POLITECNICO DI TORINO

Dipartimento di Scienze Matematiche G.L. Lagrange

Master's Degree in Mathematical Engineering

Master's Thesis

ADAPTIVE DESIGNS AND BIAS IN TREATMENT EFFECTS ESTIMATION

Supervisor:
Prof. MAURO GASPARINI
Dr. GAËLLE SAINT-HILARY

Candidate:
FULVIO DI STEFANO

Academic year 2019/2020

Contents

Chapter 1

Introduction

Every year, many pharmaceutical companies and research labs spend lots of resources researching new drugs to fight with old and new diseases and illnesses, allowing humanity to progress. Therefore, a large number of new drugs are created and tested to evaluate their efficacy and safety.

The evaluation process of a new drug in humans is well defined. In a first step called phase I, the safety of the drug is mainly assessed, to be sure that this cannot have serious undesirable side effects on people. This study is conducted on a limited number of individuals and only after the drug has passed this first phase it can proceed to the following ones. In phase II, clinical studies aim at identifying the dose of drug that maximizes the effectiveness and the patient population that benefits the most from it, comparing different alternatives and testing the drug on a bigger number of subjects. Once this phase is completed the process moves on to confirmatory phase III, where the drug is tested on a large number of subjects.

In classical Randomized Controlled Trials (RCTs), the evaluation of the effectiveness of a drug is typically performed by giving different doses and a reference treatment or a placebo to a certain population. At the end of the trial, when all individuals have completed the process, the drug effects obtained are analysed. In recent years, Adaptive Designs (ADs) trials have been developed to enhance clinical development. One of the main advantage of this proce-

dure consists in the possibility, at interim analyses during the trial, to stop the evaluation of certain treatments for lack of efficacy and to focus only on the best ones. This results in improvements both in terms of resources and ethics, because it reduces the number of patients receiving non effective treatments. Another main advantage of this procedure is the possibility to select a specific sub-population which benefits the most from the treatment, optimising the resources on the most promising group of people.

This thesis focuses on the problems that arise from the selection process in Adaptive Designs. Because of this selection, the naive Maximum-Likelihood Estimation, main reference in classical Randomized Controlled Trials, is biased. In particular, two types of biases can be identified: the always-reporting bias, a negative bias which affects the estimation of the dropped treatments, and the selection bias, a positive bias which affects the estimation of the selected treatments. In the literature, several methods have been proposed to obtain a better estimation of the treatments' effects in such contexts

In the following, we identify and handle the bias due to the use of these Adaptive Designs. In the next Chapter in particular, Adaptive Designs and related Bias will be described from a statistical point of view. In Chapter 3, the methodology used to address this issue in three different settings is described. In Chapter 4, we conduct an extensive simulation study to assess the properties of the proposed estimators. In Chapter 5, we present two practical applications inspired by real-case studies on Alzheimer's Disease and Heart Failure. Some discussion and concluding remarks are provided in the last Chapter.

Chapter 2

Adaptive Designs

At first, we try to focus on the statistical aspects of Adaptive Designs. Although they are very useful from a practical point of view, allowing more flexibility with respect to classical Clinical Trials, they produce some peculiarities from a statistical point of view. In particular, a naive estimate that does not take into account the selection process, is biased.

2.1 Clinical Trials and Adaptive Designs

Let's suppose we have have discovered a new drug and we want to identify which dose is the most effective in order to carry out new future clinical trials.

If we make use of classical Randomized Controlled Trials (RCT), a certain number of patients is recruited and randomly allocated into a number of arms equal to the doses we want to test. Then, the patients are given the selected dose of drug for a certain period of time and data regarding the efficacy and the safety of the drug are collected. At the end of the trial the data analysis is performed. A naive estimator like the Maximum-Likelihood Estimator (MLE) is a good choice, since in this case it will be *unbiased*. As a matter of fact, it can be proven that the MLE of a normal distribution, which corresponds to the sample mean, is expected to return the distribution's mean.

This process seems quite simple and straightforward, however it is limiting. First, if some of the treatments (or even all of them) are ineffective we have to wait until the end of the trial to discover it, losing resources that could be directed towards the research of new solutions. Secondly, if the drug is developed against some very serious or debilitating disease, it would be better for the patients to receive the best therapy as soon as possible.

For these reasons Adaptive Designs have been developed. They permit to make use of pre-planned modifications that include [12] refining sample size; stopping the whole trial or single doses for lack of efficacy; stopping the whole trial for success; reshuffling patients among treatment arms, selecting population that would be more likely to benefit from the treatment. Therefore, using Adaptive Designs seems a very promising option to follow, but all this flexibility has a cost. It is well known [1, 2, 12] that the selection rule applied in Adaptive Designs causes a *biased estimation*.

Back to our example, suppose that after half of the patients have completed their trial we decide to check the data and to continue the analysis only with the best two treatments. The MLE at the end of the trial is the weighted average of stage 1 and stage 2 MLE, but because of the selection rule this estimation is biased [2]. The key to understand why the estimation is biased is the fact that from a statistical point of view, the selection rule consists in choosing the *maximum* of a group of random variables, which is not distributed like the random variables. In Figure 2.1, retrieved from Pallmann et al. [12], we see that if the low treatment effects are excluded and the higher are not, the final estimation is biased optimistically. Therefore, an estimation that uses stage 1 data and does not take the selection process into account produces biased results.

In the previous example, what we do is a *treatment selection*. However, another kind of AD involves *sub-population selection*. The aim of this type of Trial is to identify the sub-population that will benefit the most from the treatment. Usually, patients are divided in sub-groups according to some biomarker or covariate values. Examples of biomarkers or covariate are: graded scores, size of

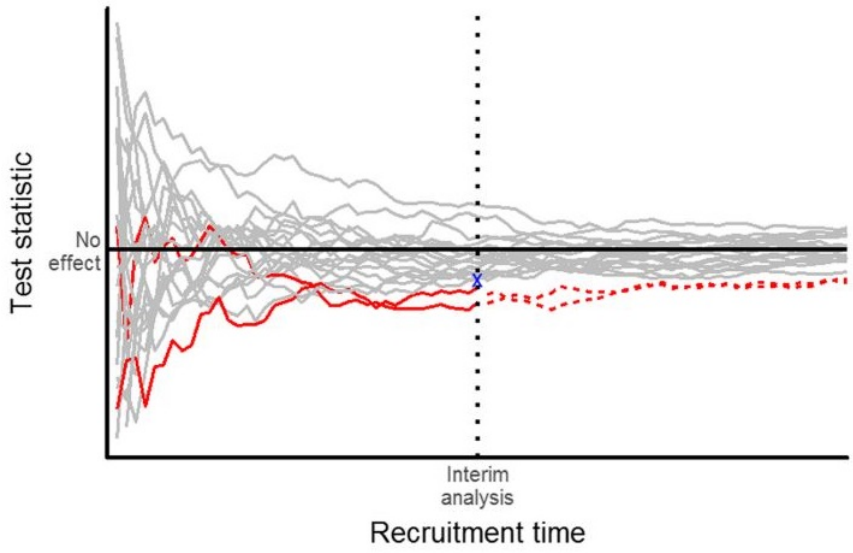

Figure 2.1: *Illustration of bias introduced by early stopping for futility. Retrieved from Pallmann et al. [12]. The red random samples are excluded because of the futility threshold, resulting in optimistic estimation of the treatment effect.*

tumor, baseline heart rate. Using some specific biomarker or covariate, a selection is made in the population, which causes the treatment efficacy's estimate to be biased.

2.2 Identification of Bias in Adaptive Designs

Many works have tried to classify the type of bias that can affect the estimations in Adaptive Designs. Here, we present the classification in the case of treatment selection, but the same principles apply for the sub-population selection. Following the work of Carreras and Brannath [5] and Bauer et al. [2] we may identify two main types of bias:

- The *selection bias*: the bias that affects the estimation of the selected treatments. The chosen treatment response is the maximum of some random variables, which has a higher expected value with respect to the single random variables. Therefore, the estimation is higher than the true

value and this bias is *positive*. However, because the treatment proceeds to the second stage, the final MLE will have lower bias with respect to the stage 1 estimate. As a matter of fact, the stage 2 estimate will be unbiased, being not affected by any selection, and this will help to reduce the selection bias. Bauer et al. [2] proved that a good way to reduce the selection bias is to *reshuffle the sample size*, i.e. redistribute among the selected treatments the sample size planned for the second stage of dropped treatments. This way, because the number of patients in second stage is higher than in first stage, the average of the two estimations will be closer to an unbiased estimation, helping to reduce the selection bias.

- The *always reporting bias*: the bias that affects the estimation of dropped treatments. Because the stage 1 realization of the dropped treatments is lower than the one of the selected treatments, their effect is underestimated. Opposite to the previous case, the always reporting bias has no chance of being reduced because the dropped treatments do not proceed to stage 2.

Bauer et al. [2] proved that fixing the number of treatments to be selected, the selection bias increases while adding treatments to the comparison. Moreover, they show that performing the interim analysis earlier reduce the maximum bias, at the expense of increasing the variability of the selection. On the other hand, Carreras and Brannath [5] prove that the selection bias is maximal when all treatments effects are equal and Cohen and Sackrowitz [6] prove that also in this case the Mean Squared Error is maximal.

2.3 Handling of Bias in Adaptive Designs

After having identified the main sources of bias, we now study some ways to reduce it. Many studies have been conducted trying to handle the bias in treatment effects estimation. In this work, some well-known [12] estimators that attempt to reduce or eradicate the bias have been compared. To make a comparison of these estimators, two main measures are taken into account: Bias and

Mean Squared Error (MSE). Of course, the Bias is the most interesting point, because to reduce or to eliminate it is the primary scope of these estimators. However, it is not the only measure to have to be taken into account. As a matter of fact, the MSE, which is the sum of the variance and the Bias squared, can give important informations on the *variability* of the estimate. A very precise but highly variable estimator may not be preferred with respect to a less precise but more stable alternative.

In the following several approaches are compared:

- *Unbiased Estimators* are developed to find estimations which have no bias. Their main drawback is that they are highly variable. A first unbiased estimator is proposed by Cohen and Sackrowitz [6]. Later, their work has been extended by Bowden and Glimm [3]. A modified version adapted for Adaptive Threshold Enrichment Clinical Trials has been proposed by Kimani et al. [9] and Roberston at al. [14].

- *Shrinkage Estimators* attempt to reduce, but not to eradicate, the bias with low impact on the MSE. Their idea is to shrink the stage 1 estimate towards the overall stage 1 mean, in order to reduce the selection bias. Then, since the stage 2 estimate is unbiased, the overall estimate has lower bias with respect to a naive estimation. We compare two approaches proposed by Carreras and Brannath [5] and Brückner et al. [4].

- *Bias-Adjusted Estimators* are mainly proposed by Whitehead [16] and Stallard and Todd [15]. The main idea of these procedures is to find an estimation of the bias that can be iteratively subtracted from the original naive estimation. The comparison is made by considering both Single-Iteration and Multi-Iteration approaches.

