at least 95% of the scenarios when $\tau \leq 0$. When $\tau > 0$, negative values also occur, and it is

seen that the Bayes factor has a larger negative range than a positive range. In other words, when $\tau > 0$, the negative values that the log Bayes factor attains are higher in absolute value than the positive ones. Again, by looking at the quantiles, it is seen that the log Bayes factor is negative for at least 95% of the scenarios only when $\tau \geq 0.9$.

For the larger dataset, it is seen that the average value of the log Bayes factor becomes negative when $\tau > 0$. By checking the quantiles, it is seen that when $\tau \geq 0.2$, the log Bayes factor is negative for at least 95% of the scenarios. For the larger dataset, the Bayes factors are a lot more concentrated around the average and take on lower values than the Bayes factors for the smaller dataset. This can be expected as the sample sizes are larger in the larger dataset, hence there is a lot more evidence to be found for/against the models.

In the Bayes factor evaluation, it was seen that the Bayes factor cannot detect properly if $\tau < 0$. This is probably due to the fact that quite a large average group size is taken for both the small and the large dataset (50 and 100 respectively), hence the negative τ values considered were quite close to 0. In retrospect, it would have been better to take smaller average group sizes. In future research, a simulation study with smaller average group sizes would hence be necessary to also show more clearly the Bayes factor performance at negative values of τ.

To conclude this chapter, it the simulation results were as expected. Inference on the model yielded reliable information about the true parameter values. With larger values of τ, the Bayes factor is seen to reject the null hypothesis more frequently. Furthermore, it was seen that with larger samples, there was more confidence about the parameters of the model. The conclusion from this is that inference under the model in Definition 2 can be efficiently and reliably performed by the method described in Chapter 5. In the future, a dataset with small expected group sizes is however still needed to assess the Bayes factor performance for negative values of τ.

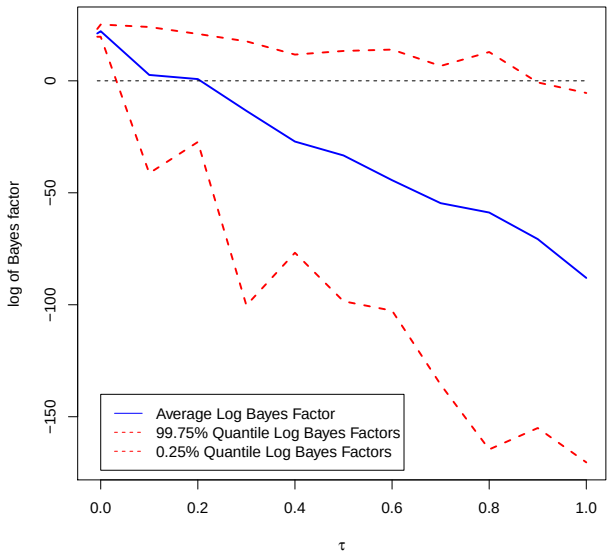


Figure 6.6: Average, 99.75% and 0.25% quantiles of the logarithm of Laplace-Metropolis estimated Bayes factors for $H_0 : \tau = 0$ against $H_1 : \tau \neq 0$. The number of groups (n) was 10 and the expected group size was 50.

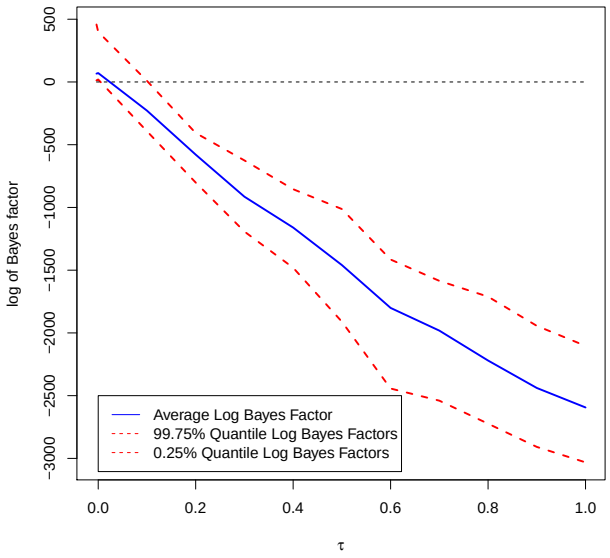


Figure 6.7: Average, 99.75% and 0.25% quantiles of the logarithm of Laplace-Metropolis estimated Bayes factors for $H_0 : \tau = 0$ against $H_1 : \tau \neq 0$. The number of groups (n) was 100 and the expected group size was 100.

Chapter 7

Conclusion and Discussion

7.1 Conclusion

In this research, alternative designs for clinical trials were investigated. After a literature research, the choice was made to investigate a multivariate survival model, where the event-times are interval censored. This model has links to the recently introduced meta-analytic approach to modeling multiple clinical trials.

It was seen that inference on the survival model can be done efficiently and reliably. The method for sampling the spline coefficients γ in this model uses a formula for a multivariate normal integral. This formula yielded an efficient method of evaluating a multivariate probability in very high dimensions, under the restriction of equal correlation. As such a probability in general cannot be evaluated efficiently, the method could also have a lot of other uses.

Lastly, the multivariate survival model also lends itself to covariance testing. If a medicine is effective, one would expect that the outcome variables for patients to either increase or decrease overall. This effect induces correlation in the outcomes of the patients, which can only be accounted for by the treatment effect. If groups of patients in the group have significantly different responses to the treatment, the additional correlation can be tested for. Detecting this correlation hence corresponds to detecting that the treatment has had an effect on the outcome variables. This testing procedure can be performed without the need for a control group. Hence, if no current medicine is available, every patient in the group now receives treatment. Only dealing with one treatment group also has monetary advantages, a lot of planning and bookkeeping in clinical trials can be avoided.

Covariance testing can also have other uses. It can also be used to determine

whether personalized medicine might be an option. If different versions of the treatment are given to different groups of patients, detecting covariance indicates that these groups had a significantly different reaction to the different treatments and the best option might be selected.

7.2 Discussion

In this section, ideas for further research are considered. First off, throughout this research, it is stated that if there are groups of patients in a trial, and if correlation is found for outcomes of patients in these groups, a treatment effect is detected. An interesting possible next step for this research would be to include uncertainty about these groups in the survival model (i.e. to find these groups). This could be quite difficult, one has to define a distribution on clusters of patient outcomes having a latent jointly multivariate Gaussian distribution. In sequential analysis, the number of patients is unbounded (in theory) and hence, an infinite amount of clusters could possibly occur. Reversible jump Markov Chain Monte Carlo techniques, or techniques in nonparametric Bayesian statistics might be an option to deal with this problem.

From Figure 6.2, it is seen that the uncertainty in the baseline increases at the larger time points. This is due to the fact that α is chosen as a linear combination of increasing functions, which are all nonzero at the later time points. In practice, it will already be difficult to determine the baseline at these points due to the decreasing number of event observations at later times. Hence, this additional uncertainty at later time points is unwanted. It could be that the definition of α can be altered such that the uncertainty decreases with time.

In this research, a method was considered to test for a treatment effect without a control group. This ”loss of control” could result in a higher probability of a type 1 error / lower power of the Bayes Factor test when compared to an RCT situation. It could hence be useful to compare the performance of a test conducted in an RCT with the Bayes factor test on the same benchmarks.

Another idea for further research could be to include multiple significantly different treatments in the trial. The model would then consider the treatment effects for these variables as drawn from different distributions, hence more elaborate covariance structures are considered. Furthermore, more outcome variables, possibly representing recurrent or dependent events could also be included in the model.

References

Amberson Jr, J. B. (1931). A clinical trial of sanocrysin in pulmonary tuberculosis. *American Review of Tuberculosis*, *24*, 401–435.

Armitage, P., Berry, G., & Matthews, J. N. S. (1971). *Statistical methods in medical research* (Vol. 449). Wiley Online Library.

Assaf, G. A., & Tsionas, M. (2018). Bayes factors vs. p-values. *Tourism Management*, *67*, 17-31.

Box, G. E., Luceno, A., & del Carmen Paniagua-Quinones, M. (2011). *Statistical control by monitoring and adjustment* (Vol. 700). John Wiley & Sons.

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, *210*, 2–21.

Eaton, M. L. (1983). Multivariate statistics: a vector space approach. *John Wiley & Sons, Inc., 605 Third Ave., New York, NY 10158, USA, 1983, 512*.

Fisher, R. A. (1936). Design of experiments. *Br Med J*, *1*(3923), 554–554.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. Springer Science & Business Media.

Fox, J.-P., Mulder, J., & Sinharay, S. (2017). Bayes factor covariance testing in item response models. *psychometrika*, *82*(4), 979–1006.

Friedman, L. M., Furberg, C., DeMets, D. L., Reboussin, D. M., Granger, C. B., et al. (2010). *Fundamentals of clinical trials* (Vol. 4). Springer.

Gassmann, H. I., Deák, I., & Szántai, T. (2002). Computing multivariate normal probabilities: A new look. *Journal of Computational and Graphical Statistics*, *11*(4), 920–949.

Gelman, A., Rubin, D. B., et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, *7*(4), 457–472.

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Grünwald, P. D. (2016). Toetsen als gokken: een redelijk alternatief voor de p-waarde. *Nieuw Archief voor Wiskunde*, *17*(4), 236–244.

Hendriks, H. (2018). Test martingales for bounded random variables. *arXiv preprint arXiv:1801.09418*.

Hill, A. B. (1990). Memories of the british streptomycin trial in tuberculosis: the first randomized clinical trial. *Controlled Clinical Trials*, *11*(2), 77–79.