Chapter 3

Methodology

At the end of the previous chapter, several methods that aim to reduce the bias in Adaptive Designs have been introduced. In this chapter, we are going to describe them in greater detail. In particular, treatment selection and subpopulation selection are analysed. Some of the proposed methods have minor differences depending on the analysis. If this were the case, these differences are pointed out.

3.1 Treatment Selection

One of the main reasons why adaptive designs have been developed is to analyse and identify the most effective treatments (or doses) out of a group. In the following, we consider comparisons between different treatments on which a selection of the best one is performed after an interim analysis. We consider N the total number of patients which take part to the trial and N_1 and N_2 those who participate at stage 1 and stage 2 respectively. We are going to compare K different treatments among the N patients. Among these treatments we identify as S the one that is the most effective, according to the stage 1 data.

We can define n_{1i} and n_{2i} for $i = (1, ..., K)$ as the number of patients assigned to each of the treatment arm in stage 1 and stage 2. In the following, we consider that the number of patients assigned to each treatment arm is the same. If we

select only one treatment, $n_{11} = n_{12} = ... = n_{1K} = \frac{N_1}{K}$, $n_{2S} = N_2$ and $n_{2i'} = 0$ $i' = (1, ..., K), i' \neq S$.

The response of the treatment is supposed to have a normal distribution with true treatment effect corresponding to δ_i for $i = (1, ..., K)$ and a common variance to all patients equal to σ^2. If $\bar{X}_{1i}$ for $i = (1, ..., K)$ and $\bar{X}_{2S}$ are the stage 1 and stage 2 sample mean, these are distributed according to a normal distribution $\bar{X}_{1i} \sim N(\delta_i, \tau^2_{1i})$, where $\tau^2_{1i} = \frac{\sigma^2}{n_{1i}}$, and $\bar{X}_{2S} \sim N(\delta_S, \tau^2_{2S})$, where $\tau^2_{2S} = \frac{\sigma^2}{n_{2S}}$.

3.1.1 Naive Estimation

The first estimator we analyse in this work is the Maximum-Likelihood Estimator. In adaptive designs' literature it is well known that this estimator is *biased* [1, 2, 12], while in classical RCTs this is the main reference.

The Maximum-Likelihood Estimator for the case of normal distribution is the sample mean. For the selected treatments we obtain two estimators, the stage 1 and stage 2 sample means, while for the dropped treatments we obtain only the stage 1 sample mean. As mentioned before, the stage 2 sample mean is an unbiased estimation, while stage 1 sample mean are biased because of selection and suffer from either the selection bias and the always reporting bias.

To calculate this estimator we define $t_S = \frac{n_{1S}}{n_{1S}+n_{2S}}$ as the information fraction at the interim analysis. Therefore, for the selected treatments the estimation is:

$$\hat{\delta}_{S,N} = t_S \bar{X}_{1S} + (1 - t_s)\bar{X}_{2S}$$

while for the dropped ones it is simply $\hat{\delta}_{i,N} = \bar{X}_{1i}$.

In the following, we are going to compare the performance of this estimation with the others'. We will see estimators that eradicate the bias but have a high variance, estimators that reduce the bias with a small effect on variance and estimators that estimate the bias of the MLE to subtract it from the MLE.

3.1.2 Uniformly Minimum Variance Conditionally Unbiased Estimator

We explore the first kind of estimator that handles the bias in treatment effects estimation: Uniformly Minimum Variance Conditionally Unbiased Estimator, or UMVCUE. This type of estimator, as the name suggests, is *unbiased*. It is based on a process of Rao-Blackwellization that actually eradicated the bias. However, it has the drawback of being highly variable, which results in high value of MSE.

Following the work of Cohen and Sackrowitz [6] and Bowden and Glimm [3] for the case of treatment selection, we use the following notations: we index by (1) the treatment for which the stage 1 sample mean $\bar{X}_{1(1)}$ is the maximum of $\bar{\mathbf{X}}_1 = (\bar{X}_{11}, ..., \bar{X}_{1K})$, $\delta_{(1)}$ its true mean and $\bar{X}_{2(1)}$ its stage 2 sample mean; we index by (2) the treatment for which the stage 1 mean $\bar{X}_{1(2)}$ is the second-largest element of $\bar{\mathbf{X}}_1$, $\delta_{(2)}$ and $\bar{X}_{2(2)}$ its true mean and stage 2 sample mean; and so on until $\bar{X}_{1(K)}$, $\delta_{(K)}$ and $\bar{X}_{2(K)}$ correspond to the minimum of $\bar{\mathbf{X}}_1$. We define the Uniformly Minimum Variance Conditionally Unbiased Estimator to be:

$$\hat{\delta}_{(j),UMVCUE} = \hat{\delta}_{(j),N} - \frac{\tau^2_{2(j)}}{\sqrt{\tau^2_{1(j)} + \tau^2_{2(j)}}} \frac{\phi(W_{j,j+1}) - \phi(W_{j,j-1})}{\Phi(W_{j,j+1}) - \Phi(W_{j,j-1})}$$

Where ϕ and Φ are the probability density function and the cumulative density function of the normal distribution and $W_{j,p} = \frac{\sqrt{\tau^2_{1(j)} + \tau^2_{2(j)}}}{\tau^2_{1(j)}} \left(\hat{\delta}_{(j),N} - \bar{X}_{1(p)}\right)$. We set conventionally that $\bar{X}_{1(0)} := +\infty$ and $\bar{X}_{1(K+1)} := -\infty$. Note that by our notation $\tau^2_{1(j)}$ is the variance of the j-th largest element of $\bar{\mathbf{X}}_1$ and $\tau^2_{2(j)}$ the stage 2 one for the corresponding element.

3.1.3 Shrinkage Estimators

The second type of estimators studied in this work are Shrinkage Estimators. The aim of this type of estimators is to have a reduction in the bias, with re-

spect to a naive estimation, not increasing the variance. The most important aspect of these estimator is that they *shrink* the stage 1 estimations towards their overall mean.

We recall from the previous chapter that while the stage 2 means provide an unbiased estimation of the effectiveness of the treatments, stage 1 estimations are biased because of selection. Since the selection bias which affects the selected treatment is positive and the always-reporting bias which affects the dropped treatments is negative, the idea behind these estimators is that shrinking the estimations towards the overall mean will reduce both.

Shrinkage Estimator following the work of Carreras and Brannath

The first shrinkage estimator we consider is the one by Carreras and Brannath [5]. We add the following notations: we define $\bar{X}_{1\cdot} = \frac{1}{K}\sum_{i=1}^{K}\bar{X}_{1i}$ as the overall stage 1 mean. The shrinkage estimator is calculated as:

$$\hat{\delta}_{i,S1} = t_s[\hat{C}_+\bar{X}_{1i} + (1-\hat{C}_+)\bar{X}_{1\cdot}] + (1-t_s)\bar{X}_{2i}$$

if i is selected, or $\hat{\delta}_{i,S1} = [\hat{C}_+\bar{X}_{1i} + (1-\hat{C}_+)\bar{X}_{1\cdot}]$ otherwise, with $\hat{C}_+$ defined as follows.

If $K \geq 4$:

$$\hat{C}_+ = max(0,\hat{C}), \qquad \hat{C} = 1 - \frac{(K-3)\sigma^2}{n\sum_{j=1}^{K}\left(\bar{X}_{1j} - \bar{X}_{1\cdot}\right)^2}$$

While if $K = 2,3$:

$$\hat{C}_+ = max(0,\hat{C}), \qquad \hat{C} = 1 - \frac{(K-1)\sigma^2}{n\sum_{j=1}^{K}\left(\bar{X}_{1j} - \bar{X}_{1\cdot}\right)^2}$$

Shrinkage Estimator following the work of Brünckner et al.

The second shrinkage estimator we analyse, which is derived in a Bayesian framework, is the one by Brückner et al. [4]. We suppose to have a prior distribution of the vector $\boldsymbol{\delta}$, which is a Multivariate Normal $MVN(\boldsymbol{\mu}, v^2\mathbf{I}_K)$, that is updated with the data $\bar{\boldsymbol{X}}_1 \sim MVN(\boldsymbol{\delta}, \boldsymbol{\Sigma}_X)$, to get a posterior estimation for $\boldsymbol{\delta}$. This estimation is $\mathbf{C}\bar{\boldsymbol{X}}_1 + (\mathbf{I}_K - \mathbf{C})\boldsymbol{\mu}$, where $\mathbf{C} = \mathbf{I}_K - \boldsymbol{\Sigma}_X(v^2\mathbf{I}_K + \boldsymbol{\Sigma}_X)^{-1}$.

The $\boldsymbol{\Sigma}_X$ is a diagonal matrix containing the τ^2_{1i} on the diagonal. We define the prior mean $\boldsymbol{\mu}$ as a vector of length K containing $\bar{X}_{1\cdot}$, the overall stage 1 mean. The core of the method is to derive a sensible value for v^2 using an iterative approach. Defining $\mathbf{D}$ to be the diagonal matrix with the eigenvalues of $\boldsymbol{\Sigma}_X$ (in this case $\mathbf{D} = \boldsymbol{\Sigma}_X$, but it is not the case in sub-population selection) and $\hat{v}^2$ as an initial guess, we iteratively proceed:

- Step 1: Define weights $w_i = (\hat{v}^2 + D^2_{ii})^{-1}$ for $i = (1, ..., K)$.
- Step 2: Update the estimate calculating $\hat{v}^2 = \frac{\sum_{i=1}^K w_i[(\bar{X}_{1i} - \bar{X}_{1\cdot}) - D^2_{ii}]}{\sum_{i=1}^K w_i}$.
- Go back to step 1 using the updated $\hat{v}^2$.

When the iterative approach converges, we get a solution $\widetilde{v}^2$. Since this solution might be negative, $\widetilde{v}^2_+ = max(0, \widetilde{v}^2)$ is used to calculate $\widetilde{\boldsymbol{C}}_+ = \mathbf{I}_K - \boldsymbol{\Sigma}_X(\widetilde{v}^2_+\mathbf{I} + \boldsymbol{\Sigma}_X)^{-1}$. Then the stage 1 estimator is:

$$\hat{\boldsymbol{\delta}}^{Stage1}_{S2} = \widetilde{\boldsymbol{C}}_+\bar{\boldsymbol{X}}_1 + (\mathbf{I}_K - \widetilde{\boldsymbol{C}}_+)\bar{X}_{1\cdot}$$

This is the estimator of $\hat{\delta}_{i,S2}$ if i is not selected. For the selected treatments instead:

$$\hat{\delta}_{i,S2} = t_s\hat{\delta}^{Stage1}_{i,S2} + (1 - t_s)\bar{X}_{2i}$$

3.1.4 Bias-Adjusted Estimator

The last type of estimator we compare in this work is Bias-Adjusted Estimator. This estimator follows the work of Whitehead [16] and Stallard and Todd [15] and the main idea behind this procedure is to estimate the bias of the current estimator and subtract it from the estimation. In this work, we compare two approaches: the single-iteration estimator and the multi-iteration estimator. The single-iteration estimator is calculated as follows:

$$\hat{\delta}_{i,SI} = \hat{\delta}_{i,N} - \hat{b}_i(\hat{\boldsymbol{\delta}}_N)$$

Where $\hat{b}_i(\hat{\boldsymbol{\delta}}_N)$ is an estimation of the bias of the Naive Estimator.

In the multi-iteration approach, we would continue estimating the bias of the new estimation and subtracting it, iteratively, until convergence.

Since the single-iteration approach is just a special case of the multi-iteration, in the following we consider a generic iteration with estimation $\widetilde{\boldsymbol{\delta}}$. In the case of treatment selection, it can be shown that the bias of the selected treatments is:

$$b_S(\widetilde{\boldsymbol{\delta}}) = t_s(E[\bar{X}_{1S}|S] - \widetilde{\delta}_S)$$

Where the expectation is conditioned on the fact that S is selected. For the other treatments:

$$b_i(\widetilde{\boldsymbol{\delta}}) = (E[\bar{X}_{1i}|S] - \widetilde{\delta}_i)$$

To estimate the value of $\mathbf{b}(\widetilde{\boldsymbol{\delta}})$, the vector of $b_i(\widetilde{\boldsymbol{\delta}})$ for $i = (1,...,K)$, we need to estimate the single $E[\bar{X}_{1i}|S]$. This is computationally expensive, because

it involves multiple-integration. However, given the treatment arms are assumed independent, using the probability density function or cumulative density function of the Normal distribution permits to simplify the calculations, allowing to compute only a single integral. In the following, an example is made to better understand how this calculation can be performed.

Example

We consider a case of adaptive design with treatment selection where $K = 3$ and $S = 1$, meaning that we have three treatments and the first is the selected. $\widetilde{\boldsymbol{\delta}}$ is our current estimate, while $\boldsymbol{\tau_1}$ is assumed to be known. We can calculate the probability of selecting the first treatment as follows:

$$P(S = x_{11}) = \int_{-\infty}^{+\infty} \int_{-\infty}^{x_{11}} \int_{-\infty}^{x_{11}} f(x_{11}, x_{12}, x_{13})\, dx_{13}\, dx_{12}\, dx_{11} =$$

$$\int_{-\infty}^{+\infty} \phi\left(\frac{x_{11} - \widetilde{\delta}_1}{\tau_{11}}\right) \Phi\left(\frac{x_{11} - \widetilde{\delta}_2}{\tau_{12}}\right) \Phi\left(\frac{x_{11} - \widetilde{\delta}_3}{\tau_{13}}\right) dx_{11}$$

Where f is a multivariate normal density with mean $\widetilde{\boldsymbol{\delta}}$ and variance-covariance matrix $\boldsymbol{\Sigma}_X = \boldsymbol{\tau_1}\mathbf{I}_K$, ϕ the probability density function of a standard normal distribution and Φ its cumulative density function.

The main concept of this integration is that while the integral over x_{11} is from $-\infty$ to $+\infty$ because it can take any value, the other two treatments must have their effect lower than x_{11}'s, for this to be selected. Moreover, since the three random variables are independent, the integral over x_{12} and x_{13} can be transformed using the cumulative density function of the normal to simplify the calculations.

Now it is straightforward to calculate the expected value of x_{11}:

$$E(x_{11}|S=x_{11}) = \frac{1}{P(S=x_{11})} \int_{-\infty}^{+\infty} \int_{-\infty}^{x_{11}} \int_{-\infty}^{x_{11}} x_{11} f(x_{11}, x_{12}, x_{13})\, dx_{13}\, dx_{12}\, dx_{11} =$$

$$\frac{1}{P(S=x_{11})} \int_{-\infty}^{+\infty} x_{11} \phi\left(\frac{x_{11}-\widetilde{\delta}_1}{\tau_{11}}\right) \Phi\left(\frac{x_{11}-\widetilde{\delta}_2}{\tau_{12}}\right) \Phi\left(\frac{x_{11}-\widetilde{\delta}_3}{\tau_{13}}\right) dx_{11}$$

It is also possible to calculate the expected value of one of the other treatments, for example x_{12}:

$$E(x_{12}|S=x_{11}) = \frac{1}{P(S=x_{11})} \int_{-\infty}^{+\infty} \int_{-\infty}^{x_{11}} \int_{-\infty}^{x_{11}} x_{12} f(x_{11}, x_{12}, x_{13})\, dx_{13}\, dx_{12}\, dx_{11} =$$

$$\frac{1}{P(S=x_{11})} \int_{-\infty}^{+\infty} \phi\left(\frac{x_{11}-\widetilde{\delta}_1}{\tau_{11}}\right) \left(\int_{-\infty}^{x_{11}} x_{12} \phi\left(\frac{x_{12}-\widetilde{\delta}_2}{\tau_{12}}\right) dx_{12} \right) \Phi\left(\frac{x_{11}-\widetilde{\delta}_3}{\tau_{13}}\right) dx_{11}$$

After these calculations are completed, the expected values estimated in this procedure can be used in the estimation of the bias to calculate the estimator.

3.2 Sub-Population Selection

Some small changes arise when dealing with sub-population selection. The main difference is that the sub-populations are not independent anymore.

Following the work of Kimani et al. [10], we divide the population into K partitions according to some biomarker value. In particular, we define K threshold values such that the patients inside a partition have a biomarker value between the upper and lower thresholds. Instead, the sub-population consists in the patients below the corresponding threshold. This way, sub-population K

will correspond to the total population; sub-population $K-1$ will be a portion of population K; and so on. Using the notation S_i for $i = (1,...,K)$ to indicate the sub-population, we notice that $S_1 \subseteq S_2 \subseteq ... \subseteq S_{K-1} \subseteq S_K$.

We suppose that the N_1 patients are split between the K partitions, with n_{1i} for $i = (1,...,K)$ the number of patients in each partition in stage 1, which are considered equal $n_{11} = n_{12} = ... = n_{1K} = \frac{N_1}{K}$. In each of these partitions, half of the patients are given a control and the other half are given the treatment. As in the previous case, the patients' outcome is considered normal with mean difference equal to $\delta_1 \geq \delta_2 \geq ... \geq \delta_K$ with a common variance σ^2. This means that the partition with smaller biomarker value is more affected by the treatment. The stage 1 sample mean will therefore be normally distributed $\bar{X}_{1i} \sim N(\delta_i, \tau_{1i}^2)$, where $\tau_{1i}^2 = \frac{4\sigma^2}{n_{1i}}$. When sub-population S is selected it means that only the first S partitions pass to the second stage. In this case $n_{21} = n_{22} = ... = n_{2S} = \frac{N_2}{S}$ and $n_{2i'} = 0$ for $i' = (S+1,...,K)$. Also stage 2 sample means will be distributed like $\bar{X}_{2i} \sim N(\delta_i, \tau_{2i}^2)$, where $\tau_{2i}^2 = \frac{4\sigma^2}{n_{2i}}$.

However, in the case of sub-population selection we are not interested in the effect that the drug has on the partition, but on the sub-population. Therefore, it is needed to have estimations about the effect on sub-population. That can be done in two different ways: estimating directly the effect on the sub-populations, or calculating it from the effect on partitions. Defining $p_i = \frac{n_{1i}}{N_1}$ and $\widetilde{p}_i = \sum_{i'=1}^{i} p_{i'}$ for $i = (1,...,K)$, we can create a matrix that links the effects on the partition δ_i to the effect on the sub-population θ_i. This matrix $\mathbf{B}$ is a KxK matrix where $B_{ij} = 0$ if $i < j$ and $B_{ij} = \frac{p_j}{\widetilde{p}_i}$ if $i \geq j$. The vector of effects on sub-population $\boldsymbol{\theta}$ is equal to $\boldsymbol{\theta} = \mathbf{B}\boldsymbol{\delta}$. Analogously, we can define the number of patients in each sub-population as $m_{1i} = \sum_{i'=1}^{i} n_{1i'}$ for $i = (1,...,K)$, $m_{2i} = \sum_{i'=1}^{i} n_{2i'}$ for $i = (1,...,S)$. Now, we see that $\bar{\boldsymbol{Y}}_1 = \mathbf{B}\bar{\boldsymbol{X}}_1$ and $\bar{\boldsymbol{Y}}_2 = \mathbf{B}_{|SxS}\bar{\boldsymbol{X}}_2$ (where $\mathbf{B}_{|SxS}$ indicates the first SxS elements of the $\mathbf{B}$ matrix) are the stage 1 and stage 2 sub-populations' means, which are also normally distributed: $\bar{Y}_{1i} \sim N(\theta_i, \sigma_{1i}^2)$, where $\sigma_{1i}^2 = \frac{4\sigma^2}{m_{1i}}$; $\bar{Y}_{2i} \sim N(\theta_i, \sigma_{2i}^2)$, where $\sigma_{2i}^2 = \frac{4\sigma^2}{m_{2i}}$.

The adaptation at the interim analysis is as follows: if all the sub-populations at the interim analysis have a mean difference lower than a futility threshold b

($\bar{Y}_{1i} \leq b \ \forall i \in (1,...,K)$), the trial is stopped; instead, if the trial is not stopped, the biggest sub-population with a mean difference greater than b is selected.

3.2.1 Naive Estimation

In this case, the naive estimation is very similar to the previous case: it is sufficient to calculate the $\hat{\delta}_{i,N}$ $i = (1,...,S)$ as in the treatment selection case and then use the matrix $\mathbf{B}$ to calculate $\hat{\boldsymbol{\theta}}_N = \mathbf{B}\hat{\boldsymbol{\delta}}_N$.

3.2.2 UMVCUE following the Work of Roberston et al. and Kimani et al.

In this setting, the estimator has to account for the stop for futility. We can proceed in two ways: estimating directly θ_S or estimating the δ_i for $i = (1,...,S)$ and then use the matrix $\mathbf{B}$ to find an estimation for θ_S.

The estimator for θ_S has been given by Robertson et al. [14]

$$\hat{\theta}_{S,UMVCUE} = \hat{\theta}_{S,N} - \frac{\sigma^2_{2S}}{\sqrt{\sigma^2_{1S} + \sigma^2_{2S}}} \frac{\phi(g(b)) - \phi(g(U))}{\Phi(g(b)) - \Phi(g(U))}$$

Where $g(b) = \frac{\sqrt{\sigma^2_{1S}+\sigma^2_{2S}}}{\sigma^2_{1S}} \left(\hat{\theta}_{S,N} - b\right)$, $g(U) = \frac{\sqrt{\sigma^2_{1S}+\sigma^2_{2S}}}{\sigma^2_{1S}} \left(\hat{\theta}_{S,N} - U\right)$ and $U = min\left\{ \frac{\tilde{p}_{S+1}b - p_{S+1}\bar{X}_{1,S+1}}{\tilde{p}_S}, \frac{\tilde{p}_{S+2}b - \sum_{i=S+1}^{S+2} p_i \bar{X}_{1i}}{\tilde{p}_S}, ..., \frac{\tilde{p}_K b - \sum_{i=S+1}^{K} p_i \bar{X}_{1i}}{\tilde{p}_S} \right\}$ or $+\infty$ when $S = K$.