Hobbs, B. P., Carlin, B. P., Mandrekar, S. J., & Sargent, D. J. (2011). Hierarchical commensurate and power prior models for adaptive incorporation of historical in-

formation in clinical trials. *Biometrics*, *67*(3), 1047–1056.

Hollander, M., Wolfe, D. A., & Chicken, E. (2013). *Nonparametric statistical methods* (Vol. 751). John Wiley & Sons.

Ibrahim, J. G., Chen, M.-H., Gwon, Y., & Chen, F. (2015). The power prior: theory and applications. *Statistics in medicine*, *34*(28), 3724–3749.

Ibrahim, J. G., Chen, M.-H., et al. (2000). Power prior distributions for regression models. *Statistical Science*, *15*(1), 46–60.

Kelly, F. P. (2011). *Reversibility and stochastic networks*. Cambridge University Press.

Korosteleva, O. (2009). *Clinical statistics: introducing clinical trials, survival analysis, and longitudinal data analysis*. Jones & Bartlett Publishers.

Lai, T. L. (2001). Sequential analysis: some classical problems and new challenges. *Statistica Sinica*, 303–351.

LaMorte, W. W. (2019). *Mann whitney u test (wilcoxon rank sum test).* http://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_nonparametric/BS704_Nonparametric4.html.

Lewis, S. M., & Raftery, A. E. (1997). Estimating bayes factors via posterior simulation with the laplace—metropolis estimator. *Journal of the American Statistical Association*, *92*(438), 648–655.

Lim, J., Walley, R., Yuan, J., Liu, J., Dabral, A., Best, N., ... others (2018). Minimizing patient burden through the use of historical subject-level data in innovative confirmatory clinical trials: review of methods and opportunities. *Therapeutic innovation & regulatory science*, *52*(5), 546–559.

Lin, M., Lucas Jr, H. C., & Shmueli, G. (2013). Research commentary—too big to fail: large samples and the p-value problem. *Information Systems Research*, *24*(4), 906–917.

Lin, X., & Wang, L. (2010). A semiparametric probit model for case 2 interval-censored failure time data. *Statistics in medicine*, *29*(9), 972–981.

Lind, J. (1757). *A treatise on the scurvy: in three parts, containing an inquiry into the nature, causes, and cure, of that disease*. A. Millar.

Louis, P. C. A. (1835). *Recherches sur les effets de la saignée dans quelques maladies inflammatoires...* J.-B. Baillière.

Ly, A., Etz, A., Marsman, M., & Wagenmakers, E.-J. (2018). Replication bayes factors from evidence updating. *Behavior research methods*, 1–11.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 50–60.

McHugh, M. L. (2013). The chi-square test of independence. *Biochemia medica: Biochemia medica*, *23*(2), 143–149.

Mehta, C. R., & Senchaudhuri, P. (2003). Conditional versus unconditional exact tests for comparing two binomials. *Cytel Software Corporation*, *675*, 1–5.

Mukerjee, N. R. R. (2003). Some aspects of matching priors. *Mathematical statistics and applications: Festschrift for Constance van Eeden*, *42*, 31.

Neuenschwander, B., Branson, M., & Spiegelhalter, D. J. (2009). A note on the power prior. *Statistics in Medicine*, *28*(28), 3562–3566.

Neuenschwander, B., Capkun-Niggli, G., Branson, M., & Spiegelhalter, D. J. (2010). Summarizing historical information on controls in clinical trials. *Clinical Trials*, *7*(1), 5–18.

Pauler, D. K., Wakefield, J. C., & Kass, R. E. (1999). Bayes factors and approximations for variance component models. *Journal of the American Statistical Association*, *94*(448), 1242–1253.

Pettitt, A. (1984). Proportional odds models for survival data and estimates using ranks. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *33*(2), 169–175.

Pocock, S. J. (1976). The combination of randomized and historical controls in clinical trials. *Journal of chronic diseases*, *29*(3), 175–188.

Pocock, S. J. (2013). *Clinical trials: a practical approach*. John Wiley & Sons.

Pollard, D. (2002). *A user's guide to measure theoretic probability* (Vol. 8). Cambridge University Press.

Raymond, M., & Rousset, F. (1995). An exact test for population differentiation. *Evolution*, *49*(6), 1280–1283.

Roberts, G. O., & Smith, A. F. (1994). Simple conditions for the convergence of the gibbs sampler and metropolis-hastings algorithms. *Stochastic processes and their applications*, *49*(2), 207–216.

Rousseeuw, P. J., & Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical association*, *85*(411), 633–639.

Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to student's t-test and the mann–whitney u test. *Behavioral Ecology*, *17*(4), 688–690.

Schmidli, H., Gsteiger, S., Roychoudhury, S., O'Hagan, A., Spiegelhalter, D., & Neuenschwander, B. (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, *70*(4), 1023–1032.

Shafer, G., Shen, A., Vereshchagin, N., Vovk, V., et al. (2011). Test martingales, bayes factors and p-values. *Statistical Science*, *26*(1), 84–101.

Spiegelhalter, D. J., Abrams, K. R., & Myles, J. P. (2004). *Bayesian approaches to clinical trials and health-care evaluation* (Vol. 13). John Wiley & Sons.

Steck, G., & Owen, D. (1962). A note on the equicorrelated multivariate normal distribution. *Biometrika*, *49*(1/2), 269–271.

Van der Vaart, A. W. (2000). *Asymptotic statistics* (Vol. 3). Cambridge university press.

van Rosmalen, J., Dejardin, D., van Norden, Y., Löwenberg, B., & Lesaffre, E. (2018). Including historical data in the analysis of clinical trials: Is it worth the effort? *Statistical methods in medical research*, *27*(10), 3167–3182.

Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N., ... others (2014). Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical statistics*, *13*(1), 41–54.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems ofp values. *Psychonomic bulletin & review*, *14*(5), 779–804.

Wahba, G. (1990). *Spline models for observational data* (Vol. 59). Siam.

Wald, A. (1945). Sequential tests of statistical hypotheses. *The annals of mathematical statistics*, *16*(2), 117–186.

Walker, J., & Almond, P. (2010). *Interpreting statistical findings: a guide for health professionals and students*. McGraw-Hill Education (UK).

Wasserstein, R. L., Lazar, N. A., et al. (2016). The asa's statement on p-values: context, process, and purpose. *The American Statistician*, *70*(2), 129–133.

Windish, D. M., Huot, S. J., & Green, M. L. (2007). Medicine residents' understanding of the biostatistics and results in the medical literature. *Jama*, *298*(9), 1010–1022.

Wu, H., & Wang, L. (2019). Normal frailty probit model for clustered interval-censored failure time data. *Biometrical Journal*.

Appendix A

List of Symbols and Their Description

Symbol	Description
$\mathbb{I}(F)$	The indicator that statement/event F is true.
$\mathbb{1}_E(x)$	The indicator that x lies in the set E (i.e. $\mathbb{1}_E(x) = \mathbb{I}(x \in E)$).
p_Y	The prior on the latent variable(s) Y.
$f_{Y\|X}$	The conditional density of the latent variable(s) Y given X.
$\mathbf{1}_m$	The vector of length m (for $m \in \mathbb{N}$) where every element is equal to 1.
$\overset{iid}{\sim}$	Independent and identically distributed according to the distribution on the right.
Φ	The standard normal cumulative distribution function.
$\Phi_k(E;\ \mu;\ \Sigma)$	The probability that a multivariate normal vector with mean μ and covariance Σ lies in a set E.
ϕ	The standard normal density function.
$\wedge, \vee$	The binary minimum and maximum operators respectively, where conventionally logical statements are taken to be either 1 or 0.
Hess$()$	The Hessian operator. When acting on a d-dimensional function, it returns a function $H : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ such that $[H(x)]_{ij} = \left(\frac{\partial^2}{\partial u_i \partial u_j} h\right)(x) \quad \forall x \in \mathbb{R}^d$.
IG(α, β)	The class of inverse gamma distributed random variables with shape and scale parameter α and β respectively.
$\mathcal{U}(a, b)$	The class of uniformly distributed random variables on the interval with endpoints a, b with $a < b$ both in $\mathbb{R}$.
$\mathcal{N}_k(\mu, \Sigma)$	The class of k-dimensional multivariate normally distributed vectors with mean μ and variance Σ. When $k = 1$, the subscript is omitted.
J_m	The $m \times m$ matrix where every element is equal to 1.
I_m	The $m \times m$ identity matrix.
$B_l^{d,\mathbf{t}}$	The l-th integrated spline function with degree d and knots $\mathbf{t}$.
e_i^j	The i-th unit vector in $\mathbb{R}^{m_j}$, where m_j is the j-th group size.
$C^q(S)$	The set of q-times continuously differentiable functions on the set S.

Appendix B

Conditional Marginal Distributions for a Truncated Multivariate Normal Vector

In this appendix, the result stated on Page 44 is shown.

Theorem 4. *Let* $X \sim \mathcal{N}_m(\mu, \Sigma)$ *with* $X = [X_1', X_2']'$, $\mu = [\mu_1', \mu_2']'$ *and* $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}' & \Sigma_{22} \end{pmatrix}$ *such that* $X_1 \sim \mathcal{N}_{m_1}(\mu_1, \Sigma_{11})$ *and* $X_2 \sim \mathcal{N}_{m_2}(\mu_2, \Sigma_{22})$ *with* $m_1 + m_2 = m$.
Then:

$$X_1 | X_2 \sim \mathcal{N}_{m_1}(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2),\ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}')$$

Proof. See Eaton (1983) pg. 116-117. □

Definition 4. A random vector $X \in \mathbb{R}^m$ has the *Truncated multivariate normal distribution* if there exist three vectors $\mu, L, R \in \mathbb{R}^m$ and a matrix $A \in \mathbb{R}^{m \times n}$ for some $n \in \mathbb{N}$ such that:

$$X \sim \mu + AY$$

where $Y \in \mathbb{R}^n$ is a random vector with density:

$$f_Y(y|\mu, A, L, R) = C \prod_{i=1}^{n} \phi(y_i) \mathbb{1}_{(L_i, R_i)}((Ay + \mu)_i).$$

In the above equation C is the normalizing constant. Defining $\Sigma = AA'$, the shorthand notation used in this research is $X \sim \mathcal{N}_m(\mu, \Sigma)|_{(L,R)}$

Corollary 4.1. Let $Z \sim \mathcal{N}_m(\mu, aI_m + bJ_m)|_{(L,R)}$ where $a \in (0, \infty)$, $\mu \in \mathbb{R}^m$ and $b \in$

$(-\frac{\alpha}{m}, \infty)$. Then, denoting with $Z^{(-i)}$ the vector Z without its i-th element, it holds that

$$Z_i|Z^{(-i)} \sim \mathcal{N}(\mu, \sigma^2)|_{(L_i,R_i)},$$

where

$$\mu = \mu_i + \frac{b}{(m-1)b+a}\sum_{j\neq i}(Z_j - \mu_j),$$

$$\sigma^2 = \frac{a(bm+a)}{(m-1)b+a}.$$

Proof. The density p_Z of Z is zero when the density $f_{Z^{(-i)}}$of $Z^{(-i)}$ is zero, hence the pdf of $Z_i|Z^{(-i)}$ is given by:

$$f_{Z_i|Z^{(-i)}} = \begin{cases} \frac{f_Z}{f_{Z^{(-i)}}} & \text{when } f_{Z^{(-i)}} > 0, \\ 0 & \text{else.} \end{cases}$$

As $f_{Z^{(-i)}} = 0$ iff $Z^{(-i)}$ is not in the truncation intervals, the truncation intervals for $Z^{(-i)}$ vanish in the above ratio. Hence, taking $X \sim \mathcal{N}_m(\mu, aI_m + bJ_m)$, it follows that

$$f_{Z_i|Z^{(-i)}} = \frac{f_X}{f_{X^{(-i)}}}\mathbb{1}_{(L_i,U_i)}(Z_i) = f_{X_i|X^{(-i)}}\mathbb{1}_{(L_i,U_i)}(Z_i).$$

Hence one can just apply Theorem (4) to get $f_{X_i|X^{(-i)}}$ and multiply the result with an indicator.

Referring to Theorem (4), it follows that $\Sigma_{22} = aI_{m-1} + bJ_{m-1}$, $\Sigma_{21} = b\mathbf{1}_{m-1}$.

Furthermore, it holds that $\mu_1 = \mu_i$ and $\mu_2 = \mu^{(-i)}$.

It can now be checked that:

$$\Sigma_{22}^{-1} = \frac{1}{a}I_{m-1} - \frac{b}{a((m-1)b+a)}J_{m-1}.$$

Hence

$$\begin{aligned}\Sigma_{12}\Sigma_{22}^{-1} &= b\mathbf{1}'_{m-1}\left(\frac{1}{a}I_{m-1} - \frac{b}{a((m-1)b+a)}J_{m-1}\right) \\ &= b\left(\frac{1}{a} - \frac{(m-1)b}{a((m-1)b+a)}\right)\mathbf{1}'_{m-1} = \frac{b}{(m-1)b+a}\mathbf{1}'_{m-1}.\end{aligned}$$

And thus

$$\mu = \mu_i + \frac{b}{(m-1)b+a}\mathbf{1}'_{m-1}(Z^{(-i)} - \mu_2) = \mu_i + \frac{b}{(m-1)b+a}\sum_{j\neq i}(Z_j - \mu_j).$$

Furthermore:

$$\sigma^2 = a + b - \frac{b^2(m-1)}{(m-1)b+a} = \frac{a(bm+a)}{(m-1)b+a}.$$

□

Appendix C

Alternative Expression of an Equicorrelated Multivariate Normal Integral

In this Appendix, the results stated on Page 49 are shown.

Theorem 5. *Let $n \in \mathbb{N}$, $L, R \in \mathbb{R}^n$ such that $L < R$ elementwise, $\mathbf{0}_n \in \mathbb{R}^n$ have every element equal to 0, and $\tau \in (-1/n, \infty)$.*
The multivariate normal probability

$$\Phi_n\left(\prod_{i=1}^{n}[L_i, R_i); \mathbf{0}_n; I_n + \tau J_n\right) \tag{C.1}$$

can be rewritten as:

$$\mathbb{E}\left[\frac{e^{\frac{\tau\left(\sum_{i=1}^{n} V_i\right)^2}{2(n\tau+1)}}}{\sqrt{n\tau+1}}\right]\prod_{i=1}^{n}\left(\Phi(R_i) - \Phi(L_i)\right), \qquad \textit{where } V_i \sim \mathcal{N}(0,1)|_{[L_i,R_i)} \textit{ for all } i.$$

Proof. Extending the result in Steck and Owen (1962), Expression (C.1) can be written as:

$$\mathcal{I} := \int_{\mathbb{R}}\left[\prod_{i=1}^{n}(\Phi(R_i - \sqrt{\tau}y) - \Phi(L_i - \sqrt{\tau}y))\right]\phi(y)dy.$$

Note that when $\tau < 0$, the term $\sqrt{\tau}y$ is imaginary.
First, consider the case $\tau \geq 0$.

In this case it is easy to see that:

$$\begin{aligned}
\mathcal{I} &= \int_{L_1}^{R_1} \cdots \int_{L_n}^{R_n} (2\pi)^{-n/2} \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\tau}} e^{-\frac{\left[\sum_{i=1}^n (x_i - z)^2\right] + z^2/\tau}{2}} dz \cdot dx_1 \cdots dx_n \\
&= \int_{\prod_{i=1}^n [L_i, R_i)} (2\pi)^{-n/2} e^{-\frac{\|x\|_2^2}{2}} \frac{1}{\sqrt{n\tau + 1}} e^{\frac{\left(\sum_{i=1}^n x_i\right)^2}{2(n+1/\tau)}} \int_{\mathbb{R}} \sqrt{\frac{(1/\tau + n)}{2\pi}} e^{-\frac{(1/\tau+n)\left(z - \frac{\sum_{i=1}^n x_i}{n+1/\tau}\right)^2}{2}} dz \cdot dx \\
&= \int_{\mathbb{R}^n} \frac{e^{\frac{\tau\left(\sum_{i=1}^n x_i\right)^2}{2(n\tau+1)}}}{\sqrt{n\tau+1}} \left[\prod_{i=1}^n \frac{e^{-x_i^2/2} \mathbb{1}_{[L_i, R_i)}(x_i)}{\sqrt{2\pi}}\right] dx.
\end{aligned}$$

Now, consider the case $\tau < 0$.
As ϕ is symmetric around zero, the integral can be written in the following manner:

$$\mathcal{I} = \int_{\mathbb{R}} \frac{1}{2} \left[\prod_{i=1}^n (\Phi(R_i - \sqrt{\tau}y) - \Phi(L_i - \sqrt{\tau}y)) + \prod_{i=1}^n (\Phi(R_i + \sqrt{\tau}y) - \Phi(L_i + \sqrt{\tau}y))\right] \phi(y) dy.$$

Now, for all $z \in \mathbb{R}$:

$$\Phi(z - \sqrt{\tau}y) = e^{|\tau|y^2/2} \int_{-\infty}^{z} e^{ix\sqrt{|\tau|}y} \phi(x) dx.$$

Hence:

$$\begin{aligned}
&\frac{1}{2} \left[\prod_{i=1}^n (\Phi(R_i - \sqrt{\tau}y) - \Phi(L_i - \sqrt{\tau}y)) + \prod_{i=1}^n (\Phi(R_i + \sqrt{\tau}y) - \Phi(L_i + \sqrt{\tau}y))\right] \\
&= \frac{1}{2} \left[\prod_{i=1}^n e^{|\tau|y^2/2} \int_{L_i}^{R_i} e^{ix\sqrt{|\tau|}y} \phi(x) dx\right] + \frac{1}{2} \left[\prod_{i=1}^n e^{|\tau|y^2/2} \int_{L_i}^{R_i} e^{-ix\sqrt{|\tau|}y} \phi(x) dx\right] \\
&= \frac{1}{2} e^{n|\tau|y^2/2} \int_{L_1}^{R_1} \cdots \int_{L_n}^{R_n} \left[\prod_{i=1}^n e^{ix_i\sqrt{|\tau|}y} \phi(x_i)\right] dx_1 \cdots dx_n \\
&+ \frac{1}{2} e^{n|\tau|y^2/2} \int_{L_1}^{R_1} \cdots \int_{L_n}^{R_n} \left[\prod_{i=1}^n e^{-ix_i\sqrt{|\tau|}y} \phi(x_i)\right] dx_1 \cdots dx_n \\
&= e^{n|\tau|y^2/2} \int_{L_1}^{R_1} \cdots \int_{L_n}^{R_n} \frac{e^{i\left[\sum_{i=1}^n x_i\right]\sqrt{|\tau|}y} + e^{-i\left[\sum_{i=1}^n x_i\right]\sqrt{|\tau|}y}}{2} \left[\prod_{i=1}^n \phi(x_i)\right] dx_1 \cdots dx_n \\
&= e^{n|\tau|y^2/2} \int_{L_1}^{R_1} \cdots \int_{L_n}^{R_n} \cos\left(\sqrt{|\tau|}y \sum_{i=1}^n x_i\right) \left[\prod_{i=1}^n \phi(x_i)\right] dx_1 \cdots dx_n.
\end{aligned}$$