The estimator for δ_i for $i = (1,...,S)$ is given by Kimani et al. [9]:

$$\hat{\delta}_{i,KIMANI} = \hat{\delta}_{i,N} - \frac{\tau^2_{2i}}{\sqrt{\tau^2_{1i} + \tau^2_{2i}}} \frac{\phi(g'(V_i)) - \phi(g'(Q_i))}{\Phi(g'(V_i)) - \Phi(g'(Q_i))}$$

Where we set $g'(V_i) = \frac{\sqrt{\tau_{1i}^2+\tau_{2i}^2}}{\tau_{1i}^2}\left(\hat{\delta}_{i,N} - V_i\right)$, $g'(Q_i) = \frac{\sqrt{\tau_{1i}^2+\tau_{2i}^2}}{\tau_{1i}^2}\left(\hat{\delta}_{i,N} - Q_i\right)$,

$V_i = \frac{1}{p_i}\left(\widetilde{p}_S b - \sum_{j=1, j\neq i}^{S} p_j \bar{X}_{1j}\right)$. Q_i is equal to $+\infty$ if $S = K$ or in any other case

$Q_i = min\left\{\frac{\widetilde{p}_{S+1} b - \sum_{j=1, j\neq i}^{S+1} p_j \bar{X}_{1j}}{p_i}, \frac{\widetilde{p}_{S+2} b - \sum_{j=1, j\neq i}^{S+2} p_j \bar{X}_{1j}}{p_i}, ..., \frac{\widetilde{p}_K b - \sum_{j=1, j\neq i}^{K} p_j \bar{X}_{1j}}{p_i}\right\}$.

Eventually, $\hat{\theta}_{S,KIMANI}$ is the S-th element of the vector $\mathbf{B}_{|SxS}\,\hat{\boldsymbol{\delta}}_{KIMANI}$.

3.2.3 Shrinkage Estimators

In this case, the Shrinkage Estimator of Carreras and Brannath can be calculated as in the treatment selection case and then $\hat{\boldsymbol{\theta}}_{S1}$ can be estimated via the matrix $\mathbf{B}$.

Also the procedure of Brünckner et al. is analogous. The only differences are that $\boldsymbol{\delta}$ is substituted by $\boldsymbol{\theta}$, which we need to estimate, and $\bar{\boldsymbol{X}}_1$ and $\bar{\boldsymbol{X}}_2$ are substituted by $\bar{\boldsymbol{Y}}_1$ and $\bar{\boldsymbol{Y}}_2$.

3.2.4 Bias-Adjusted Estimator

In the case of sub-population selection, since we are selecting a sub-population and not a partition, the bias at a general iteration is shown to be equal to:

$$b_i(\widetilde{\boldsymbol{\theta}}) = t_S(E[\bar{Y}_{1i}|S] - \widetilde{\theta}_i) \qquad i = (1, ..., S)$$

$$b_i(\widetilde{\boldsymbol{\theta}}) = (E[\bar{Y}_{1i}|S] - \widetilde{\theta}_i) \qquad i = (S+1, ..., K),$$

Now we have to calculate the $E[\bar{Y}_{1i}|S]$ for $i \in (1, ..., K)$ to estimate this bias. To better understand the procedure we make an example.

Example

Consider the case where $K = 3$ sub-populations are compared and where sub-population $S = 2$ is selected. We can calculate the probability of selecting that sub-population as follows:

$$P(S = y_{12}) = \int_{-\infty}^{+\infty}\int_{b}^{+\infty}\int_{-\infty}^{b} f(y_{11}, y_{12}, y_{13})\, dy_{13}\, dy_{12}\, dy_{11}$$

Where f is the density of a multivariate normal distribution with mean $\tilde{\boldsymbol{\theta}}$ and variance-covariance matrix $\boldsymbol{\Sigma}_Y = \mathbf{B}\boldsymbol{\Sigma}_X\boldsymbol{B}^T$, where $\boldsymbol{\Sigma}_X = \boldsymbol{\tau_1}\mathbf{I}_K$ and $\mathbf{B}$ are known. In this case since $S = 2$, y_{11} can take any value and its integral is from $-\infty$ to $+\infty$. On the other hand, y_{12} is greater than b since it is selected, while y_{13} is lower than b since it is not selected. Analogous arguments can be made while calculating the expected values:

$$E(y_{12}|S = y_{12}) = \frac{1}{P(S = y_{12})}\int_{-\infty}^{+\infty}\int_{b}^{+\infty}\int_{-\infty}^{b} y_{12} f(y_{11}, y_{12}, y_{13})\, dy_{13}\, dy_{12}\, dy_{11}$$

$$E(y_{13}|S = y_{12}) = \frac{1}{P(S = y_{12})}\int_{-\infty}^{+\infty}\int_{b}^{+\infty}\int_{-\infty}^{b} y_{13} f(y_{11}, y_{12}, y_{13})\, dy_{13}\, dy_{12}\, dy_{11}$$

More details about the calculations can be found in the supplementary material of Kimani et al. [10].

3.3 Sub-Population Selection with Time to Event Data

Now we investigate the setting of sub-population selection with time to event data. This setting is very common in oncological or cardiovascular illnesses

treatment. In this case, we follow the work of Kimani et al. [11] on time to event data.

The idea is, like in the previous case, to split the data according to some biomarker value and analyse the different sub-populations separately. For simplicity, we consider the different partitions coincident with the sub-populations: we call P_i for $i = (1, ..., K)$ the sub-population/partition and we notice that $P_i \cap P_j = \emptyset \ \forall (i \neq j)$. We again define K threshold values such that the patients inside a partition have a biomarker value between the upper and lower thresholds and n_{1i} and n_{2i} for $i = (1, ..., K)$ the number of patients in each partition at stage 1 and stage 2 respectively. Note that these findings could be easily extended to the case presented in the previous section, where the largest population is selected.

In each partition, there are some patients who have been given the placebo and some patients who have been given the treatment. The measure we want to estimate is the log hazard ratio (LHR), defined as $\delta_i = ln\left(\frac{h_{ti}(t)}{h_{ci}(t)}\right)$ for $i = (1, ..., K)$, with $h_{ti}(t)$ and $h_{ci}(t)$ are the hazard functions of the treatment and the control in group i, respectively. This estimation is done via a Cox proportional hazard model. We consider here that a negative value of the LHR corresponds to a reduction of risk of event with the treatment, i.e. that the treatment is effective with respect to the placebo; if the LHR in one group is lower than in the other, it means the treatment is more effective in that sub-population. The LHR is usually assumed to be normally distributed, with stage 1 and stage 2 estimators $\hat{\delta}_{1i} \sim N(\delta_i, \tau^2_{1i})$ and $\hat{\delta}_{2i} \sim N(\delta_i, \tau^2_{2i})$ for $i = (1, ..., K)$; also $\hat{\tau}^2_{1i}$ and $\hat{\tau}^2_{2i}$ are estimated from the Cox model.

In this setting, there is one main aspect that needs to be pointed out: at the interim analysis, some stage 1 patients may not have had the event of interest yet. If we continue the analysis carrying these patients to stage 2, stage 1 and stage 2 data will be correlated, inducing some bias in the estimation. In many analyses, this correlation is considered as negligible, the stages are considered independent and patients are followed until the end of the trial. To avoid any correlation, patients from stage 1 would have to stop the study at the interim

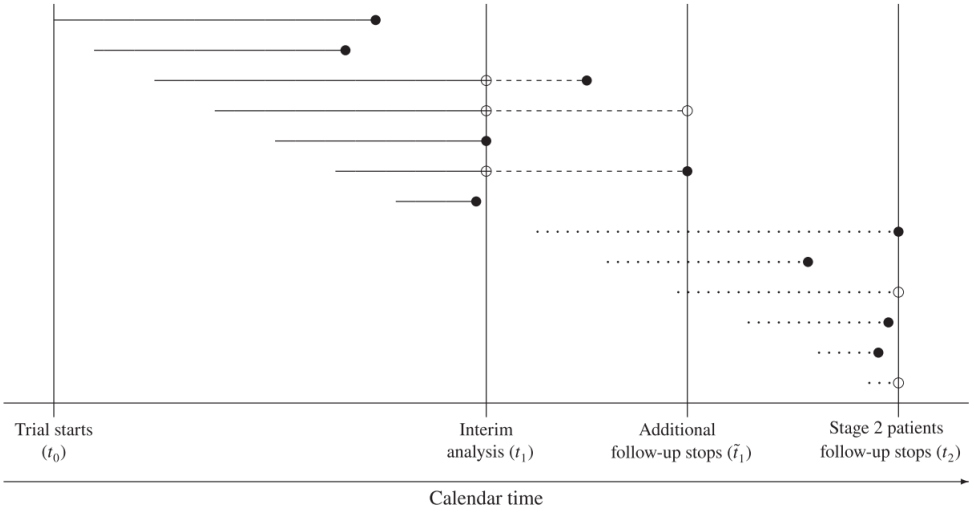

Figure 3.1: *The follow-up of the patients. Retrived from Kimani et al. [11]. The black dots represent patients who had the event, the white dots those who have not. At the interim analysis, some stage 1 patients have had the event, while others have not. Those last ones are followed up to* $\tilde{t}_1$*. Stage 2 patients, recruited after the interim analysis, are followed until the end of the trial.*

analysis. However, this is obviously not ethical and not applicable in practice. Instead, we use an intermediate rule. Consider t_1 the time of the interim analysis and t_2 the end time of the trial, usually defined when a certain number of patients had the event. We define $\tilde{t}_1$ (such that $t_1 \leq \tilde{t}_1 \leq t_2$) the time until which the stage 1 patients are followed. This way, we avoid much of the correlation and we can obtain more accurate estimations. In Figure 3.1 we have a graphical representation of this rule, retrieved from Kimani et al. [11].

Also in this case we specify a selection rule: given a threshold value b, each sub-population P_i $i \in (1,...,K)$ does not continue to stage 2 if its stage 1 estimation is greater than b ($\hat{\delta}_{1i} \geq b$); if all sub-populations satisfy this condition the trial is stopped for futility. We identify as $\mathscr{S}$ the set of selected sub-populations.

3.3.1 Naive Estimation

The naive estimation, which do not take into account the selection process, can be retrieved from the estimators of Cox proportional hazard model:

$$\hat{\delta}_{S,N} = \frac{\hat{\tau}_{1S}^2 \hat{\delta}_{1S} + \hat{\tau}_{2S}^2 \hat{\delta}_{2S}}{\hat{\tau}_{1S}^2 + \hat{\tau}_{2S}^2} \qquad \forall S \in \mathscr{S}$$

while it is simply $\hat{\delta}_{i,N} = \hat{\delta}_{1i}\ \forall i \notin \mathscr{S}$.