When the above expression is integrated against $\phi(y)$ over $\mathbb{R}$, one obtains:

$$\begin{aligned}
\mathcal{I} &= \int_{\mathbb{R}} \int_{\mathbb{R}^n} e^{n|\tau|y^2/2} \cos\left(\sqrt{|\tau|} y \sum_{i=1}^n x_i\right) \left[\prod_{i=1}^n \phi(x_i) \mathbb{1}_{[L_i,R_i)}(x_i)\right] \phi(y)\, dx_1 \cdots dx_n dy \\
&= \int_{\mathbb{R}^n} \int_{\mathbb{R}} \frac{e^{-(1-n|\tau|)y^2/2}}{\sqrt{2\pi}} \cos\left(\mathbf{1}_n' x \sqrt{|\tau|} y\right) dy \cdot f_X(x)\, dx \\
&= \frac{1}{\sqrt{1-n|\tau|}} \int_{\mathbb{R}^n} \int_{\mathbb{R}} \frac{e^{-u^2/2}}{\sqrt{2\pi}} \cos\left(\frac{\mathbf{1}_n' x \sqrt{|\tau|}}{1-n|\tau|} u\right) du \cdot f_X(x) dx \\
&= \frac{1}{\sqrt{1-n|\tau|}} \int_{\mathbb{R}^n} \mathbb{E}\left[\cos\left(\frac{\mathbf{1}_n' x \sqrt{|\tau|}}{\sqrt{1-n|\tau|}} Y\right)\right] \cdot f_X(x) dx \qquad Y \sim \mathcal{N}(0,1)
\end{aligned}$$

In the above, $f_X(x) = \left[\prod_{i=1}^n \phi(x_i) \mathbb{1}_{[L_i,R_i)}(x_i)\right]$ for all $x \in \mathbb{R}^n$.
Now, for every $a \in \mathbb{R}$, by using Fubini's Theorem and the Taylor series for the cosine function, it holds that :

$$\begin{aligned}
&\mathbb{E}[\cos(aY)] = \mathbb{E}\left[\sum_{k=0}^{\infty} \frac{(-1)^k (aY)^{2k}}{2k!}\right] = \sum_{k=0}^{\infty} \frac{(-1)^k a^{2k} \mathbb{E}\left[Y^{2k}\right]}{2k!} = 1 + \sum_{k=1}^{\infty} \frac{(-1)^k a^{2k} (2k-1)!!}{2k!} \\
&= 1 + \sum_{k=1}^{\infty} \frac{(-1)^k a^{2k}}{2k!!} = \sum_{k=0}^{\infty} \frac{(-1)^k a^{2k}}{2k!!} = e^{-a^2/2}.
\end{aligned}$$

In the above, the fact is used that as Y is standard normally distributed,
$\mathbb{E}[Y^{2k}] = (2k-1)!!$ for all $k \in \mathbb{N}$. Here, $k!! = \prod_{i=1}^{\lceil k/2 \rceil} (2i - k\%2)$.
Hence:

$$\mathcal{I} = \frac{1}{\sqrt{1-n|\tau|}} \int_{\mathbb{R}^n} e^{-\frac{S(x)^2}{2(1/|\tau|-n)}} f_X(x) dx = \int_{\mathbb{R}^n} \frac{e^{\frac{\tau(\sum_{i=1}^n x_i)^2}{2(n\tau+1)}}}{\sqrt{n\tau+1}} \left[\prod_{i=1}^n \frac{e^{-x_i^2/2} \mathbb{1}_{[L_i,R_i)}(x_i)}{\sqrt{2\pi}}\right] dx.$$

Hence, for both $\tau < 0$ and $\tau \geq 0$, $\mathcal{I}$ takes on the same expression.

Lastly, by noticing that the term in square brackets above is the unnormalized density of independent truncated normally distributed variables, it holds that:

$$\mathcal{I} = \mathbb{E}\left[\frac{e^{\frac{\tau(\sum_{i=1}^n V_i)^2}{2(n\tau+1)}}}{\sqrt{n\tau+1}}\right] \prod_{i=1}^n (\Phi(R_i) - \Phi(L_i)), \qquad \text{where } V_i \sim \mathcal{N}(0,1)|_{[L_i,R_i)} \text{ for all } i.$$

□

A corollary now follows:

Corollary 5.1. Let $n \in \mathbb{N}$, $L, R \in \mathbb{R}^n$ such that $L < R$ elementwise, let $\mathbf{0}_n$ have every element equal to 0 and $\tau \in (-1/n, 0)$. Lastly, let:

$$P = \Phi_n \left(\prod_{i=1}^{n} [L_i, R_i); \mathbf{0}_n; I_n + \tau J_n \right).$$

Then:

$$\frac{e^{\frac{\tau(I_1+I_2)}{2(n\tau+1)}}}{\sqrt{n\tau+1}} \prod_{i=1}^{n} (\Phi(R_i) - \Phi(L_i)) \leq P \leq \frac{1}{\sqrt{n\tau+1}} \prod_{i=1}^{n} (\Phi(R_i) - \Phi(L_i)) \tag{C.2}$$

where

$$I_1 = \left[\sum_{i=1}^{n} \frac{L_i e^{-\frac{L_i^2}{2}} - R_i e^{-\frac{R_i^2}{2}}}{\sqrt{2\pi} \left(\Phi(R_i) - \Phi(L_i) \right)} \right] + n,$$

$$I_2 = \sum_{i=1}^{n} \sum_{j \neq i} \left(\frac{e^{-\frac{L_i^2}{2}} - e^{-\frac{R_i^2}{2}}}{\sqrt{2\pi} \left(\Phi(R_i) - \Phi(L_i) \right)} \right) \left(\frac{e^{-\frac{L_j^2}{2}} - e^{-\frac{R_j^2}{2}}}{\sqrt{2\pi} \left(\Phi(R_j) - \Phi(L_j) \right)} \right).$$

Proof. The upper bound is trivial when one considers the result of Theorem 5 and notices that $\frac{\tau}{\sqrt{n\tau+1}}$ is negative and $(\sum_{i=1}^{n} V_i)^2$ has to be positive, hence the exponent is always less than 1 for $\tau < 0$.
The lower bound is found by applying Jensen's inequality in Theorem 5, as e^x is convex, it holds that:

$$e^{\frac{\tau \mathbb{E}\left[(\sum_{i=1}^{n} V_i)^2 \right]}{2(n\tau+1)}} \leq \mathbb{E} \left[e^{\frac{\tau (\sum_{i=1}^{n} V_i)^2}{2(n\tau+1)}} \right].$$

It now holds that:

$$\mathbb{E} \left[\left(\sum_{i=1}^{n} V_i \right)^2 \right] = \mathbb{E} \left[\sum_{i=1}^{n} V_i^2 \right] + \mathbb{E} \left[\sum_{i=1}^{n} \sum_{j \neq i} V_i V_j \right].$$

Now, define I_1 to be the left expectation in the expression above, and I_2 to be the rightmost expectation. The result then follows by applying ordinary rules of integration.

□

Appendix D

The falsely claimed error in the method of Lin and Wang

In this section, the inference method used in X. Lin and Wang (2010) to sample γ will be examined. This method was used as a starting point for this research. However, during our research, it was falsely claimed that the method was not correct, and in this section, it will be explained what part of the method we thought was wrong.
In the paper by Lin and Wang, it was the case that, like Z in the ultimately chosen method, the vector γ is not sampled jointly. Similar to Z in the described MH algorithm, the whole vector γ is not sampled. Instead, each element γ_i is Gibbs sampled conditionally on the other values $\gamma^{(-i)}$ and one obtains convergence to the joint posterior in the limit. A prior p_{γ_i} is taken on γ_i, and $\mu_j := X_j\beta$. Furthermore, define

$$T_{jl} = (L_{jl} - R_{jl})\mathbb{I}(L_{jl} > -\infty) + R_{jl}$$

and $\xi_{jl} = \mathbb{I}(R_{jl} < \infty \wedge L_{jl} > -\infty)$ which indicates no censoring.
Now take $V_j = Z_j - \alpha(T_j)$. The idea is to base inference on γ_i on V. It now holds that $V_j \sim \mathcal{N}_{m_j}(\mu_j - \alpha(T_j), \Sigma_j)|_{(\alpha(L_j)-\alpha(T_j),\alpha(R_j)-\alpha(T_j))}$, where $\Sigma_j = I_{m_j} + \tau J_{m_j}$ and hence by Bayes' rule:

$$\begin{aligned} &f_{\gamma_i|(\gamma^{(-i)},\beta,X,V,\tau,L,R)} \\ &\propto p_{\gamma_i} \prod_{j=1}^{n} \frac{e^{-\frac{1}{2}(V_j+\alpha(T_j)-\mu_j)'\Sigma_j^{-1}(V_j+\alpha(T_j)-\mu_j)} \prod_{l=1}^{m_j} \mathbb{1}_{(\alpha(L_{jl})-\alpha(T_{jl}),\, \alpha(R_{jl})-\alpha(T_{jl}))}(V_{jl})}{\mathbb{P}\left(V_{jl} \in [\alpha(L_{jl}) - \alpha(T_{jl}), \alpha(R_{jl}) - \alpha(T_{jl})) \quad \forall l \in \{1, \ldots, m_j\}\right)} \end{aligned}$$

The denominator above was not included in the derivation of the Gibbs sampler provided by Lin and Wang. This is the case as the eventual Gibbs sampling step (D.1) for γ_i will match with that of the one provided in X. Lin and Wang (2010). However, we thought

the denominator **had** to be included to normalize the likelihood of V used in Bayes' rule. However, the likelihood function is with respect to both V and L, R and can be seen to integrate to 1 with respect to these variables.
Now, it is shown how continuing without the denominator yields the Gibbs sampling step provided by Lin and Wang:

$$\begin{aligned} & p_{\gamma_i} \prod_{j=1}^{n} e^{-\frac{1}{2}(V_j+\alpha(T_j)-\mu_j)'\Sigma_j^{-1}(V_j+\alpha(T_j)-\mu_j)} \prod_{l=1}^{m_j} \mathbb{1}_{(\alpha(L_{jl})-\alpha(T_{jl}),\, \alpha(R_{jl})-\alpha(T_{jl}))}(V_{jl}) \\ & \propto p_{\gamma_i} \prod_{j=1}^{n} e^{-\frac{1}{2}(V_j+\alpha(T_j)-\mu_j)'\Sigma_j^{-1}(V_j+\alpha(T_j)-\mu_j)} \prod_{l=1}^{m_j} \mathbb{1}_{(0,\, \alpha(R_{jl})-\alpha(L_{jl}))}(V_{jl})^{\xi_{jl}}. \end{aligned}$$

The last step holds as in the cases of left and right censoring, the indicators do not depend on γ_i. Now, define:

$$\begin{aligned} \sigma_l^j &= \Sigma_j^{-1/2} B_l^{d,\mathbf{t}}(T_j), \\ \tilde{\mu}_j &= \Sigma_j^{-1/2} \mu_j, \\ \tilde{V}_j &= \Sigma_j^{-1/2} V_j, \\ \delta_{lm}^j &= B_m^{d,\mathbf{t}}(R_{jl}) - B_m^{d,\mathbf{t}}(L_{jl}), \\ \hat{\mu}_{ij} &= \tilde{\mu}_j - \sum_{l \neq i} \sigma_l^j \gamma_l - \tilde{V}_j, \end{aligned}$$

and let $\sigma_l, \hat{\mu}_i$ be the concatenation of σ_l^j and $\hat{\mu}_{ij}$ for all j respectively.
It follows that:

$$\begin{aligned} & f_{\gamma_i|(\gamma^{(-i)},\beta,X,Z,\tau,L,R)} \\ & \propto p_{\gamma_i} \prod_{j=1}^{n} e^{-\frac{1}{2}(\sigma_i^j\gamma_i+\sum_{l\neq i}\sigma_l^j\gamma_l-\tilde{\mu}_j+\tilde{V}_j)'(\sigma_i^j\gamma_i+\sum_{l\neq i}\sigma_l^j\gamma_l-\tilde{\mu}_j+\tilde{V}_j)} \prod_{l=1}^{m_j} \left(\mathbb{1}_{\left(-\delta_{li}^j\gamma_i,\, \sum_{m\neq i}\delta_{lm}^j\gamma_m\right)}(V_{jl} - \delta_{li}^j\gamma_i) \right)^{\xi_{jl}} \\ & = p_{\gamma_i} e^{-\frac{1}{2}\sum_{j=1}^{n}(\sigma_i^j\gamma_i-\hat{\mu}_{ij})'(\sigma_i^j\gamma_i-\hat{\mu}_{ij})} \prod_{j=1}^{n}\prod_{l=1}^{m_j} \left(\mathbb{1}_{\left(\frac{V_{jl}-\sum_{m\neq i}\delta_{lm}^j\gamma_m}{\delta_{li}^j},\, \gamma_i+\frac{V_{jl}}{\delta_{li}^j}\right)}(\gamma_i) \right)^{\xi_{jl}} \\ & = p_{\gamma_i} e^{-\frac{1}{2}(\sigma_i\gamma_i-\hat{\mu}_i)'(\sigma_i\gamma_i-\hat{\mu}_i)} \mathbb{I}\left(\gamma_i > \max_{\{(j,l):\xi_{jl}=1\}} \frac{V_{jl} - \sum_{m\neq i}\delta_{lm}^j\gamma_m}{\delta_{li}^j} \right). \end{aligned}$$

The last equation follows from the fact that for $\xi_{jl} = 1$ it will a.s. be the case that $V_{jl} > 0$ hence $\gamma_i + \frac{V_{jl}}{\delta_{li}^j} > \gamma_i$ a.s. and one can drop this requirement and get the indicator with the maximum.

The exponent above also appears in Bayesian linear regression with one covariate, hence it is well known (Gelman et al., 2013) that the last expression can be rewritten as:

$$f_{\gamma_i|(\gamma^{(-i)},\beta,X,Z,\tau,L,R)} \propto p_{\gamma_i} e^{-\frac{\sigma_i'\sigma_i(\gamma_i-\sigma_i'\hat{\mu}_i)^2}{2}} \mathbb{I}\left(\gamma_i > \max_{\{(j,l):\xi_{jl}=1\}} \frac{V_{jl} - \sum_{m\neq i}\delta_{lm}^j\gamma_m}{\delta_{li}^j}\right).$$

From the structure of the above posterior, we see that the truncated normal distribution is a conjugate prior for γ_i and hence if one sets a $\mathcal{N}(\mu_0,\sigma_0^2)|_{(a_0,b_0)}$ prior on γ_i, it holds that:

$$\begin{aligned}
&\gamma_i|(\gamma^{(-i)},\beta,X,Z,\tau,L,R) \sim \mathcal{N}(\mu_1,\sigma_1^2)|_{(a_1,b_1)} \qquad \text{(D.1)}\\
&\sigma_1^2 = \frac{1}{\sigma_i'\sigma_i + (\sigma_0^2)^{-1}},\\
&\mu_1 = \sigma_1^2\left(\frac{\mu_0}{\sigma_0^2} + \sigma_i'\hat{\mu}_i\right),\\
&a_1 = \max\left(\left(\max_{\{(j,l):\xi_{jl}=1\}} \frac{V_{jl} - \sum_{m\neq i}\delta_{lm}^j\gamma_m}{\delta_{li}^j}\right), a_0\right),\\
&b_1 = b_0.
\end{aligned}$$

Appendix E

Mathematical Formulation

E.1 Introduction

In this Appendix, the situation of a group of randomized controlled trials in medicine is modeled. This was done in order to get a grip on the problem at hand. Eventually, this model/formulation was not used in this research. After inspection of Chapter 2, some common aspects of clinical trials are:

1. Subjects enter a clinical trial at a random time point σ_i for every patient i, and leave the trial at a random time point τ_i.

2. According to some (possibly randomized) assignment rule, based on observed patient features (and possibly all patient/outcome data collected up to that point), the patient is selected for one of the treatment groups.

3. Call $Y_{\mathcal{G}}|k$ the possible outcomes of a subgroup $\mathcal{G}$ of patients when they would have been assigned to treatment group k. Based on the recorded outcomes, treatment assignments and patient features, a test (often based on the mean) is now performed with the null hypothesis that $Y_{\mathcal{G}}|k$ is distributed as $Y_{\mathcal{G}}|0$. Testing can be done on multiple subgroups, treatment groups, and at multiple time points.

This leads to the following mathematical description of an RCT, which is expressed in terms of measure theoretic probability.

E.2 description

First, let $(\Omega, \mathcal{F}, \mathcal{F}_t)$ be a filtered measurable space, where t denotes time. Furthermore, let $k \in \mathbb{N} \cup \{\infty\}$ be the number of features having an influence on trial outcomes. It is assumed that each of these features lie on/have an injective mapping to the real line. Hence, the features of patient i at time t can be represented by an $\mathcal{F}_t-$adapted process:

$$\mathcal{X}^i : \Omega \times \mathbb{R}_+ \to \mathbb{R}^k \qquad \forall i \in \mathbb{N}.$$

In the above definition (and in the following definitions), the first input variable induces (the possibility of) randomness in the process, and the second input variable denotes time. It is seen from this formulation that every feature has the possibility to stochastically change in time. Note also that as i can take on any value in $\mathbb{N}$, the patient population is infinite.
Now, let there be $d \in \mathbb{N}$ decision variables (e.g. the assigned treatment group, the location of the trial center), which (like the features) are assumed to lie in/have an injective mapping to the real line. An *assignment policy* is now defined as a $\mathcal{F}_t$-adapted stochastic process

$$\pi^i : \Omega \times \mathbb{R}_+ \to \mathbb{R}^d$$

where i denotes the patient for which the decision is made.
The outcome process is now defined:

$$\mathcal{Y}^i : \Omega \times \mathbb{R}_+ \to \mathbb{R}^l \qquad \forall i \in \mathbb{N}$$

is the $\mathcal{F}_t-$adapted, $(l)-$dimensional outcome process for patient i, where $l \in \mathbb{N} \cup \{\infty\}$.