This estimation, like in the previous cases, is biased because of selection process. However, there is also a component of the bias which comes from the use of stage 1 patients also in stage 2, which induces a correlation between the stages. In the following, we focus only on handling the selection bias while the correlation bias is left.

3.3.2 UMVCUE following the Work of Kimani et al.

Following the work of Kimani et al. [11] we calculate an estimator which handles the selection bias. However, because of the correlation bias, this estimator would not be perfectly unbiased.

$$\hat{\delta}_{S,UMVCUE} = \hat{\delta}_{S,N} - \frac{\hat{\tau}_{2S}^2}{\sqrt{\hat{\tau}_{1S}^2 + \hat{\tau}_{2S}^2}} \frac{\phi(g(L)) - \phi(g(b))}{\Phi(g(L)) - \Phi(g(b))}$$

Where $g(x) = \frac{\sqrt{\hat{\tau}_{1S}^2 + \hat{\tau}_{2S}^2}}{\hat{\tau}_{1S}^2} \left(\hat{\delta}_{S,N} - x\right)$ and $L = -\infty$.

3.3.3 Shrinkage Estimators

In this case, the Shrinkage Estimators of Carreras and Brannath and Brünckner et al. can be calculated as in the treatment selection case, using the estimations

from Cox proportional hazard model. The LHRs will be shrinked towards the overall mean LHR.

3.3.4 Bias-Adjusted Estimator

The procedure for the bias-adjusted estimator is similar to the previous cases: we iteratively subtract an estimation of the bias from the estimation. Also in this case, we compare both single-iteration and multiple-iteration approaches. The bias at a general iteration is equal to:

$$b_i(\widetilde{\boldsymbol{\delta}}) = t_i(E[\hat{\delta}_{1i}|S] - \widetilde{\delta}_i) \qquad i \in \mathscr{S}$$

$$b_i(\widetilde{\boldsymbol{\delta}}) = (E[\hat{\delta}_{1i}|S] - \widetilde{\delta}_i) \qquad i \notin \mathscr{S}$$

Now we have to calculate the $E[\hat{\delta}_{1i}|S]$ $i \in (1,...,K)$ to estimate this bias. Since in this setting we consider independent sub-populations, the calculation is as follows:

$$E[\hat{\delta}_{1i}|S] = \int_{-\infty}^{b} x\, \phi\left(\frac{x-\widetilde{\delta}_i}{\hat{\tau}_{1i}}\right) dx \qquad i \in \mathscr{S}$$

$$E[\hat{\delta}_{1i}|S] = \int_{b}^{\infty} x\, \phi\left(\frac{x-\widetilde{\delta}_i}{\hat{\tau}_{1i}}\right) dx \qquad i \notin \mathscr{S}$$

where ϕ is the probability density function of a normal distribution. As a reminder, in this formula, a sub-population is selected if the treatment has a LHR lower than b, while it is dropped if not.

Chapter 4

Simulations

In this Chapter, we are comparing via simulations the performances of the estimators presented in the previous chapter. The results are compared in three different settings: treatment selection with normally distributed endpoint, sub-population selection with normally distributed endpoint, sub-population selection with time to event endpoint.

We compare the various estimators in terms of Bias, Variance and Mean Squared Error (MSE). The bias is our main interest: the objective of this report is to identify and handle bias. However, another crucial point is the variance: an Estimator which has low Bias but high Variance may not be informative enough on the treatment effect to make a decision, and may not be preferred to one with higher Bias but low Variance. The sum of the Variance and the Bias Squared is the MSE, which summarises up the information previously obtained.

4.1 Treatment selection with normally distributed endpoint

In the Treatment Selection case, to evaluate the performance of the estimators, we consider a base-case adaptive design setting in which we make vary some characteristics to assess their impact on the results.

Following the methodology in section 3.1, we consider a clinical trial in

which we compare $K = 4$ treatments (e.g. doses) to a control, with a primary endpoint normally distributed and a common standard deviation of $\sigma = 5.4$. One interim analysis is conducted during the trial at which all treatments are compared to the control, and only the best treatment and the control arms continue in stage 2. In the control and the selected arm, $N = 100$ patients per arm will be included. We analyse the performances of the estimators when the interim analysis is made, with three different scenarios: information fraction of $\frac{1}{3}$, $\frac{1}{2}$ and $\frac{2}{3}$, i.e. when $N_1 = 33$, $N_1 = 50$ and $N_1 = 66$ are included in each treatment arm in stage 1, respectively. We also consider three scenarios regarding the treatment effect: all treatments are ineffective $\boldsymbol{\delta} = (0,0,0,0)$; only one treatment is effective $\boldsymbol{\delta} = (0,0,0,3)$; treatments are linearly increasing in effectiveness $\boldsymbol{\delta} = (1,2,3,4)$. In all scenarios, the control group has no effect $\delta_c = 0$. In each scenario we run 50000 simulations and the results (bias, variance and MSE) are given in units of approximate standard errors $SE = \sqrt{\frac{\sigma^2}{N}}$.

In Figure 4.1 the results of the simulations are shown:

- The Naive Estimation (N) has the highest Bias but low Variance, resulting in overall average MSE.
- The Unbiased Estimator (UMVCUE) has zero Bias but has high Variance, resulting in high MSE.
- The Single-Iteration Bias-Adjusted Estimator (SI) has low positive Bias and slightly more Variance with respect to the MLE, resulting in an overall lower MSE with respect to MLE.
- The Multiple-Iteration Bias-Adjusted Estimator (MI) has high negative bias (almost the same magnitude as the MLE) and high Variance, resulting in high MSE.
- The Shrinkage Estimator of Carreras and Brannath [5] (S1) has a positive Bias in the case of no effective treatments and linear effectiveness, but this is lower in magnitude with respect to MLE. In the case of only one effective treatment it has negative Bias. This may be due to the fact that

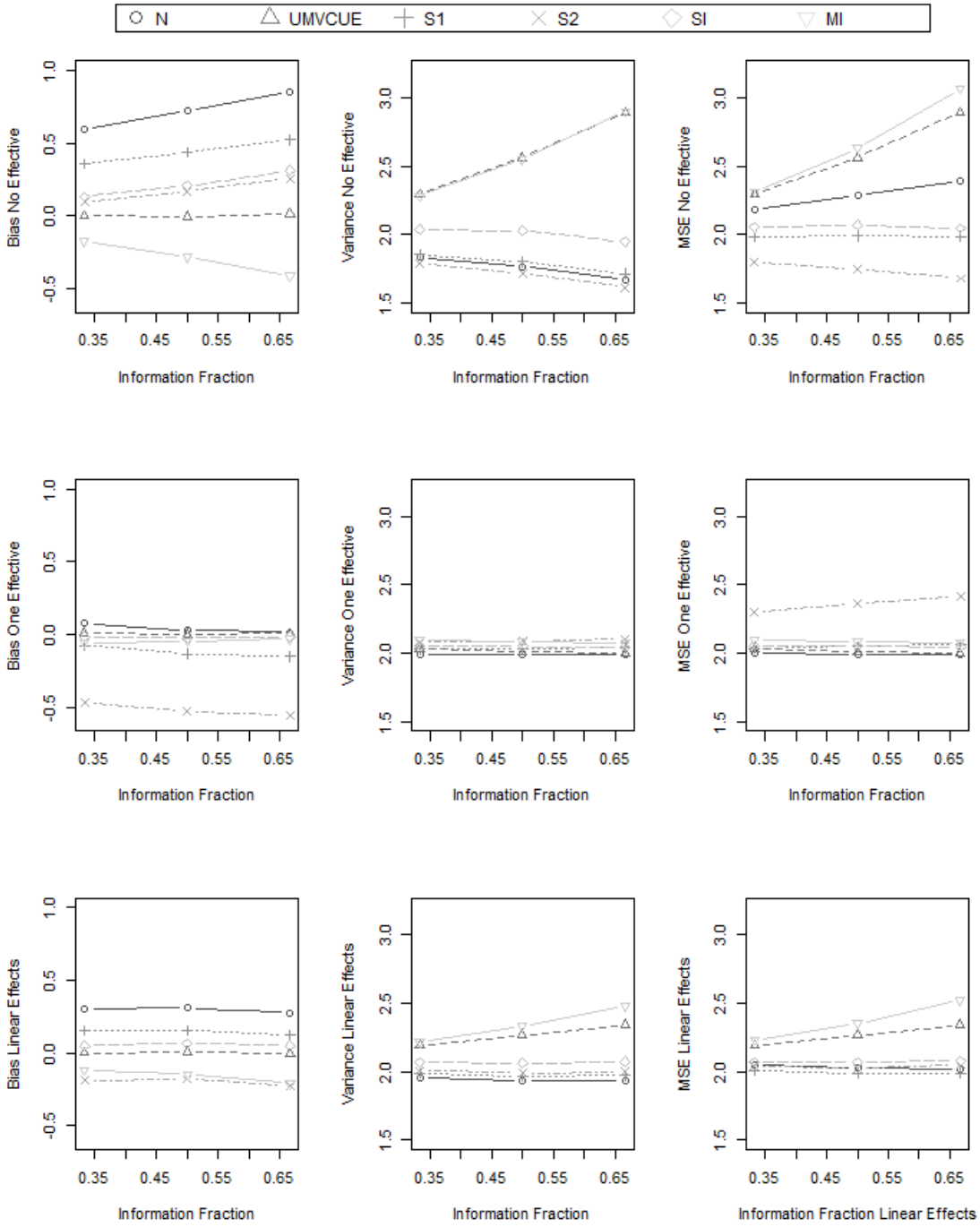

Figure 4.1: *Estimators' Performance in Treatment Selection with Normally Distributed Endpoints. Top Row: No Effective Treatment; Middle Row: One Effective Treatment; Bottom Row: Linear Effects of Treatments. Left Column: Bias; Centre Column: Variance; Right Column: Mean Squared Error.*

it tends to Shrink towards the overall mean, which would be small in this case (driven by the non-effective treatments). It is one of the best performing in terms of Variance, resulting in low MSE too.

- The Shrinkage Estimator of Brunckner et al. [4] (S2) has low positive Bias in the no effective treatment scenario, but negative bias in the linear effectiveness scenario and higher negative bias in the one effective scenario. It performs well in terms of Variance, but the MSE is greatly affected by the Bias.

In this case, we suggest that the best performing Estimators are the Single-Iteration Bias-Adjusted Estimator and the Unbiased Estimator. The SI performs well in terms of Bias and has relatively low Variance, resulting in good MSE. Instead, the UMVCUE has high Variance and MSE, but completely eradicates the Bias, which is our primary scope and is of interest to understand the extent of the bias induced by the selection.