In the following, if Z^i is a stochastic process for patient i, Z^i_t denotes the process evaluated at t (which is a random vector) and Z denotes the matrix-valued process where the rows correspond to the patients. Note that dependencies between $\mathcal{X}, \pi^i$ and $\mathcal{Y}^i$ can all be modeled with the probability measure that is eventually chosen for the measurable space $(\Omega, \mathcal{F}, \mathcal{F}_t)$.
Next, censoring of information is modelled. Let N be the total number of patients recorded in the trial, let $(j_i)_{i=1}^N$ be a sequence in $\mathbb{N}$ of length N denoting the patients included in the trials. Assume that from these patients, $X^1, \ldots, X^N$ and $Y^1, \ldots, Y^N$ are measured, where X^i only measures $m_i \leq k$ number of features in $\mathcal{X}^{j_i}$ and Y^i only measures $n_i \leq l$ outcome variables in $\mathcal{Y}^{j_i}$.
Let σ_i denote the arrival/diagnosis time of patient i, and let τ_i be the "dropout" time for

patient i, where $\sigma_i \leq \tau_i$. Assume they are both stopping times with respect to $\mathcal{F}_t$. It then holds that

$$\nu_t^i = (t \wedge \tau_i) \cdot \mathbb{1}_{[0,t]}(\sigma_i)$$

is also a stopping time w.r.t $\mathcal{F}_t$ for every $t \in \mathbb{R}_+$. In the above, $\wedge$ denotes the binary minimum operator. ν_t^i is a stopping time (for all t, i) which remains zero until patient i enters the trial, after which it jumps to σ_i and increases linearly with slope 1 in t until the patient leaves the trial at time τ_i. The total filtration evaluated at these stopping times hence exactly measures information of patient i up until time t. For an example of a path of ν_t^i, see Figure E.1.

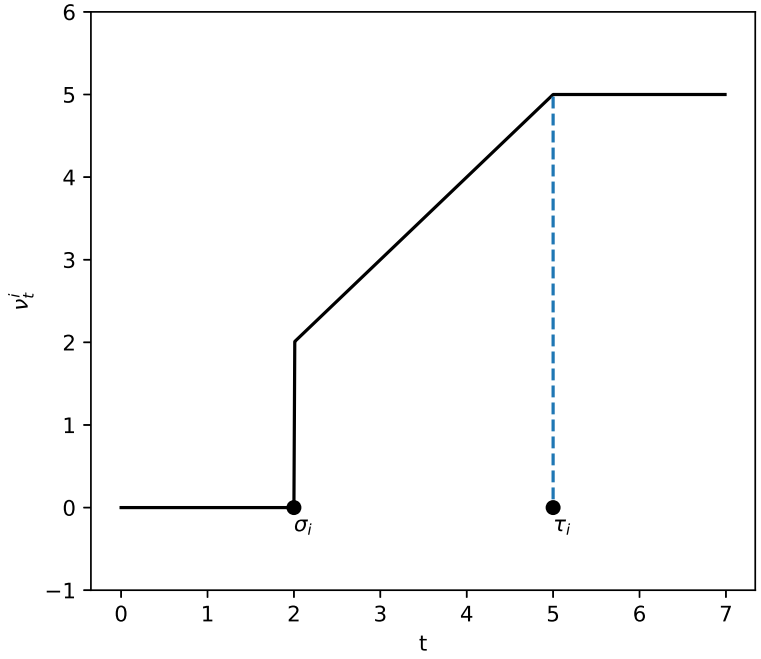


Figure E.1: Example of a path of ν_t^i in the case $\sigma_i = 2$ and $\tau_i = 5$.

Hence, define

$$\mathcal{G}_t = \prod_{i=1}^{n} \mathcal{F}_{\nu_t^i}^{(X^i, Y^i)} \cap \mathcal{F}_t^{\pi^i}$$

to be the filtration where the observed patient features and outcomes are observed inside the observation intervals (σ_i, τ_i) as well as the treatment policies up to time t. In the above definition, the product is Cartesian. To explain some notation, for an $\mathcal{F}_t$-adapted process X, the filtration $\mathcal{F}_t^X$ contains the $\sigma-$algebra's generated by X up to

each time point t. It hence only contains the minimal required information to measure X. Furthermore, for an $\mathcal{F}_t$ stopping time τ, $\mathcal{F}_\tau = \{A \in \mathcal{F}_\infty : A \cap \{\tau \leq t\} \in \mathcal{F}_t \quad \forall t \in \mathbb{R}_+\}$. In the case of statistical hypothesis testing based on trial outcomes, there is a fixed/pre-determined probability measure $\mathbb{P}^\pi$ over the treatment policies. A treatment policy π^* is then drawn from this probability measure and patient features and outcomes are collected up to some finite time T in a trial based on this policy. One now chooses a set of models $\{\mathcal{M}_1^{\pi^*}, \ldots, \mathcal{M}_r^{\pi^*}\}$ based on the sampled policy and determines for which of these models the likelihood of the observations (X, Y) is the largest.
In the case where decision making is applied (e.g. sequential testing or a multi-armed bandit approach), the situation is different. Assume a decision has to be made at time t and one has observed $(X_s^i, Y_s^i)_{s \in [0, \nu_t^i]}$, as well as π^i up to time t for all followed patients i. Let $E \in \mathcal{G}_t$ be the corresponding set in $\mathcal{G}_t$ on which the observed event takes place. Let $\mathcal{Q}$ be the set of all possible treatment policies. One now chooses a class of probability measures as the *model* for the data

$$\mathcal{P} = \{\mathbb{P}_\pi | \mathbb{P}_\pi \text{ is a probability measure on } (\Omega, \mathcal{F}) \text{ for every } \pi \in \mathcal{Q}\}.$$

Let $\mathcal{O} := \cup_{i=1}^{\infty} \{f \mid f : \mathbb{R}_+ \mapsto \mathbb{R}^l\}$ be the set of all possible outcomes for the process $\mathcal{Y}$. Let $\mathcal{L} : \mathcal{O} \to \mathbb{R}$ be an $\mathcal{F}-$measurable loss function based on all outcome processes. The objective in decision making in clinical trials is now to find a probability measure $\mathbb{Q}$ over all possible policies π that minimizes (possibly approximately):

$$\int_{\mathcal{Q}} \mathbb{E}_\pi \left[\mathcal{L}(Y) \mathbb{1}_E\right] d\mathbb{Q}(\pi).$$

In words, find the policy distribution, that given the observed event E minimizes the expected loss induced by the outcome process under the chosen model $\mathcal{P}$ for the data.

Appendix F

Test Martingales

This section relates Bayes factors to a recently studied class of stochastic processes called *test martingales*. If the clinical trial is regarded as a prospective study rather than a retrospective study (the latter of which is mostly assumed throughout this thesis), one can view the Bayes factor value based on all information up to that time point as a stochastic process B_n (where discrete time is assumed for convenience). This process is related to a class of processes called test martingales (Shafer et al., 2011), (Grünwald, 2016), (Hendriks, 2018). Hence, studying these types of processes could provide results applicable on both clinical trials and other situations of sequential testing. First, a definition is given:

Definition 5. A test martingale is a stochastic process M on $\mathcal{L}^1(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n=1}^{\infty}, \mathbb{P})$ for some set Ω, σ-algebra $\mathcal{F}$ for Ω, filtration $(\mathcal{F}_n)_{n=1}^{\infty}$ and probability measure $\mathbb{P}$ such that:

1. $M_1 = 1, \mathbb{P} - a.s.$
2. $M_n \geq 0$ for all $n \in \mathbb{N}$,
3. $\mathbb{E}[M_{n+1}|\mathcal{F}_n] = M_n \quad \mathbb{P} - a.s.$.

By Doob's martingale inequality for nonnegative (sub)martingales, denoting $M^* = \sup_{n\in\mathbb{N}} M_n$, it holds that:

$$\mathbb{P}\left(\frac{1}{M^*} \leq \alpha\right) = \mathbb{P}\left(M^* \geq \frac{1}{\alpha}\right) \leq \alpha.$$

By Doob's first convergence theorem, M_∞ exists $\mathbb{P}$-almost surely. Furthermore, because of monotonicity, M^* and $1/M^*$ are well-defined (possibly infinite) random variables. The latter variable satisfies (3.4), and it is hence a p-value. Furthermore, one can ask whether more links exist between p-values and test martingales, as well as

Bayes factors and test martingales. The following relations show that a test martingale is a generalization of p-values and Bayes factors:

Theorem 6. *Let M be a test martingale, P be an exact p-value and B be a Bayes factor. The following statements hold:*

1. $(M^*)^{-1}$ *is a p-value.*
2. $(n\mathbb{1}_{[0,1/P)}(n))_{n=1}^{\infty}$ *is a test martingale.*
3. M_n^{-1} *for all n and M_∞^{-1} are Bayes factors.*
4. $(\mathbb{E}[B_k^{-1} \mid \mathcal{F}_n])_{n=1}^{\infty}$ *is a test martingale in n for all k. Furthermore, if the observations are independent, $(B_n)_{n=1}^{\infty}$ is a test martingale.*

Proof. see (Shafer et al., 2011). □

In 3.7, it was seen that the Bayes factor for a total dataset can be written as the product of conditional Bayes factors. It is now shown that this property holds for all test martingales.

Theorem 7. *Let $(M^i)_{i=1}^k$ be a sequence of independent* **stopped** *test martingales on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with $k \in \mathbb{N} \cup \{\infty\}$. Let $\mathcal{F}^i$ be the filtrations generated by the martingales, and let τ_i be the stopping time corresponding to M^i.*
Define $s_0 = 0$, $s_i = \sum_{j=1}^{i} \tau_j$ for $i \in \{1, \dots, k\}$.
Let $\mathcal{T} = \{\emptyset, \Omega\}$ be the trivial sigma algebra, and

$$\mathcal{F}_n^{i,l} = \begin{cases} \mathcal{T} & n \leq l, \\ \mathcal{F}_{n-l}^i & n > l. \end{cases}$$

Be a right-shifted version of $\mathcal{F}_n^i$ for every $n, l \in \mathbb{N}$.
Now, let $\mathcal{F}_n^{s_1} = \mathcal{F}_n^1$, $\mathcal{F}_n^{s_0} = \mathcal{T}$ for all n, and recursively define the filtrations:

$$\mathcal{F}_n^{s_i} = \{A \in \mathcal{F} : A \cap \{s_{i-1} = l\} \in \mathcal{F}_n^{i,l} \cap \mathcal{F}_l^{s_{i-1}} \quad \forall l \in \mathbb{N} \cup \{0\}\} \qquad \text{for } i \in \{2, \dots, k\},\ n \in \mathbb{N}.$$

Then

$$M_n = \prod_{i=1}^{k} \left(\mathbb{1}_{\{s_{i-1} \geq n\}} + M_{n-s_{i-1}}^i \mathbb{1}_{\{s_{i-1} < n\}} \right)$$

is a test martingale with respect to the filtration $(\mathcal{F}_n^{s_k})_{n=1}^{\infty}$.

Proof. The measurability statement $\{s_i = n\} \in \mathcal{F}_n^{s_i}$ holds by induction.
Furthermore, $\mathcal{F}_n^{s_i} \subset \mathcal{F}_n^{s_{i+1}}$ for all i, n and as the term $\mathbb{1}_{\{s_{i-1} \geq n\}} + M^i_{n-s_{i-1}} \mathbb{1}_{\{s_{i-1} < n\}} \in \mathcal{F}_n^{s_i}$, it holds that the product M_n is $\mathcal{F}_n^{s_k}$ measurable. Furthermore, it trivially holds that M_n nonnegative and $M_1 = M_1^1 = 1$.
Now all that's left is to prove the martingale inequality, as this also shows that M is integrable by nonnegativity and the tower rule:

$$\mathbb{E}\left[M_{n+1}|\mathcal{F}_n^{s_k}\right] = \mathbb{E}\left[\sum_{i=1}^{k} \mathbb{1}_{\{s_{i-1}<n<s_i\}} \prod_{j=1}^{i-1} M^j_{\tau_j} M^i_{n+1-s_{i-1}} | \mathcal{F}_n^{s_k}\right] + \mathbb{E}\left[\sum_{i=1}^{k-1} \mathbb{1}_{\{n=s_i\}} \prod_{j=1}^{i} M^j_{\tau_j} | \mathcal{F}_n^{s_k}\right]$$
$$+ \mathbb{E}\left[\mathbb{1}_{\{n \geq s_k\}} \prod_{j=1}^{k} M^j_{\tau_j} | \mathcal{F}_n^{s_k}\right] = \sum_{i=1}^{k} \mathbb{1}_{\{s_{i-1}<n<s_i\}} \prod_{j=1}^{i-1} M^j_{\tau_j} \mathbb{E}\left[M^i_{n+1-s_{i-1}} | \mathcal{F}_n^{s_k}\right] + \sum_{i=1}^{k-1} \mathbb{1}_{\{n=s_i\}} \prod_{j=1}^{i} M^j_{\tau_j}$$
$$+ \mathbb{1}_{\{n \geq s_k\}} \prod_{j=1}^{k} M^j_{\tau_j}.$$

The second equality holds as $\mathbb{1}_{\{s_i \leq n\}} M^i_{\tau_i}$ is $\mathcal{F}_n^{s_i}$ (and thus $\mathcal{F}_n^{s_k}$) measurable, namely:

$$\mathbb{1}_{\{s_i \leq n\}} M_{\tau_i} \mathbb{1}_{\{s_{i-1}=k\}} = \mathbb{1}_{\{\tau_i \leq n-k\}} M_{\tau_i} \mathbb{1}_{\{s_{i-1}=k\}} \in \mathcal{F}_n^{i,k} \cap \mathcal{F}_k^{s_{i-1}} \qquad \forall k \in \mathbb{N}.$$

Now, for all $A \in \mathcal{F}_n^{s_k}$, by the definition of conditional expectation and as M^i_{n-l} is a martingale on $\mathcal{F}_n^{i,l}$ for $n > l$:

$$\mathbb{E}\left[\mathbb{1}_{\{s_{i-1}<n<s_i\}} M^i_{n+1-s_{i-1}} \mathbb{1}_A\right] = \mathbb{E}\left[\sum_{l=1}^{n-1} \mathbb{1}_{\{s_{i-1}=l\}} M^i_{n+1-l} \mathbb{1}_{A \cap \{s_i > n\}}\right]$$
$$= \mathbb{E}\left[\sum_{l=1}^{n-1} \mathbb{1}_{\{s_{i-1}=l\}} M^i_{n-l} \mathbb{1}_{A \cap \{s_i > n\}}\right] = \mathbb{E}\left[\mathbb{1}_{\{s_{i-1}<n<s_i\}} M^i_{n-s_{i-1}} \mathbb{1}_A\right].$$

The second equation holds as $A \cap \{s_i > n\} \in \mathcal{F}_n^{s_i}$ for $i \in \{1, \ldots, k\}$ by induction, hence

$$A \cap \{s_i > n\} \cap \{s_{i-1} = l\} \in \mathcal{F}_n^{i,l} \cap \mathcal{F}_l^{s_{i-1}}.$$

Now, by independence of the martingales:

$$\mathbb{E}[M^i_{n+1+l} | \mathcal{F}_n^{i,l} \cap \mathcal{F}_l^{s_{i-1}}] = \mathbb{E}[M^i_{n+1+l} | \mathcal{F}_n^{i,l}] = M^i_{n+l}.$$

Hence $\mathbb{E}\left[\mathbb{1}_{\{s_{i-1}<n<s_i\}} M^i_{n+1-s_{i-1}} | \mathcal{F}_n^{s_k}\right] = \mathbb{1}_{\{s_{i-1}<n<s_i\}} M^i_{n-s_{i-1}}$ and the martingale equation holds.

□

The interpretation of the filtration $\mathcal{F}_n^{s_k}$ is that it measures the martingale M_n up to $\min(n, s_k)$.

Appendix G

Frequentist Hypothesis Tests

In this appendix, more information is given about the frequentist hypothesis tests often used in clinical trials.

G.0.1 Qualitative responses

Qualitative responses are recorded in cases where it is difficult/not possible to record outcomes numerically. This can be for instance in the case that the outcome is e.g. a test result or a psychological effect. Mathematically speaking, the outcome set of a qualitative response is a finite countable set. Assume without loss of generality (wlog) that the responses are denoted by $\{1, \dots, K\}$ for some $K \in \mathbb{N}$. The two treatment groups are denoted by $\{0, 1\}$. Let X_i be the outcome for subject i, and let $T_i \in \{0, 1\}$ be his treatment group. Let N be the sample size, N_0 be the number of subjects in treatment group 0, and $N_1 := N - N_0$. Lastly, let N_T^k be the number of subjects in treatment group $T \in \{0, 1\}$ having response k, and let $N^k := N_0^k + N_1^k$.
In this section, the null hypothesis is as follows :

$$\mathbb{P}_0(X_i = k) = \mathbb{P}_0(X_j = k) = p_k \qquad \forall i, j, k. \tag{G.1}$$

In the above, $\mathbb{P}_0$ denotes the probability under the null hypothesis.

Fisher exact test

A Fisher exact test is often performed in clinical trials when the sample size is small. Under the null hypothesis (G.1), it follows that:

$$\mathbb{P}_0(N_0^1 = m_1, \ldots, N_0^K = m_K | N_T) = \binom{N_0}{m_1, \ldots, m_K} \prod_{i=1}^{n} p_i^{m_i} \binom{N_1}{N^1 - m_1, \ldots, N^K - m_K} \prod_{i=1}^{n} p_i^{N^i - m_i},$$

$$\mathbb{P}_0(N^1 = l_1, \ldots, N^K = l_K) = \binom{N}{l_1, \ldots, l_K} \prod_{i=1}^{K} p_i^{l_i},$$

hence $\mathbb{P}_0(N_0^1 = m_1, \ldots, N_0^K = m_K | N^1, \ldots, N^K)$

$$= \frac{\binom{N_0}{m_1,\ldots,m_K} \prod_{i=1}^{n} p_i^{m_i} \binom{N_1}{N^1-m_1,\ldots,N^K-m_K} \prod_{i=1}^{n} p_i^{N^i-m_i}}{\binom{N}{N^1,\ldots,N^K} \prod_{i=1}^{n} p_i^{N^i}} = \frac{\binom{N_0}{m_1,\ldots,m_K} \binom{N_1}{N^1-m_1,\ldots,N^K-m_K}}{\binom{N}{N^1,\ldots,N^K}}.$$

Call the latter conditional probability $\pi(m, n)$ for $m, n \in \mathbb{N}^K$ (taking $\binom{l}{k} = 0$ for $k > l$). If $m, n \in \mathbb{N}^K$ are now observed (i.e. $n_k = N_1^k$ and $m_k = N_0^k$), the p-value P in the Fisher exact test is calculated as:

$$P = \sum_{k,l \in \mathbb{N}^K : \pi(k,l) \leq \pi(m,n)} \pi(k, l).$$

In words, the sum of probabilities of all less likely outcomes given $N^1, \ldots, N^K$.

The upside of this test is that it does not assume knowledge of the probabilities p_k, the downside of the test is that the p-value is calculated conditional on the distribution of subjects over outcomes. This makes the test more conservative, and decreases the statistical power as compared to similar tests (see e.g. Barnard's test in Mehta and Senchaudhuri (2003)). The p-value in the Fisher exact test can also be based on other statistics (see for instance Mehta and Senchaudhuri (2003)), the p-value in this section is considered standard, and based on e.g. Raymond and Rousset (1995).