4.2 Sub-population selection with normally distributed endpoint

The second context in which we are going to assess the estimators' performances via simulations is the Sub-Population Selection (for the methodology used we refer to section 3.2). Our setting is very similar to the one of Kimani et al. [10], where we have 4 sub-populations and two treatment arms: effective dose and placebo. One interim analysis is done with sub-population selection, based on a normally distributed endpoint. We consider three different scenarios: linear effects in the partitions $\boldsymbol{\delta} = (-0.2, -0.1,\ 0\ , 0.1)$, effectiveness in only one partition $\boldsymbol{\delta} = (0\ , 0\ , 0\ , 0.1)$ and no effects in all partitions $\boldsymbol{\delta} = (0\ , 0\ , 0\ , 0)$. The overall variance is $\sigma^2 = 1$, the threshold is set to $b = 0$ and we have a total of $N = 720$ patients. Also in this case, we consider the interim analysis at three different moments, when the information fraction is equal to $\frac{1}{3}$ ($N_1 = 240$), $\frac{1}{2}$ ($N_1 = 360$) and $\frac{2}{3}$ ($N_1 = 480$) and performed 50000 simulations for each case.

Note that the N_1 patients are evenly assigned to each arm and partition and, after the interim analysis, the $N_2 = N - N_1$ patients are evenly assigned to each arm and selected partition.

In Figure 4.2 we see the performances of the sub-population effect estimation compared in terms of Bias, Variance and MSE in units of approximate Standard Error:

- The Naive Estimation (N) has the highest bias in almost all the scenarios, but also one of the lowest Variances, good performing in terms of MSE.
- The Unbiased Estimator by Robertson et al. [14] (UMVCUE) has zero Bias but High Variance, resulting in high MSE.
- The Unbiased Estimator by Kimani et al. [9] (UMV) has also zero Bias but Higher Variance with respect to the previous one, resulting in higher MSE.
- The Single-Iteration (SI) and Multiple-Iteration (MI) Bias-Adjusted Estimators have small Bias, but slightly more Variance with respect to Naive Estimation, resulting in little more MSE than the latter. Among the two, the Single-Iteration is preferred because it is faster to compute and performs better in terms of Bias.
- The two Shrinkage Estimators (S1, S2) performs very similar to the Naive Estimation: among the two, the one from Carreras and Brannath [5] is preferable (lower MSE when effects are linear).

In this case, we also recommend the Single-Iteration Bias-Adjusted Estimator and the Unbiased Estimator by Robertson et al. [14]. The former performs well in terms of Bias, Variance and MSE; the latter eradicates the Bias at the expense of higher Variance and MSE.

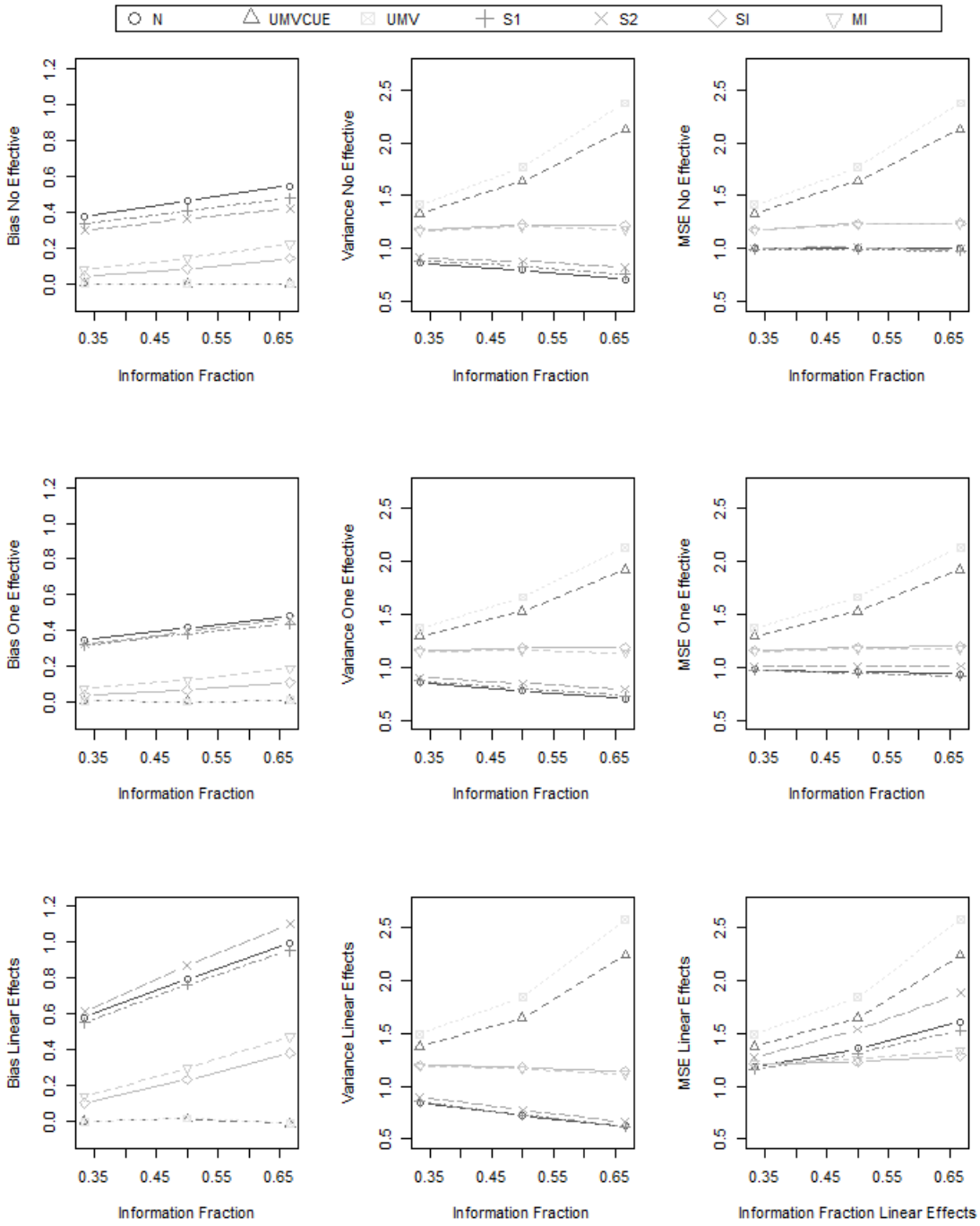

Figure 4.2: *Estimators' Performance in Sub-Population Selection with Normally Distributed Endpoints. Top Row: No Effective Treatment; Middle Row: One Effective Treatment; Bottom Row: Linear Effects of Treatments. Left Column: Bias; Centre Column: Variance; Right Column: Mean Squared Error.*

4.3 Sub-population selection with time to event endpoint

The last context we consider is a Sub-Population Selection with Time to Event endpoint, following the methodology of section 3.3. In this case, let consider 3 sub-populations and two treatment arms: effective dose and placebo. Patients are recruited evenly from the 3 sub-populations during a period of maximum 3 years and equally assigned to the arms. We suppose that the hazard function is constant for all the treatments and equal to $h_c = 0.0005$ for the control. Three cases of LHR are analysed: treatment ineffective in all partitions $\boldsymbol{\delta} = (0,0,0)$; treatment effective only in one partition $\boldsymbol{\delta} = (0\ ,0\ ,-0.3)$; linear effect on the partitions $\boldsymbol{\delta} = (-0.1,-0.2,-0.3)$. For all the cases, the threshold is set to $b = -0.1$ and 5000 simulations are performed.

The total number of events is calculated to detect a hazard ratio of 0.8 reaching a power of 80% with a two-sided type 1 error of 5%. Thus, we set that: the time of the interim analysis t_1 is after 316 events; the time until stage 1 patients are followed in stage 2 $\tilde{t}_1$ is set 6 months after the first interim analysis; the ending time of the trial t_2 is set when all 632 events occur.

In Figure 4.3 the results are shown:

- The Naive Estimation (N) has the highest Bias but very low Variance, resulting in overall average MSE.
- The Unbiased Estimator (UMVCUE) has approximately zero Bias but the highest Variance, resulting in the highest MSE.
- The Single-Iteration (SI) and Multiple-Iteration (MI) Bias-Adjusted Estimator have similar performances: they have very small Bias, but more Variance with respect to Naive Estimation, resulting in MSE comparable to Naive Estimation. Among the two, the Single-Iteration performs better in terms of Bias.
- The two Shrinkage Estimators (S1, S2) perform similarly: they have some Bias (lower than Naive Estimation) but the lowest Variance, resulting in the lowest MSE.

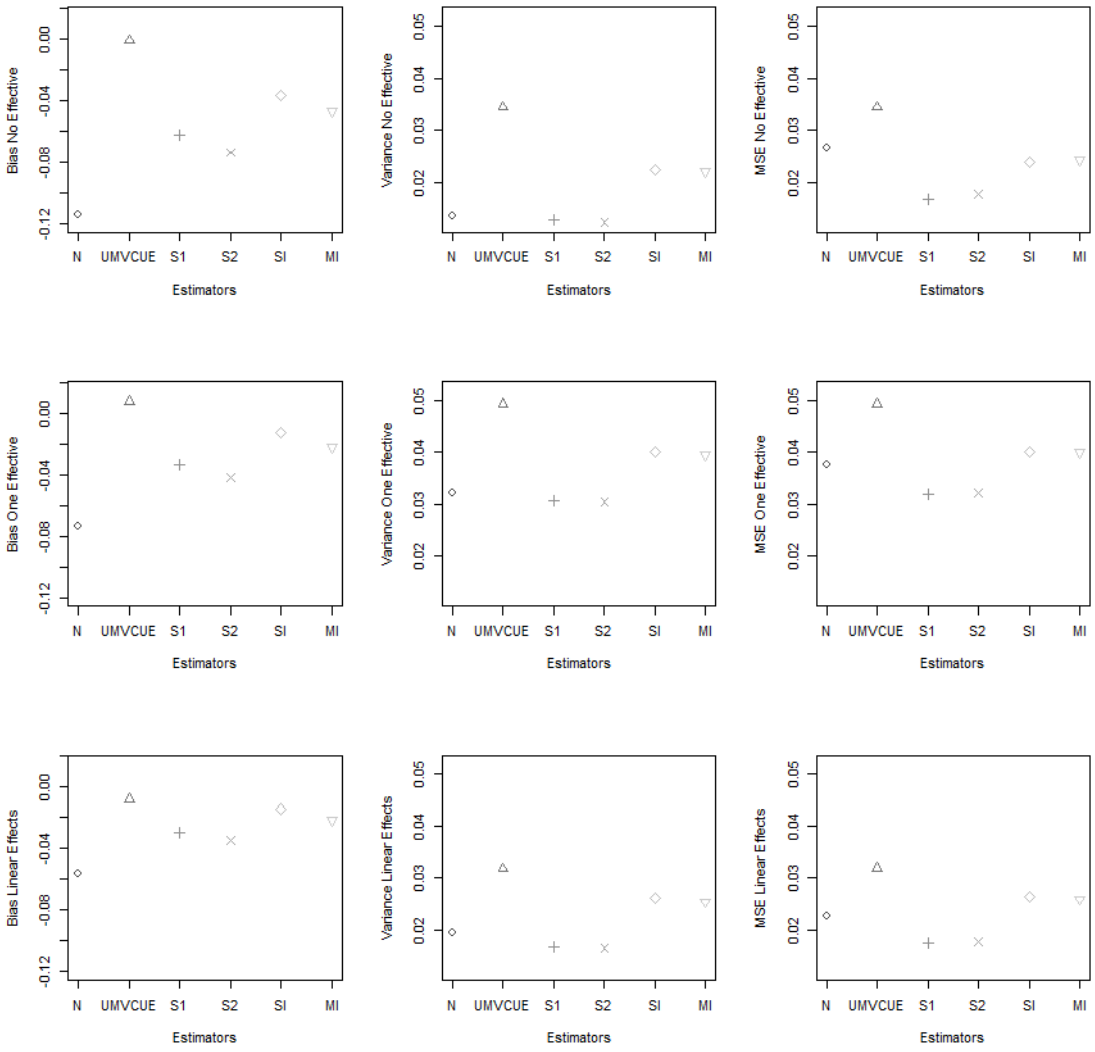

Figure 4.3: *Estimators' Performance in Sub-Population Selection with Time to Event Endpoint. Top Row: No Effective Treatment; Middle Row: One Effective Treatment; Bottom Row: Linear Effects of Treatments. Left Column: Bias; Centre Column: Variance; Right Column: Mean Squared Error.*

In this third case, the suggestion is again to use the Single-Iteration Bias-Adjusted Estimator and the UMVCUE. The first has low Bias and reasonable Variance, resulting in MSE similar to Naive Estimation; the latter eradicates the Bias, but has very high Variance and MSE.

Chapter 5

Case Studies

In this Chapter, we describe two main applications of the estimators analysed previously on Alzheimer Disease and Heart Failure. Both of these case studies have been used to construct the previous simulations, where similar effects have been used. The methodology of the case-study Alzheimer's Disease follows Section 3.1, while the methodology for the case-study Heart Failure follows Section 3.3.

5.1 Alzheimer Disease

The first case study is about a trial for Alzheimer's disease treatment. In this trial, $K = 4$ doses of an experimental treatment are compared against a control. The primary endpoint is the ADAS-Cog change from baseline at week 24, assumed normally distributed. The interim analysis is planned after $N_1 = 270$ patients, equally split in the 5 arms. The stage 1 maximum-likelihood estimations of the treatment-placebo differences are reported in Table 5.1. In this case, from MLE, the dose 4 is selected to proceed with the placebo to stage 2. The number of patients recruited in stage 2 is $N_2 = 109$, split in the two arms. The result of the stage 2 estimation is also reported in Table 5.1.

Using these data, the different estimators for the dose 4 effect versus placebo at the end of the study are reported in Figure 5.1.

	Dose 1	**Dose 2**	**Dose 3**	**Dose 4**
Stage 1	**1.178**	**1.159**	**2.041**	**3.157**
Mean (SE)	(1.062)	(1.062)	(1.062)	(1.062)
Stage 2	-	-	-	**3.334**
Mean (SE)				(1.270)

Table 5.1: *Stage 1 and Stage 2 Treatment-Placebo Difference Estimations for Alzheimer Disease*

	Estimators
N	3.246
UMVCUE	3.23
S1	3.119
S2	2.877
SI	3.168
MI	3.141

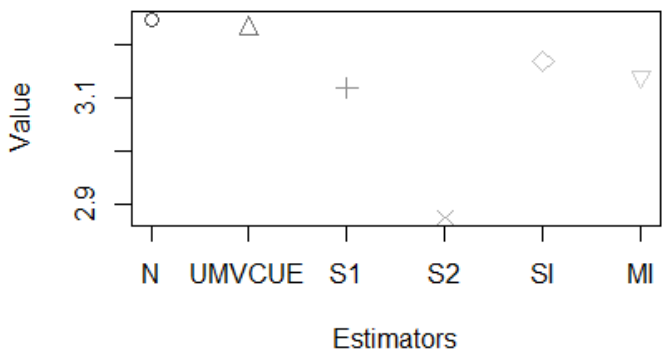

Figure 5.1: *Estimations for Alzheimer Disease*

As can be seen in Figure 5.1, both the UMVCUE and single-iteration bias-adjusted estimator reduce the optimism of the naive estimation. The first shrinkage estimator and the multi-iteration bias-adjusted estimator are even more conservative. The second shrinkage estimator seems outside the box giving a very conservative estimation: this seems in accordance with the simulations made in the case of only one effective treatment, which is similar to our case.

5.2 Heart Failure

The second case study regards a treatment on heart failure. In this case, an experimental treatment is compared to placebo, and the initial patient popu-

lation is partitioned in $K = 3$ sub-populations according to the baseline heart rate: low heart rate (below 75 bpm); medium heart rate (between 75 and 81 bpm); high heart rate (above 81 bpm). The primary endpoint is the time from randomisation to cardiovascular death or hospital admission for worsening of heart failure. The main analysis is a Cox proportional hazard model adjusted for previous beta-blocker intake at randomisation, and the treatment effect is estimated using the hazard ratio between the two treatment arm. The interim analysis is done after 630 events occurred and the stage 1 patients are followed for 13 months after the analysis; the trials stops when 1260 events have occurred. The futility threshold on the log-hazard ratio scale is set to $b = -0.1$.

In Table 5.2 we can see that only the low heart rate sub-population is dropped at the interim analysis, while the others continue to stage 2. With the data from Table 5.2 we can proceed with the estimation of the treatment effect in the selected sub-populations at the end of the study, the results of which are shown in Figure 5.2:

- For the sub-population with medium heart rate, we see that the UMVCUE and single-iteration bias-adjusted estimator provide more conservative estimations with respect to the naive one. Also, the multi-iteration bias-adjusted estimator is less optimistic but does not vary very much from the previous two. On the other hand, the two shrinkage estimators provide much more conservative estimations.

- For the high heart rate sub-population, we see that the naive estimation is indeed the most optimistic, but the UMVCUE, single-iteration and multiple-iteration bias-adjusted estimators do not provide a hugely different estimation, remaining close to each others. The shrinkage estimators provide a more conservative estimation, instead.

Log HR	Low Heart Rate	Medium Heart Rate	High Heart Rate
Stage 1	**-0.075**	**-0.397**	**-0.358**
(SE)	(0.155)	(0.150)	(0.121)
Stage 2	-	**-0.086**	**-0.363**
(SE)		(0.122)	(0.107)

Table 5.2: *Stage 1 and Stage 2 Estimations for Heart Failure*

	Medium	High
N	-0.27319	-0.36024
UMVCUE	-0.26222	-0.35979
S1	-0.19417	-0.34132
S2	-0.20594	-0.34339
SI	-0.25605	-0.35783
MI	-0.26018	-0.35856

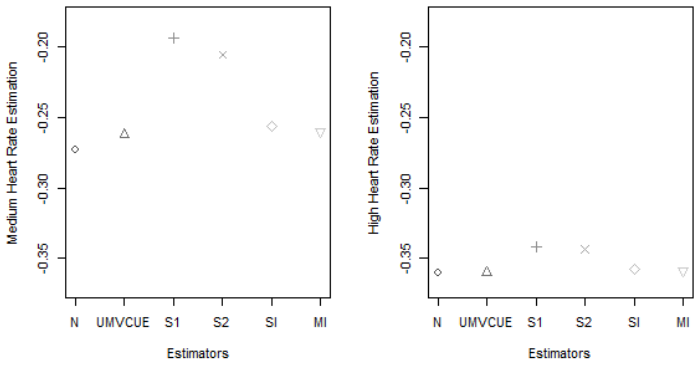

Figure 5.2: *Estimations for Heart Failure*

Chapter 6

Conclusions

In this work, we investigated how to identify and handle the bias in treatment effect estimation in adaptive designs. We identified the source of bias in the selection process that occurs at interim analyses. Two main type of bias are described: the selection bias, a positive bias that affects the selected treatments and the always reporting bias, a negative bias that affects the dropped treatments. We were mainly interested in handling the selection bias, using several estimators found in literature [3, 4, 5, 15]. We studied the properties of these estimators and their performances in different ways: several scenarios were analysed, like treatment selection, sub-population selection with normally distributed endpoint and sub-population selection with time-to-event endpoint; also, we described selection rules based on the most effective treatment, or determined by a futility threshold.

We conducted an extensive simulation study in order to compare the performances of the estimators in terms of bias, variance and mean squared error. Of course, the bias is of primary interest, but also the variability is of great interest: a low biased but highly variable estimation may not be preferred over a slightly more biased but much less variable one. We recommend in all the cases to present the unbiased estimator and the single-iteration bias-adjusted estimator: the former completely eradicates the bias, but is highly variable with respect to a naive estimation; the latter is less biased than a naive estimation,

but only slightly more variable. Thus, this two estimators are the suggested for a general application.

For completeness, two appendices are added to this work: the first focuses on the construction of simultaneous confidence intervals for the MLE to avoid overoptimism in treatment effect estimation; the second one focuses on bias in adaptive designs with sample size reassessment. These set up the ground for further exploration.

Acknowledgements

We would like to thank Dr. Maximo Carreras and Professor Peter K. Kimani for providing the code from their works [5, 10] that helped developing this thesis in the early phases.

Bibliography

[1] Bassler D, Briel M, Montori VM, Lane M, Glasziou P, et al. *Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis*. JAMA. 2010;303:1180–1187.

[2] Bauer P, Koenig F, Brannath W, Posch M. *Selection and bias – two hostile brothers*. Statistics in Medicine 2010; 29(1): 1–13

[3] Bowden J, Glimm E. *Unbiased estimation of selected treatment means in two-stage trials*. Biometrical Journal 2008; 50(4):515–527.

[4] Brückner M, Titman A, Jaki T. *Estimation in multi-arm two-stage trials with treatment selection and time-to-event endpoint*. StatistMed. 2017;36:3137-3153.

[5] Carreras M, Brannath W. *Shrinkage estimation in two-stage adaptive designs with midtrial treatment selection*. Statist Med. 2013;32(10):1677-1690.

[6] Cohen A, Sackrowitz HB. *Two stage conditionally unbiased estimators of the selected mean*. Stat Probab Lett. 1989;8:273-278.

[7] Friede T and Kieser M. *Blinded sample size re-estimation in superiority and noninferiority trials: bias versus variance in variance estimation*. Pharmaceut Stat 2013; 12: 141–146.

[8] Glimm E. *Adjusting for selection bias in assessing treatment effect estimates from multiple subgroups*. Biometrical Journal. 2019;61:216–229.