Chi-square test

The Chi-square test is often performed in trials with large sample sizes. It is based on the following result:

Theorem 8. *Let $\hat{p}_k = \frac{N^k}{N}$ for all k, then under (G.1) :*

$$T_N := \sum_{k=1}^{K} \frac{(N_0^k - N_0 \hat{p}_k)^2}{N_0 \hat{p}_k} + \frac{(N_1^k - N_1 \hat{p}_k)^2}{N_1 \hat{p}_k} \xrightarrow{d} \chi^2_{(K-1)} \text{ as } N \to \infty.$$

Proof. See e.g. Van der Vaart (2000), page 247. □

In the above $\xrightarrow{d}$ denotes convergence in distribution. The p-value of this test is now taken to be $1-F_{\chi^2_{(K-1)}}(T_N)$. For more information about the Chi-square test, see McHugh (2013).

Mann-Whitney U Test

If there exists a binary ordering relation $<$ on the qualitative outcome space, we can assume wlog that the numbers linked to the outcomes have the same ordering, and we are dealing with an *ordinal outcome space*. In this case, the Mann-Whitney U test can also be used to compare outcomes in two different treatment groups, either for a small or large sample.

In this test, every subject i is given a rank $R_i \in \mathbb{Q}_+$, where there exists $\delta \in \mathbb{Q}_+$ such that for all patients i, j :

$$\begin{aligned} X_i < X_j &\iff R_i < R_j, \\ X_i = X_j &\iff R_i = R_j, \\ X_i = k,\ X_{i+1} = k+1 &\implies R_{i+1} - R_i = \delta, \\ \sum_{i=1}^{N} R_i &= \frac{N(N+1)}{2}. \end{aligned}$$

Let R'_k be the ranking of subjects having outcome k, these conditions lead to an invertible linear system in $R'_1, \ldots, R'_K$ and hence the ranks are uniquely defined.

Let R_i^0 be the ranks for the i-th subject in treatment group 0, and let R_i^1 be defined similarly. Following Hollander, Wolfe, and Chicken (2013), the Mann-Whitney U statistic is now equal to:

$$U = \left[\sum_i R_i^0\right] - \frac{N_0(N_0+1)}{2}.$$

Remarkably, many other definitions of U are available in literature (Hollander et al. (2013), LaMorte (2019)). The test on the above statistic are equivalent or more standard than tests based on other statistics.

Assuming (G.1), the statistic U should be distributed as if the sum of ranks were taken from any given subset of size N_0 of the total set of rankings. Hence, to get an exact conditional p-value, the frequency of all subsets of outcomes such that the resulting value of the statistic U is smaller/larger than the observed value (one-sided) or both (two-sided) is determined.

Sometimes, in large samples, a test based on asymptotic normality of the scaled/translated U statistic is performed (which was shown by Mann and Whitney Mann and Whitney (1947)). This mainly has to do with the fact that the possible number of subsamples grows fast with sample size.

G.0.2 Quantitative Responses

Let X_i now be the responses for the subjects in treatment group 0, and Y_i those for treatment group 1. Let $\overline{X}_N$ denote the sample average for the outcomes in group 0, and $\overline{Y}_N$ the average for group 1. Lastly, let s_X, s_Y denote the sample standard deviation for group 0 and 1 respectively.

Independent Two-sample t-test

If the data is assumed to follow a normal distribution, the independent two-sample t-test can be conducted[1]. We take as null hypothesis:

$$\exists \mu \in \mathbb{R}, \sigma_0, \sigma_1 \in \mathbb{R}_+, \qquad \begin{bmatrix} X_i \\ Y_j \end{bmatrix} \overset{iid}{\sim} N_2 \left(\begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & 0 \\ 0 & \sigma_1^2 \end{bmatrix} \right), \qquad \forall i \in \{1, \dots, N_0\},\ j \in \{1, \dots, N_1\}.$$

Two independent-sample t-tests can be performed, one is Welch's t-test (Pocock, 2013) and one is the independent-sample Student's t-test (Walker & Almond, 2010).

1. independent-sample Student's t-test

In this case, it is furthermore assumed under the null hypothesis that $\sigma_0 = \sigma_1$. The test statistic:

$$T = \frac{\overline{X}_N - \overline{Y}_N}{s_p \sqrt{\frac{1}{N_0} + \frac{1}{N_1}}}, \quad \text{where} \quad s_p = \sqrt{\frac{(N_0 - 1)s_X^2 + (N_1 - 1)s_Y^2}{N_0 + N_1 - 2}}$$

is known (see e.g. Armitage et al. (1971)) to have a $T_{N_0+N_1-2}-$distribution under the null hypothesis, which can be used to form a p-value for the test.

2. Welch's independent-sample $t-$test

In Welch's independent-sample t-test, it is assumed that $\sigma_1 \neq \sigma_0$. The test statistic in Welch's test is now defined as:

$$T = \frac{\overline{X}_N - \overline{Y}_N}{\sqrt{s_X^2/N_0 + s_Y^2/N_1}}.$$

[1] If this assumption is violated, the nonparametric Mann-Whitney U test is sometimes also performed.

The statistic is *approximately* follows a $T_\nu-$distribution where $\nu \in \mathbb{N}$ is the largest natural number such that:

$$\nu \leq \frac{(s_X^2/N_0 + s_Y^2/N_1)^2}{s_X^4/(N_0^2(N_0-1)) + s_Y^4/(N_1^2(N_1-1))}.$$

A downside of Welch's t-test is that it is based on an approximation of a distribution. The test is however seen to outperform the Student $t-$test under unequal variances in practical applications (Ruxton, 2006).

G.0.3 Time to Event Responses

Time to event responses are analyzed using so called *survival analysis*. The events of interest are e.g. recurrence to the clinic or time of death. The subjects arrive to the trial at some time, and are followed until the event of interest is observed, or the response is *right censored*. This censoring entails basically all events that causes the time-to-event measurement to stop early. Examples of this are e.g. patients leaving the trial, stopping of the trial or death due to other causes than the disease. Next to right censoring, left censoring and interval censoring can also occur. Left censoring occurs when only an upper bound on the event time is known. This can happen for instance when patients are added to a clinical trial retrospectively (e.g. when the inclusion criteria are redefined) and the disease has already recurred before the patients enter the trial. Interval censoring occurs when only the interval in which the event time lies is known. This can happen when monitoring is not done continuously but only between certain time points. Denote the event times for subjects in the control group with T_i^0 and the times for subjects in the treatment group with T_i^1. The *observations* for subject i in survival analysis are now time intervals $[L_i, R_i)$ in which T_i lies. When no censoring occurs, this interval becomes degenerate, but in the case of left/right/interval censoring, it contains more than one point. In frequentist hypothesis testing, the following null hypothesis is now often posed:

$$S_X(t) := \mathbb{P}_0(T_i^0 > t) = \mathbb{P}_0(T_i^1 > t) =: S_Y(t) \qquad \forall t \in \mathbb{R}_+.$$

The functions $S_X(t)$ and $S_Y(t)$ are often called the survival functions for groups X and Y respectively. The null hypothesis hence states that the survival functions for both groups are identical.

Logrank Test for Event-time Data

The logrank test is often performed in cases where interval censoring takes place. It is hence the case that the observation for each patient i in treatment group $j \in \{0, 1\}$ is

the tuple (L_i^j, R_i^j) with $R_i^j \geq L_i^j$ and $L_i^j, R_i^j \in \{t_1, \ldots, t_m\}$ for $t_i \in \mathbb{R}\ \forall i$.
To construct the test statistic, define

$$\begin{aligned}
O_j^0 &= \sum_{i=1}^{N_0} \mathbb{I}((R_i^0 < t_j) \wedge (L_i^0 \geq t_{j-1})), \\
O_j^1 &= \sum_{i=1}^{N_1} \mathbb{I}((R_i^1 < t_j) \wedge (L_i^1 \geq t_{j-1})), \\
N_j^0 &= O_j^0 + \sum_{i=1}^{N_0} \mathbb{I}(L_i^0 \geq t_{j-1}), \\
N_j^1 &= O_j^1 + \sum_{i=1}^{N_1} \mathbb{I}(L_i^1 \geq t_{j-1}), \\
O_j &= O_j^0 + O_j^1, \\
N_j &= N_j^0 + N_j^1.
\end{aligned}$$

In words, O_j^0, O_j^1, O_j are the number of subjects in group $0, 1$ or both groups experiencing the event between t_{j-1} and t_j, and N_j^0, N_j^1, N_j are the number of subjects at risk of experiencing the event between t_{j-1} and t_j. Note that N_j^0, N_j^1, N_j does not count the number of censored subjects in the period $[t_{j-1}, t_j]$.
Similarly to the situation in Fisher's exact test above, under the null hypothesis, $Z_j := O_j^0 | O_j, N_j, N_j^0$ are independently drawn (for all j) from the hypergeometric distribution:

$$\mathbb{P}_0(O_j^0 = k | O_j, N_j, N_j^0) = \frac{\binom{N_j^0}{k}\binom{N_j^1}{O_j - k}}{\binom{N_j}{O_j}} \qquad \forall k \in \mathbb{N} \cup \{0\}.$$

Like in the Fisher exact test, all outcomes with a smaller probability could be summed up to obtain a p-value. However, with large sample sizes the number of possible combinations of outcomes grows very fast. Hence an asymptotic approximation is often performed. The expectation and variance of Z_j are

$$\mu_j = \frac{O_j N_j^0}{N_j}, \quad \mathbb{V}_j = \frac{N_j^0 N_j^1 (N_j - O_j) O_j}{N_j^2 (N_j - 1)}.$$

Hence, by Lindenberg's central limit theorem (Pollard, 2002), as $m \to \infty$:

$$\frac{\sum_{j=1}^m (Z_j - \mu_j)}{\sqrt{\sum_{j=1}^m \mathbb{V}_j}} \xrightarrow{d} N(0, 1).$$

Notice that $m \to \infty$ also means that the sample size goes to infinity. For more information on survival analysis, see Korosteleva (2009).