[9] Kimani PK, Todd S, Stallard N. *Estimation after subpopulation selection in adaptive seamless trials*. Statist Med. 2015;34:2581-2601.

[10] Kimani PK, Todd S, Renfro LA, Stallard N. *Point estimation following two-stage adaptive threshold enrichment clinical trials*. Stat Med. 2018;37:3179-3196.

[11] Kimani PK, Todd S, Renfro LA, et al. *Point and interval estimation in two-stage adaptive designs with time to event data and biomarker-driven subpopulation selection*. Statistics in Medicine. 2020;1–19.

[12] Pallmann P, Bedding AW, Choodari-Oskooei B, Dimairo M, Flight L, Hampson LV, et al. *Adaptive designs in clinical trials: why use them, and how to run and report them*. BMC Med 2018 Dec 28;16(1):29

[13] Posch M, Klinglmueller F, König F, Miller F. *Estimation after blinded sample size reassessment*. Stat Methods Med Res. 2018;27(6):1830–46.

[14] Robertson DS, Prevost AT, Bowden J. *Unbiased estimation in seamless phase II/III trials with unequal treatment effect variances and hypothesis-driven selection rules*. Statist Med. 2016a;35(22):3907-3922.

[15] Stallard N, Todd S. *Point estimates and confidence regions for sequential trials involving selection*. J Stat Plann Infer. 2005;135:402-419.

[16] Whitehead J. *On the bias of maximum likelihood estimation following a sequential test*. Biometrika. 1986;73(3):573-581.

Appendix A

Confidence Intervals

While the previous discussion focused on the methods to obtain a Point Estimation that handles the Selection Bias in Adaptive Designs, we now move our attention towards the construction of accurate Confidence Intervals with joint $(1-\alpha)$ coverage for the naive estimator (α being the type 1 error), addressing the multiplicity issue (multiplicity of treatment arms or of populations). Using Confidence Intervals with good coverage is of fundamental importance to obtain a good appreciation of the uncertainty in the treatment effect estimation. Glimm [8] suggests to use simultaneous confidence intervals for the MLE to avoid overoptimism towards one or more treatments.

We may consider two options: simultaneous confidence intervals with Sidak correction or Bonferroni correction. The idea of Sidak is that considering K tests, independent from each other, the overall coverage of the corresponding will be the product of the single coverages: $(1-\alpha_i)^K$, where α_i is the significance level of the single test. Since we want to get Simultaneous Confidence Intervals with joint $(1-\alpha)$ coverage, we find that:

$$(1-\alpha) = (1-\alpha_i)^K \implies \alpha_i = 1-(1-\alpha)^{\frac{1}{K}}$$

Therefore, using the significance level of $\alpha_i = 1-(1-\alpha)^{\frac{1}{K}}$ for the confidence interval of the single comparison, we obtain an overall coverage of $(1-\alpha)$.

A more stringent correction is the Bonferroni's one. The idea of Bonferroni is to have the single Confidence Intervals with a coverage of $(1-\frac{\alpha}{K})$. Then, the Simultaneous Confidence Intervals, considering independence between the test statistics, would have $(1-\frac{\alpha}{K})^K$ coverage. If we suppose to have $K=4$ tests and an $\alpha=0.05$ level of significance, we see that with Sidak Correction the single Interval has significance of $\alpha_{i,S}=1-(1-\alpha)^{\frac{1}{K}}=1-(1-0.05)^{\frac{1}{4}}\approx 0.0127$, while with Bonferroni correction it is of $\alpha_{i,B}=\frac{\alpha}{K}=\frac{0.05}{4}=0.0125$. The overall significance would be $\alpha_S=1-(1-\alpha_{i,S})^K=0.05$ for Sidak and $\alpha_B=1-(1-\alpha_{i,B})^K\approx 0.049$ for Bonferroni, which is indeed more conservative.

In Figure A.1 and Figure A.2 are reported the Sidak and Bonferroni simultaneous confidence intervals for the two case studies analysed in Chapter 5 on Alzheimer Disease and Heart Failure.

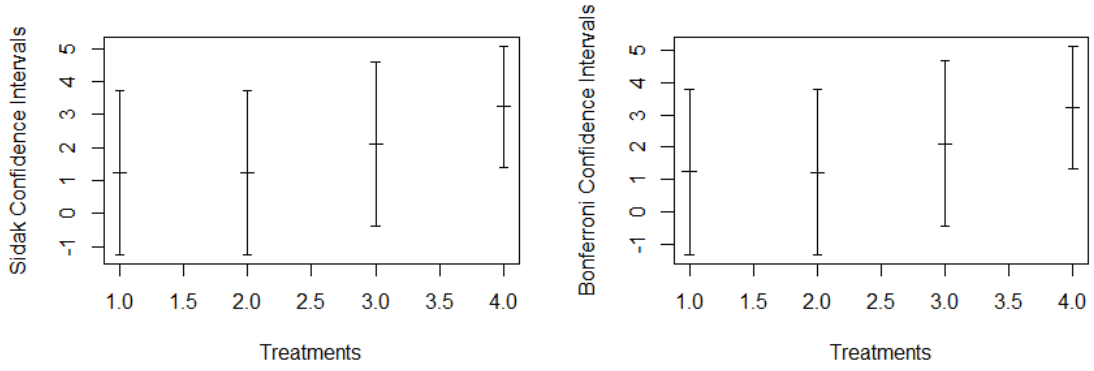

Figure A.1: *Simultaneous Confidence Intervals for Alzheimer Disease*

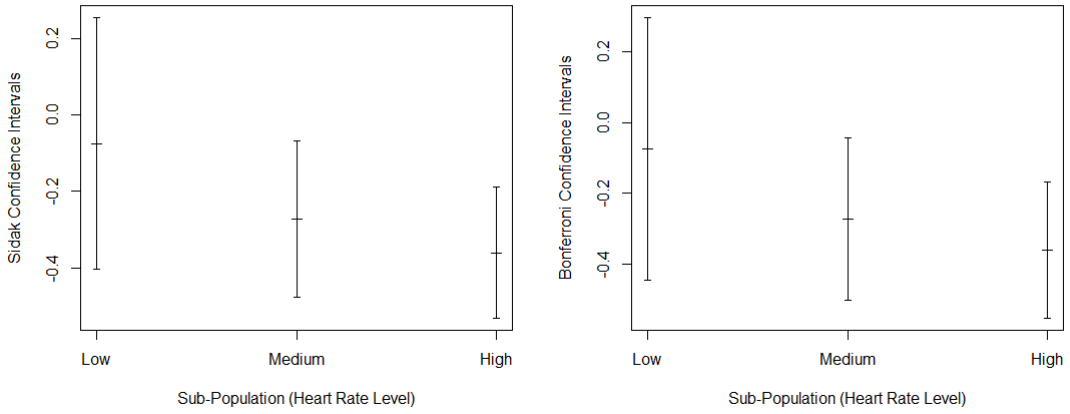

Figure A.2: *Simultaneous Confidence Intervals for Heart Failure*

Appendix B

Sample Size Reassessment

Suppose we have a two group comparison with common variance σ^2 unknown and mean μ_a and μ_b, respectively. We define $\delta = \mu_a - \mu_b$ as the true effect size and we want to test the one-sided null hypothesis $H_0 : \delta \leq 0$ against the alternative $H_A : \delta > 0$ at level α. Let δ_0 denote the alternative for which the trial is powered and σ_0^2 as an initial guess of the true variance. A first stage per group sample size equal to $n_1 \geq 2$ is chosen, based on σ_0^2. At the end of the first stage, a blinded one-sample variance estimation is calculated:

$$S_{1,OS}^2 = \frac{1}{2n_1 - 1} \left[\sum_{i=a,b} \sum_{k=1}^{n_1} (X_{i1k} - \bar{X}_{\cdot 1 \cdot})^2 \right]$$

where X_{i1k} is observation $k = 1, ..., n_1$ in group $i = a, b$ in stage 1 and $\bar{X}_{\cdot 1 \cdot}$ is the overall stage 1 mean. This estimation is an unbiased estimation of the true variance σ^2 when the true effect size $\delta = 0$; otherwise, it has a positive bias equal to $\delta^2 n_1 / (4n_1 - 2)$. Using this estimator, a second stage sample size $n_2(S_{1,OS}^2)$ is calculated and used at stage 2: the higher the variance estimate, the higher the second stage sample size.

Friede and Keiser [7] analysed the performance of the estimator $S_{1,OS}^2$ against a blinded unbiased estimator that exploited block randomization to eradicate the bias. They found that this unbiased estimator has higher variability with

respect to one-sample variance estimator, thus resulting in higher stage 2 sample size and lower power; also, type I error rate would be inflated. On the other hand, using $S^2_{1,OS}$ the type I error rate is controlled (superiority trials) or can be controlled (non-inferiority trials) with available methods. Another option would be to adjust $S^2_{1,OS}$ using an expected treatment effect, but also in this case we would have a reduction in power. Therefore, they suggest to use $S^2_{1,OS}$ for sample size reassessment.

Posch et al. [13] analysed the performance of two sample size reassessment rules: the first uses $S^2_{1,OS}$ as an estimation of variance; the second uses an adjusted variance estimation which is unbiased under the effect size δ_0, i.e. $S^2_{1,OS} - \delta_0^2 n_1/(4n_1-2)$. They are:

$$n_2^u(S^2_{1,OS}) = min\left\{n_{2max}, max\left\{n_{2min}, 2(z_{1-\alpha}+z_{1-\beta})^2\frac{S^2_{1,OS}}{\delta_0^2} - n_1 + 1\right\}\right\}$$

$$n_2^a(S^2_{1,OS}) = min\left\{n_{2max}, max\left\{n_{2min}, 2(z_{1-\alpha}+z_{1-\beta})^2\left(\frac{S^2_{1,OS}}{\delta_0^2} - \frac{n_1}{4n_1-2}\right) - n_1 + 1\right\}\right\}$$

where $1-\beta$ is the desired power and n_{2min} and n_{2max} are pre-specified second stage minimum and maximum sample size.

What they find is that with both rules the bias of confidence intervals may be large for small first stage sample sizes, but is small otherwise. Under the null hypothesis, confidence intervals do not exhibit an inflation of non-coverage probabilities, even for small first stage sample size, while very small inflations are observed otherwise. Moreover, for positive δ the lower bound of the confidence interval is conservative while the upper one is anti-conservative; for negative δ the upper bound of the confidence interval is conservative while the lower one is anti-conservative.

As regards the bias, they find that for the adjusted rule the bias of mean and variance estimate is bigger than with the unadjusted rule. In both cases,

the bias is of opposite sign of the true effect δ and decreasing with first stage sample size, but always noticeable in the worst cases. Moreover, with increasing effect size the maximum bias of the blinded sample size reassessment rule approaches the one of the unblinded one. Therefore, also in this case the recommendation would be to use the unadjusted rule with $S^2_{1,OS}$ as estimation.